

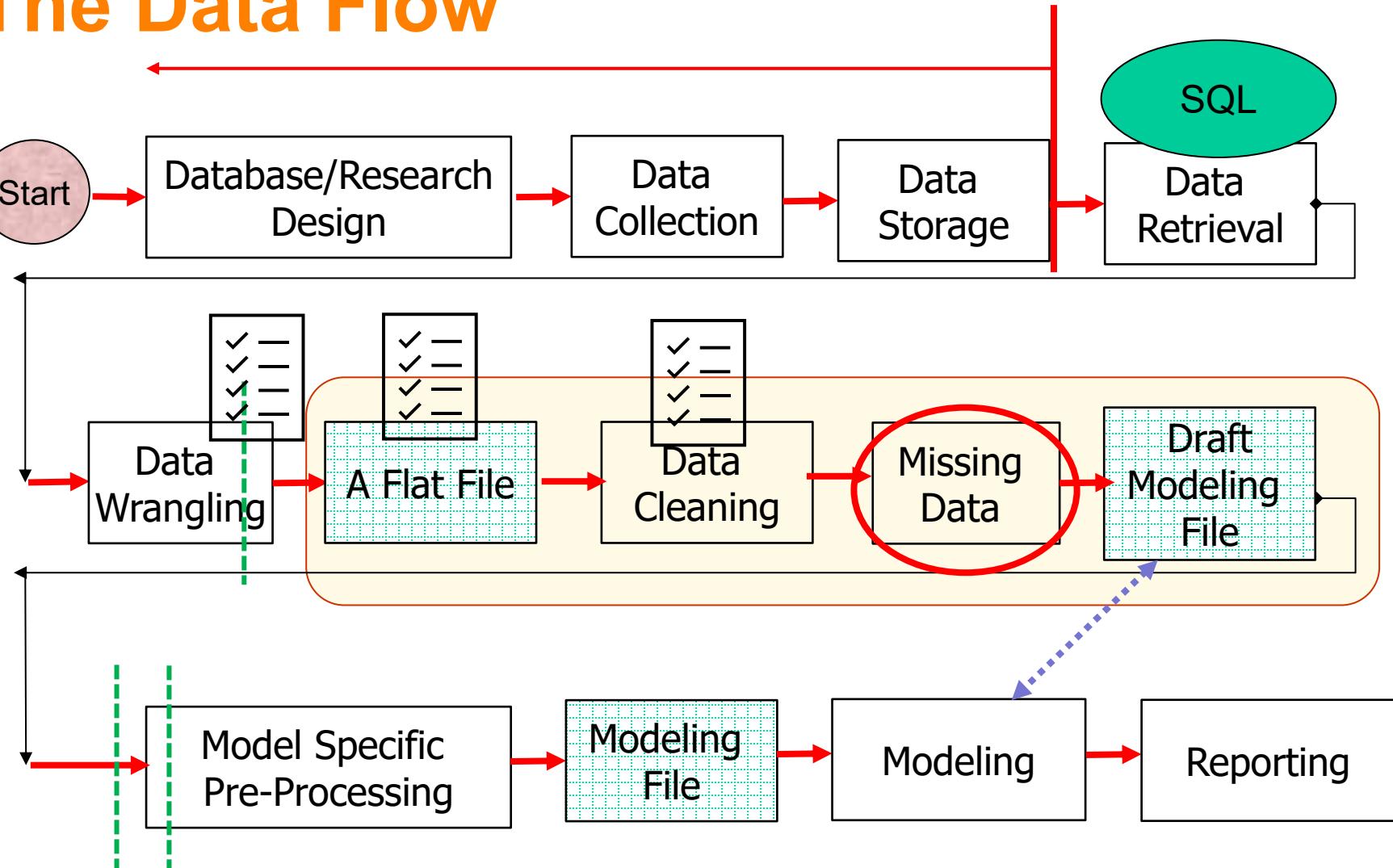
# Missing Data: Over-re-view



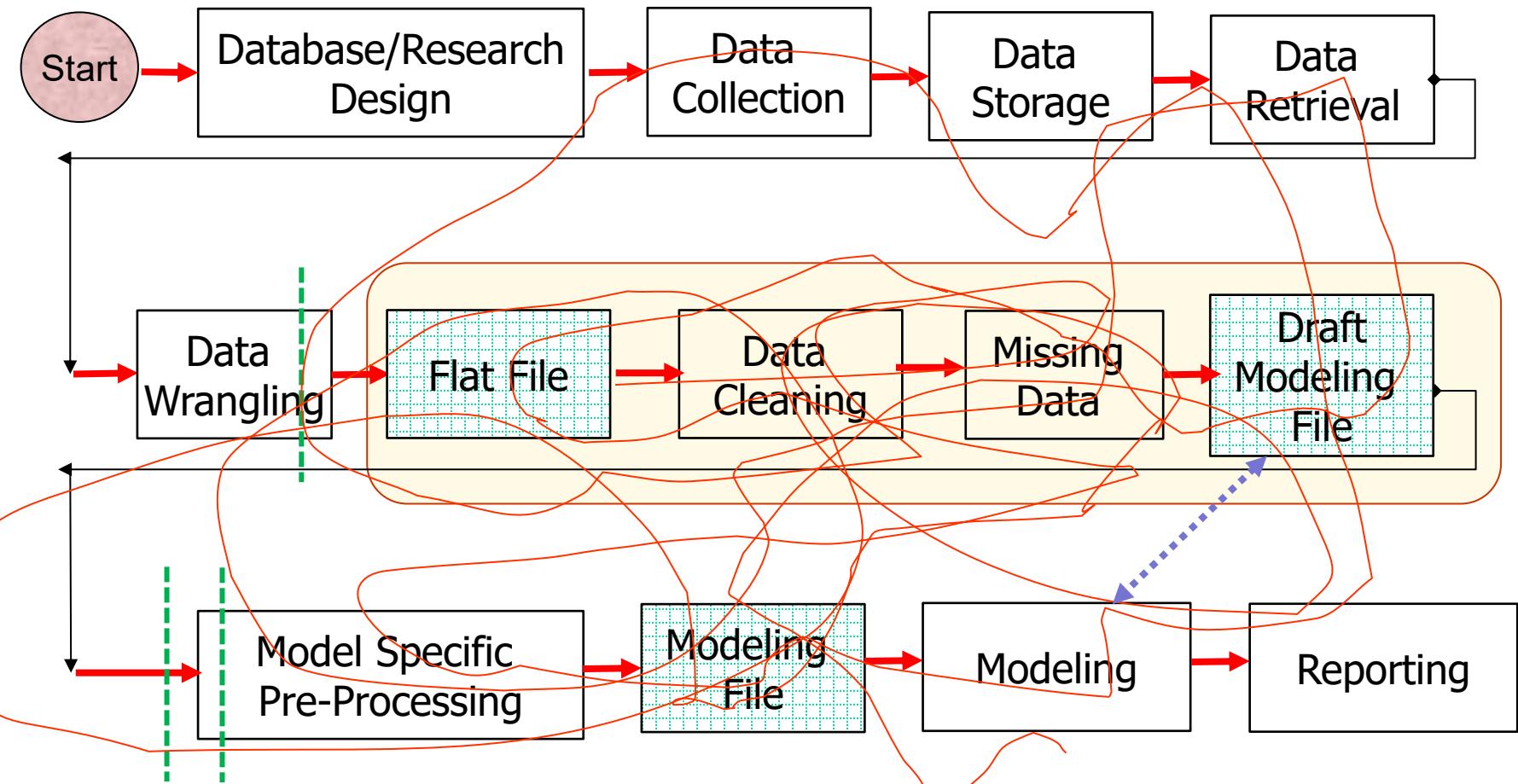
# Most Important Lesson

- Handling missing data is more of a conceptual problem than a technical one.
  - YOU must make decisions.
  - YOU must know the consequences of your decisions
  - YOU must communicate decisions.

# The Data Flow



# The Data Flow: Is it Really Linear?



# Common Missing Data Nomenclature

Missing Completely at Random	Missing at Random	Missing Not at Random
<p>The missingness of data is unrelated to both the observed and unobserved data.</p> <p>The data are literally, <b>MCAR</b>.</p>	<p>The missingness of data is related to the observed data but not to the unobserved data.</p> <p><b>MAR</b> missingness depends only on the observed data and can be predicted by other variables in the dataset.</p>	<p>The missingness of data is related to the unobserved data, even after conditioning on the observed data.</p> <p>This means that the probability of missingness depends on the values of the missing data itself, even after accounting for the observed data.</p> <p><b>MNAR</b> is the most problematic missing data mechanism as it implies that the missing data cannot be explained by the observed data</p>

