## Lab Exercise One Data Preprocessing with SPSS Modeler

## Handling Missing Data

- 1. Download data file *small\_telco.xls* from course website, save it on desktop or a folder of your choice.
- 2. Open IBM SPSS Modeler, choose Create a new stream.

BM® SPSS® Modeler	
IBM SPSS Modeler	IBM.
What would you like to do?	
O Open a demo stream	© Create a <u>n</u> ew stream
More files Introduction to modeling Automated modeling for a binary target Automated modeling for a numeric target	C Launch the Application Examples tutorial
O Open an existing stream	Create a new model with IBM® SPSS® Modeler Advantage
More files Stream1_DataProcessing.str modelingintro.str	Pind out what's new in this release
© Open an existing project	
More files	Don't show this dialog in future
	OK Cancel

3. Put an Excel Source Node on the stream canvas.

😻 Stream1* - IBM© SPSS© Modeler 📃 🗖 🔀
Eile Edit Insert Yiew Tools SuperNode Window Help
🛅 🚍 🔒 🙏 🖏 🖺 🖆 🖛 🛥 🏦 🗹 🌚 🕨 🎓 🞸 Strems Outputs Models
Excel  Stream  CRSP-DM  Classes  CRSP-DM  CRSP-D
Favorites Sources Record Ops Field Ops Agraphs Modeling Output Export OBM® SPSS® Statistics BM® SPSS® Text Analytics
Enterprise View Database Var. File Fixed File Statistics File Data Collection IBM Cognos BI SAS File Excel XML User Input
Server: Local Server 259MB / 362MB

4. Import data file into the stream, keep the default settings, and click Preview to check the data.

😵 small_telco.xls						
EXCET Previe	W 2 Refresh	0				
D:\Fordham	University\CS4631\Spring2012\SPSS\small_	_telco.xls				
Data Filter Types	Annotations					
File type:	Excel 97-2003 (*.xls)	*				
Import file:	D:/Fordham/University/CS4631\Spring2012\SPSS\small_telco.xts					
🗾 Use named range						
Choose worksheet:	By index	0 🗧				
	© By name					
Range on worksheet:	Range starts on first non-blank row					
	C Explicit range of cells					
On blank rows:	Stop reading 📃 👻					
📝 First row has colum	in names					
OK Cancel		Apply Reset				

5. Add a Type Node on the stream canvas, and connect the Source Node with the Type Node.



6. Double click the Type Node, set the appropriate measurement level for every field of the data. Set the role of the last field Churn as Target.

😰 🚺	Preview				0
Types Formet	Appotetions				
	Read Vi	alues Clea	r Values	Clear All V	alues
Field 💳	Measurement	Values	Missing	Check	Role
💞 internet	Flag	1.0/0.0		None	
Ioglong	Continuous	[-0.10536	-	None	
Iogequi	Continuous	[2.73436		None	> Input
Iogcard	Continuous	[1.01160		None	🔪 Input
🛞 logwire	Continuous	[2.70136		None	🔪 Input
Ininc	Continuous	[2.19722		None	🔪 Input
🎒 custcat	🂑 Nominal	1.0,2.0,3		None	🔪 Input
🤀 churn	🎖 Flag	1.0/0.0		None	🔘 Target
View current	t fields 🔘 View unu	sed field setting	gs		

7. Add a Data Audit Node on the stream canvas, connect it with the Type Node.



8. Double click the Data Audit Node, keep the default settings, click Run button. The statistics and charts are shown below.