# Data Cleaning Steps

1. Open data in your software of choice
2. Review variables for common sense based on SME knowledge
3. Review how the software coded the variables (nominal, continuous)
4. Perform data integrity/validation checks (misspelled levels, bogus values, combine levels which represent the same thing, etc.)
5. Handle dates (extract relevant information, e.g. day of week, hour of day)
6. Handle categorical variables - keep as is, combine rare levels, combine similar levels
7. Handle zero-variance predictors, i.e., columns that contain the same number throughout
8. Handle near zero-variance predictors, i.e., columns that contain very little variety in values or who mostly contain a single value (very low information density)
9. Eliminate redundant columns and columns that are weighted sums of others.
10. Search for outliers and initial search for missing values
11. Sanity check using Decision Tree (1 to 2 splits)

# Missing Data Steps

We join this program already in progress> DF = **MDdf**

1. Identify missing data
2. “Mark” missing data
3. Clean up any obvious mistakes (missing = 0)
4. Make easy decisions on Rows/Columns
5. Assess missingness pattern
6. Apply simple techniques
7. Apply complex techniques (MICE)
8. Sensitivity analysis
9. Report process
10. Export “Draft Modeling File”

# Step 1: Identify missing data

- Pretty much already done by now
  - Nice table to summarize
  - Will look slightly different than first look do to elimination of rows/columns

**NOTE:** `is.na()`, `flextable`, `md.pattern()`, `md.pareto()`

## Step 2: “Mark” missing data (with 0,1)

- Variation among the experts on importance (book doesn't seem to make important). I think it is.
  - Why is this important?
- Line or two of code, various functions
- How are you going to store the identifiers?

## Step 3: Clean up any obvious mistakes (missing = 0)

- May have done in Cleaning steps
  - Does U = unknown, thus missing?
  - Does blank = 0?

## Step 4: Make easy decisions on Rows/Columns

- If you haven't already, mark for elimination various rows and columns
  - Do you want to “delete” the column or just not use?
    - Could be helpful for imputation?

# Now the Hard Work Begins

- P.S. Don't forget Continuous Improvement

## Step 5: Assess missingness pattern

- We now must determine MCAR/MAR
  - Little has a method, some things I read weren't complimentary
  - I choose to do the tests manually
    - Learn more about the data
    - There is an increased Type I error rate, but the consequences are low
  - Logistic regression, Chi Square and other methods can be used
  - I find a Decision Tree to be just fine and nice visually
    - I used judgement, not statistical significance
  - Again consequences are low for mistake because ....
    - If we use more advanced methods, they assume MAR and thus handle MCAR just fine

## Step 6: Apply simple techniques

- These are univariate (one variable at a time) methods
  - Check on which methods work only for MCAR
  - First check .... “small” number of missing data
    - This gets that variable “out of the way” for Multiple Imputation
  - Second check could be variables with more missing, but SME indicates univariate imputation should be fine
    - Need there not to be “many” missing values in the variable used to impute if a modeling method is used

## Step 6: Apply simple techniques (cont.)

- Non “modeling” methods
  - Mean, median, mode
  - Simple stochastic
  - LOCF for time series
- Modeling methods
  - Any number of regression/ML techniques
    - Choose based on understand of variables and class of variable being imputed
    - Can be point estimate or stochastic
- MICE package has options, but coding yourself will create understanding. It's not hard.

## Step 7: Multiple Imputation

- Of MICE and Business Analytics Professionals
- First and most different aspect
  - We create  $m$  datasets, not one dataset
  - Each dataset has the missing data imputed “differently”
- We model on each dataset
- We don’t “average” the datasets, we “average” the models

## Step 7: Multiple Imputation – Some Terminology

- $m$ : the number of imputed datasets. For this class,  $m=10$ .
    - Some theory can be used to determine  $m$ , but practical implications probably rule unless totally automated
  - Works for **MCAR** and **MAR**
  - I'm ignoring ignorability (MNAR?)
- 
- $\hat{Q}$ : quantity(ies) of interest, our parameters
  - $Q$ : our sample estimate of  $Q$
  - **Unbiased**: All estimates are wrong, but they can be right “on average”
  - **Confidence Valid**: “conservative” on variability and **df** estimates
    - **???? What does this mean????**

## Step 7: Multiple Imputation – More Terminology

- **U:** Variance or covariance matrix (sampling variation)
  - **B:** Variance caused by missing data
  - **B/m:** Variance caused by mputing the missing data
  - **T:** Total variance of  $\hat{Q}$
- 
- **Equation 2.20:**  $T = U + B + B/m$
- 
- $\lambda$  = % of variation due to missingness  
 $= (B + B/m)/T$

# Step 7: Multiple Imputation - Patterns

- **Univariate and multivariate**
  - Don't need MICE for univariate
- **Monotone or general**
  - Monotone is a unique situation not covered in this class
- **Connected or unconnected**
  - ???? No complete overlap of missingness????

## Step 7: Multiple Imputation – “some” challenges (pgs. 111-112)

- Predictors of missing values can have missing values
- Circular dependence. I need  $x_2$  to predict  $x_2$  but  $x_2$  has missingness and I need  $x_3$  to predict  $x_2$  but it has missingness
- Different types of variables (categorical, quantitative, etc)
- Imputation can create impossible combinations of variables

## Step 7: Multiple Imputation: 2 of 3 methods

- I skipped **monotonic**
- **Joint Modeling:** imputations drawn from a multivariate model
- **Fully Conditional Specification:** aka *chained equations*. Iterated conditional models.
  - MICE is a Fully Conditional model.

# Of MICE and Imputation



# Toy Data

MICE - JMP Pro

File Edit Tables Rows Cols DOE Analyze Graph Tools View Window Help

MICE

x2 vs. X1  
x2 vs. X12  
Fit Leas...ares X1

Σ X1 x2 x3 x4 x1Miss x2Miss Pred Formula X1

Columns (7/0)

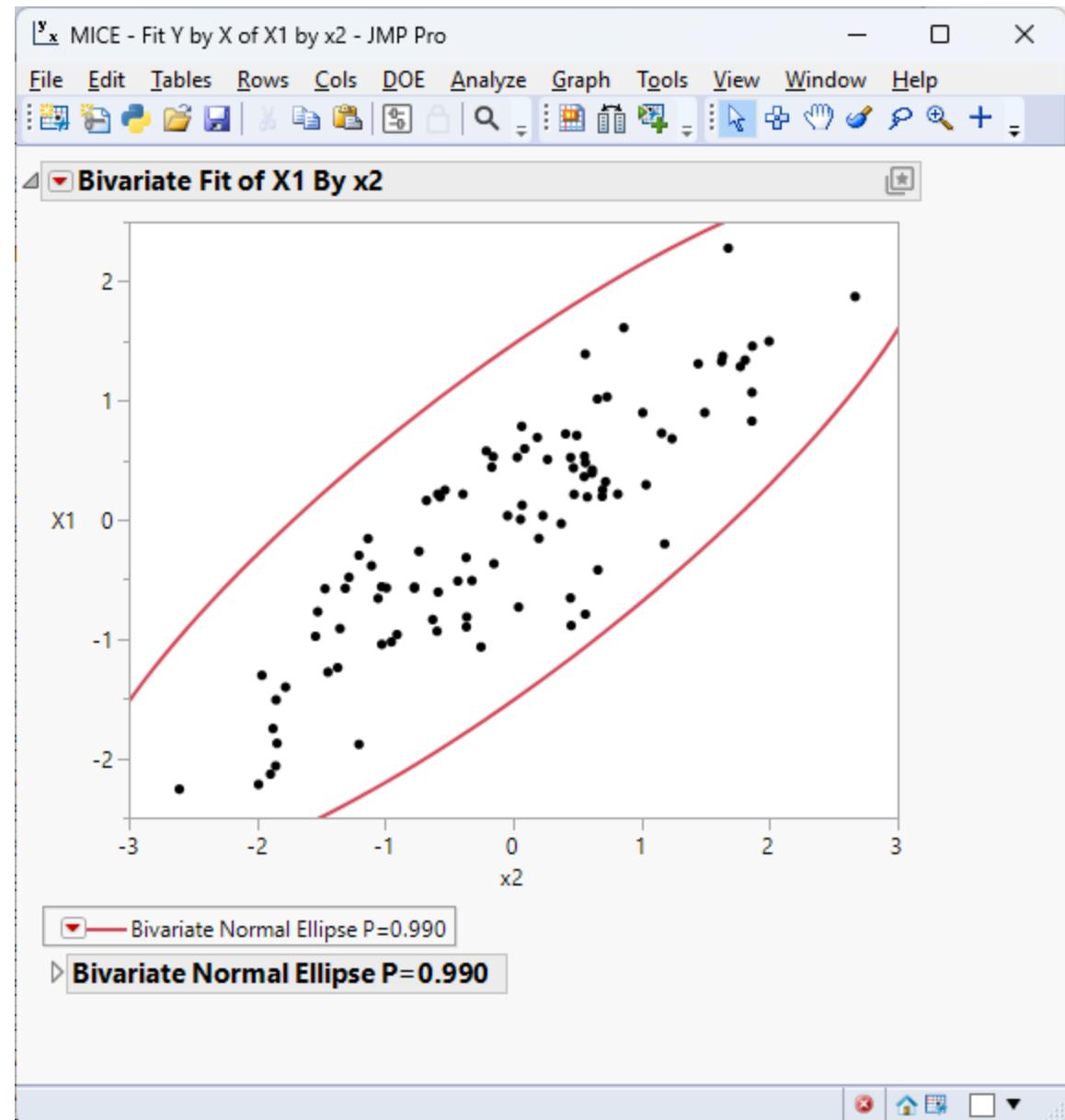
X1  
x2  
x3  
x4  
x1Miss  
x2Miss  
Pred Formula X1