Audit Guali	ty Annotations								
Field -	Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
🌮 region		💑 Nominal	1.000	3.000			-	3	100
🖗 tenure		🛷 Continuous	1.000	72.000	35.526	21.360	0.112	-	100
🖗 age		🛷 Continuous	18.000	77.000	41.684	12.559	0.357	-	100
🖗 marital		🎖 Flag	0.000	1.000	5		-	2	10
address 🤇		🔗 Continuous	0.000	55.000	11.551	10.087	1.106	-	10
🖗 income		🔗 Continuous	9.000	1668.000	77.535	107.044	6.643	52	10
🖗 ed		📲 Ordinal	1.000	5.000		-	-	5	10
🖗 employ		🔗 Continuous	0.000	47.000	10.987	10.082	1.061		100

9. Click Quality Tab, then specify impute method for missing values in fields *logequi, logcard, logwire.* 

Complete fields	(%): 86.36% Co	mplete recor	rds (%): 16.	7%					
Field	Measurement	Outliers	Extremes	Action	Impute Missi	. Method	% Complete	Valid Records	Null Value
🚯 region	💑 Nominal			2	Never	Fixed	100	1000	
tenure	🔗 Continuous	0	0	None	Never	Fixed	100	1000	
🖗 age	Continuous	0	0	None	Never	Fixed	100	1000	
🚯 marital	🎖 Flag				Never	Fixed	100	1000	
🚯 address	🔗 Continuous	12	0	None	Never	Fixed	100	1000	
income	Continuous	9	6	None	Never	Fixed	100	1000	
🚯 ed	Ordinal				Never	Fixed	100	1000	
🚯 employ	🔗 Continuous	8	0	None	Never	Fixed	100	1000	
🚯 retire	🎖 Flag				Never	Fixed	100	1000	
🚯 gender	💑 Nominal	12			Never	Fixed	100	1000	
🚯 reside	💑 Nominal				Never	Fixed	100	1000	
🚯 longmon	🔗 Continuous	18	4	None	Never	Fixed	100	1000	
🚯 longten	🔗 Continuous	20	4	None	Never	Fixed	100	1000	
👂 internet	🎖 Flag			-	Never	Fixed	100	1000	
🚯 ebill	🎖 Flag	<u>11</u>	<u>.</u>	-	Never	Fixed	100	1000	
🚯 loglong	🔗 Continuous	4	0	None	Never	Fixed	100	1000	
🚯 logequi	🗌 🔗 Continuous	1	0	None	Blank & Null	. Fixed	38.6	386	
Iogcard	🔗 Continuous	2	0	None	Blank & Null	. Fixed	67.8	678	
logwire 😥	🔗 Continuous	1	0	None	Blank & Null	. Fixed	29.6	296	
🚯 Ininc	🔗 Continuous	9	0	None	Never	Fixed	100	1000	
🚯 custcat	💑 Nominal			-	Never	Fixed	100	1000	
🖗 churn	🎖 Flag		<u>.</u>	123	Never	Fixed	100	1000	

10. Generate Missing Values SuperNode with all fields, then connect it with the Type Node.

Stream1* - IBM® SPSS® Modeler	
File Edit Insert ⊻iew Tools SuperNode Window Help	4
🖻 🚍 🖶 🖨 🗶 🍡 🖺 🕿 🛥 🛤 🗹 🚳 🕨 🌒 🗶	Streams Outputs Models
small_telcoxis	E Stream
Missing Value Imputa.	CRISP-DM Classes CIRISP-DM Classes CIRISP-CM (unsaved project) CirisPublic Standard
Severites Sources Record Ops Field Ops A Graphs Modeling Output Export OIBM® SPSS® Star	istics BM® SPSS® Text Analytics
Image: Constraint of the second se	4
Server: Local Server	281MB / 402MB

11. Generate Missing Values Filter Node with quality percentage higher than 50%, then connect the node with the Missing Value Impute SuperNode.



12. Select two fields with Outliers and Extreme values, choose appropriate Actions, and generate Outlier and Extreme SuperNode. Then connect it with the Filter Node just created. You could Zoom In the SuperNode to have a look at its details.