Rows

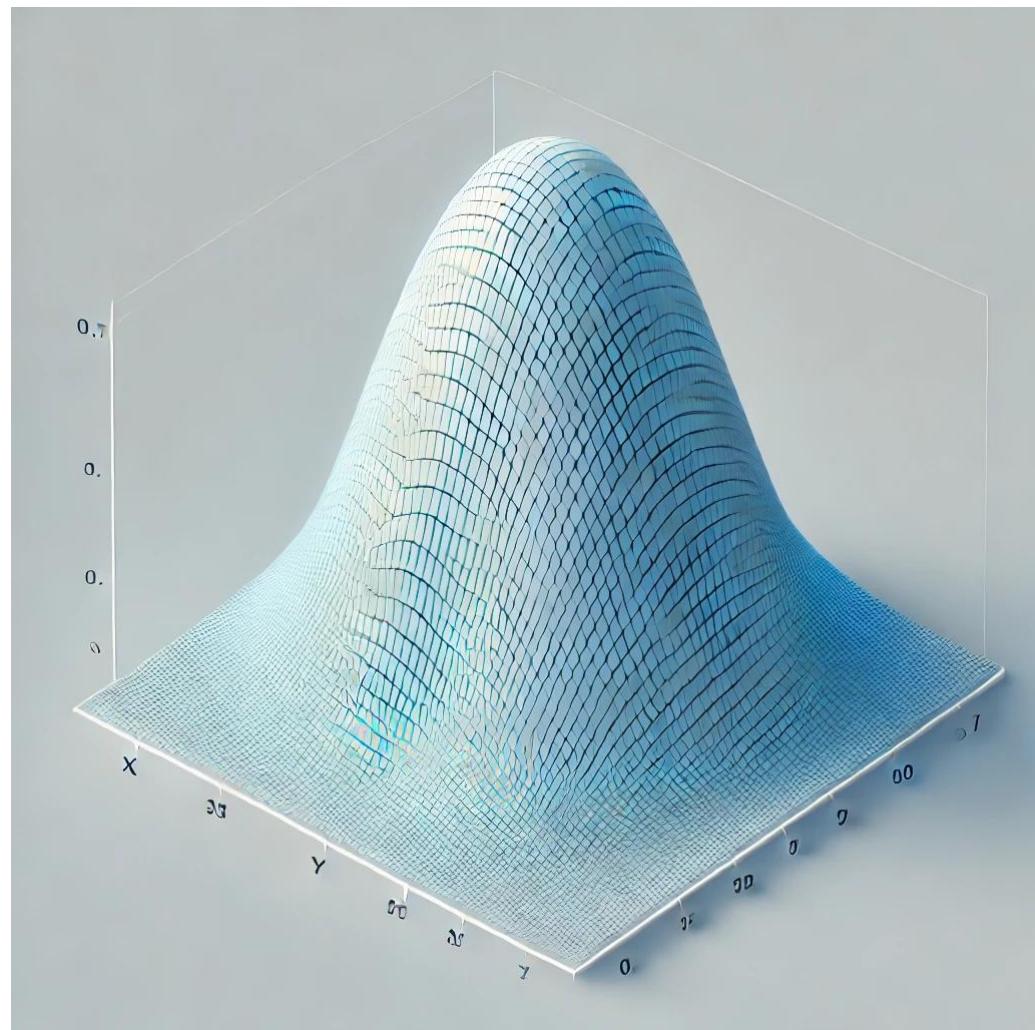
All rows 100  
Selected 0  
Excluded 0  
Hidden 0  
Labeled 0