Complete fields	: (%): 86.36% Co	mplete recor	ds (%): 16.7%					
Field	Measurement -	Outliers	Extremes Action	Impute Missing	Method	% Complete	Valid Records	Null Val
😥 region	💑 Nominal			Never	Fixed	100	1000	
tenure	Continuous	0	0 None	Never	Fixed	100	1000	
🚯 age	Continuous	0	0 None	Never	Fixed	100	1000	
😥 marital	🎖 Flag			Never	Fixed	100	1000	
address	Continuous	12	0 Coerce	Never	Fixed	100	1000	
income	Continuous	9	6 Coerce	Never	Fixed	100	1000	
😥 ed	- Ordinal			Never	Fixed	100	1000	
employ	🔗 Continuous	8	0 None	Never	Fixed	100	1000	
👰 retire	🔓 Flag			Never	Fixed	100	1000	
🚯 gender	💑 Nominal	<u></u>	C2 22	Never	Fixed	100	1000	
😥 reside	💑 Nominal		52 L2	Never	Fixed	100	1000	
Iongmon	🔗 Continuous	18	4 None	Never	Fixed	100	1000	
Iongten	🔗 Continuous	20	4 None	Never	Fixed	100	1000	
internet	🎖 Flag			Never	Fixed	100	1000	
훶 ebill	🎖 Flag			Never	Fixed	100	1000	
🚯 loglong	🔗 Continuous	4	0 None	Never	Fixed	100	1000	
👂 logequi	Continuous	1	0 None	Blank & Null Val	Fixed	38.6	386	
🚯 logcard	🔗 Continuous	2	0 None	Blank & Null Val	Fixed	67.8	678	
🚯 logwire	🔗 Continuous	1	0 None	Blank & Null Val	Fixed	29.6	296	
💫 Ininc	🔗 Continuous	9	0 None	Never	Fixed	100	1000	
🚯 custcat	💑 Nominal			Never	Fixed	100	1000	
😥 churn	🎖 Flag		22 <u>2</u> 2	Never	Fixed	100	1000	





13. Add another Type Node on the stream canvas, connect it with the SuperNode. Then reset the measurement levels of these two fields you just processed.

Types Format	Annotations				0
4. 00	PRead Va	alues Clea	r Values	Clear All V	alues
Field =	Measurement	Values	Missing	Check	Role
region	💑 Nominal	1.0,2.0,3.0	_	None	🔪 Input
tenure	🖉 Continuous	[1.0,72.0]		None	🔪 Input
age 😥	🖉 Continuous	[18.0,77.0]		None	🔪 Input
marital	🎖 Flag	1.0/0.0		None	🔪 Input
address	🖌 <default></default>	<read></read>		None	🔪 Input
income		<read></read>		None	🔪 Input
ned 🚯	Ordinal	1.0,2.0,3		None	🔪 Input
employ	🔗 Continuous	[0.0,47.0]		None	🔪 Input
A ratira	V Flan	10/00		None	N Innut
OK Cancel	fields 🔘 View unu:	sed field setting	jS		Apply Res

14. Then you could save the preprocessed data into an Excel file by adding an Excel Export Node on the stream canvas, then connect it with the Type Node. Double click the Excel Export Node to choose a location for the export data file. Save it as *processed\_telco.xls*.



15. Open *processed\_telco.xls*, compare it with the original *small\_telco.xls*, what are the differences?

## Binning

- 16. Create a new stream and load processed\_telco.xls into SPSS Modeler, connect it with a Type node, defining the measurement levels for fields.
- 17. Create a Binning node and add it on the stream.

Image: Stream1* - IBM⊚ SPSS⊙ Modeler	
File Edit Insert View Tools SuperNode Window Help	
🔁 🚍 🖶 🖨 X 🍡 🖺 🖝 🛥 🏙 🗹 🌚 🕨 🌒 🛠	Streams Outputs Models
processed_telco_labO Type	CRISP-DM Classes (CRISP-DM Classes (unawed project) Data Under standing Data Under standing Data Videing Evaluation Evaluation
💈 Favorites 💿 Sources 💿 Record Ops 💿 Field Ops 🛕 Graphs 💿 Modeling 🔳 Output 🔲 Export ④ IBM® SPSS® Stat	istics IBM® SPSS® Text Analytics
	E E
Auto Data Prep Type Filter Derive Ensemble Filler Anonymize Reolassify Binning RFM Analysis Partition	n SetToFlag Restructure Transpose Tir
Server: Local Server There are no executable nodes	281MB / 417MB

18. Choose age field to performing binning.

😵 Binning				
	erate Preview		0	
Settings Bin Value	Annotations			
₿in fields:	🔗 age			×
Binning method: Fix	(ed-width			
Fixed-width Binning				
Name extension:	_BIN	Add as:	🔘 Suffix	O Prefix
Bin width	10.0 🗲			
O No. of bins	10 🖨			
Bin thresholds: 🛛 🔘	Always recompute			
0	Read from Bin Values tab if available			
OK Cancel			Abt	ly <u>R</u> eset

🛞 Binning		×
Settings Bin Values Annotation	Preview	<b>0</b> - <b>-</b>
Binning method: Fixed-width (Bin v	vidth = 10.0)	
Binned field: age Tile:	es shown in the table	•
Bin	Lower	Upper
1	>= 13	< 23
2	>= 23	< 33
3	>= 33	< 43
4	>= 43	< 53
5	>= 53	< 63
6	>= 63	< 73
7	>= 73	<= 82
OK Cancel		Apply Reset

19. Click preview button to see the result.

🖩 Preview from Binning Node (21 fields, 10 records) #3										
📦 <u>F</u> ile	📄 <u>E</u> dit (	🕙 <u>G</u> ene	erate 🔂 🕒	14	<b>2 2 2</b>			0 X		
Table ,	Annotations									
	llong		logcard		Ininc	custcat	churn	age_BIN		
1		1.308	2.0	15	4.159	1.000	1.000	4		
2		1.482	2.7	25	4.913	4.000	1.000	3		
3		2.899	3.4	-09	4.754	3.000	0.000	4		
4		2.246	0.0	100	3.497	1.000	1.000	3		
5		1.841	0.0	00	3.401	3.000	0.000	2		
6		2.468	2.6	03	4.357	3.000	0.000	3		
7		2.389	2.1	69	2.944	2.000	1.000	1		
8		1.800	3.1	46	4.331	4.000	0.000	3		
9		2.277	2.4	85	5.112	3.000	0.000	5		
10		3.184	2.8	03	4.277	2.000	0.000	3		
	4									
								<u>ok</u>		

- 20. Export the modified data file with the new field added.
- 21. Could you remove the age field with the age\_BIN field only? Which node should you add? Adding a Filter Node.



22. Discard the old age field, save the new age\_BIN field.

😻 Filter		×					
Preview		0					
Fitter Annotations							
7	Fields:	21 in, 1 filtered, 0 renamed, 20 out					
Field -	Filter	Field					
region	$\rightarrow$	region 🖌					
tenure	$\rightarrow$	tenure					
age	<b>→</b>	age					
marital	$\rightarrow$	marital					
address	$\rightarrow$	address					
income	$\rightarrow$	income					
ed	$\rightarrow$	ed					
employ	$\rightarrow$	employ					
retire	$\rightarrow$	retire					
gender	$\rightarrow$	gender 🔽					
OK Cancel		Apply Reset					

🕅 Filter 🛛 🔀								
Preview)		0-						
Filter Annotations								
7.	Fields:	21 in, 1 filtered, 0 renamed, 20 o	ut					
Field 📼	Filter	Field						
longmon	$\rightarrow$	longmon 🛃						
longten	$\rightarrow$	longten						
internet	$\rightarrow$	internet						
ebill	$\rightarrow$	ebill						
loglong	$\rightarrow$	logiong						
logcard	$\rightarrow$	logcard						
Ininc	$\rightarrow$	Ininc						
custcat	$\rightarrow$	custcat						
churn	$\rightarrow$	churn						
age_BIN	$\rightarrow$	age_BIN	-					
View current fields     O View unused field settings								
OK Cancel Apply Reset								