	X1	x2	x3	x4	x1Miss	x2Miss	Pred Formula X1
61	0.7272017785	1.1573709454	0.6679870034	-0.540779478	0.7272017785	1.1573709454	0.8082532783
62	1.873062272	2.6671957865	5.534167169	-1.476006661	1.873062272	2.6671957865	2.1180040688
63	0.362187858	0.5530526317	1.9404688946	-0.438578612	0.362187858	0.5530526317	0.4856805818
64	0.7079389715	0.4950434086	0.1688987228	-0.212359795	0.7079389715	0.4950434086	0.3338596486
65	0.6911339283	0.1872359357	0.8398205008	-0.804899535	0.6911339283		• 0.1967860698
66	1.6123075691	0.8616579201	4.6400347092	-0.022844443	1.6123075691	0.8616579201	0.8300132254
67	-0.26451626	-0.736471388	1.9942436839	0.9615207666	-0.26451626	-0.736471388	-0.407741028
68	0.4414984327	-0.168029756	-1.620869131	-0.055144926	0.4414984327	-0.168029756	-0.21059638
69	0.0347608012	0.2310397684	0.2361824848	-0.297433886	0.0347608012	0.2310397684	0.1692060513
70	0.215246471	-0.589439247	0.3641124231	0.7865467001		•	• -0.401449995
71	0.0345768541	-0.044854159	-1.479975537	-1.283946011	0.0345768541	-0.044854159	-0.072644399
72	-1.278169474	-1.446657655	-4.546570219	-1.712052972	-1.278169474	-1.446657655	-1.150670632
73	0.2136152269	0.4741926496	-1.711273961	-2.073650463	0.2136152269	0.4741926496	0.2836278894
74	-2.13399653	-1.894927954	-7.569148448	0.7937281643		• -1.894927954	-1.722444719
75	0.252645175	0.6967548423	-0.75301626	-1.500994385	0.252645175	0.6967548423	0.4624440721
76	-1.304739384	-1.962194742	-2.009920975	2.4230867402	-1.304739384	-1.962194742	-1.502410908
77	0.5963860424	0.0879066476	2.5171918432	-0.428373401	0.5963860424	0.0879066476	0.2162901648
78	1.0690242008	1.8623950506	4.9447965689	4.5744941425	1.0690242008	1.8623950506	1.3156922358
79	0.720939421	0.4076529632	4.3316385568	1.1262439694	0.720939421	0.4076529632	0.469734394
80	0.4787819311	0.5625894477	0.0570461815	-0.104178261	0.4787819311	0.5625894477	0.3669118893
81	1.0135998514	0.6563030981	1.7133748357	-1.738420238		• 0.6563030981	0.5917024322
82	0.5297942641	-0.158066331	3.6053546042	1.4241279809	0.5297942641	-0.158066331	0.0460701274
83	-0.513497407	-0.432745692	-1.414799955	-0.080705171	-0.513497407	-0.432745692	-0.369882885
84	-0.606029566	-0.586889507	-5.855367831	-2.676429943	-0.606029566	-0.586889507	-0.629156278
85	0.5341635487	0.5555881235	-1.755711261	0.161488327	0.5341635487	0.5555881235	0.2443101354
86	-0.202045345	1.1804935705	0.4333079884	0.7644207216	-0.202045345	1.1804935705	0.7570233751
87	0.8286848336	1.8617162633	-4.388268511	-1.663656651	0.8286848336	1.8617162633	1.0128700531
88	0.6801946024	1.2386951943	-0.283886185	-1.135556994	0.6801946024	1.2386951943	0.8287498006
89	0.5252333122	0.0306316307	2.9777063536	-0.549156431		•	• 0.2110895551
90	-0.385805731	-1.106393348	3.9950972146	1.6284968777	-0.385805731	-1.106393348	-0.557085327
91	0.5064000769	0.2667748658	2.2196037491	0.0859767262	0.5064000769	0.2667748658	0.2945848262
92	0.2493437452	-0.533424837	-1.983690803	-1.111261634	0.2493437452	-0.533424837	-0.427820499
93	-2.260300344	-2.606288707	-4.04807963	-0.108179692	-2.260300344		• -1.9411279644

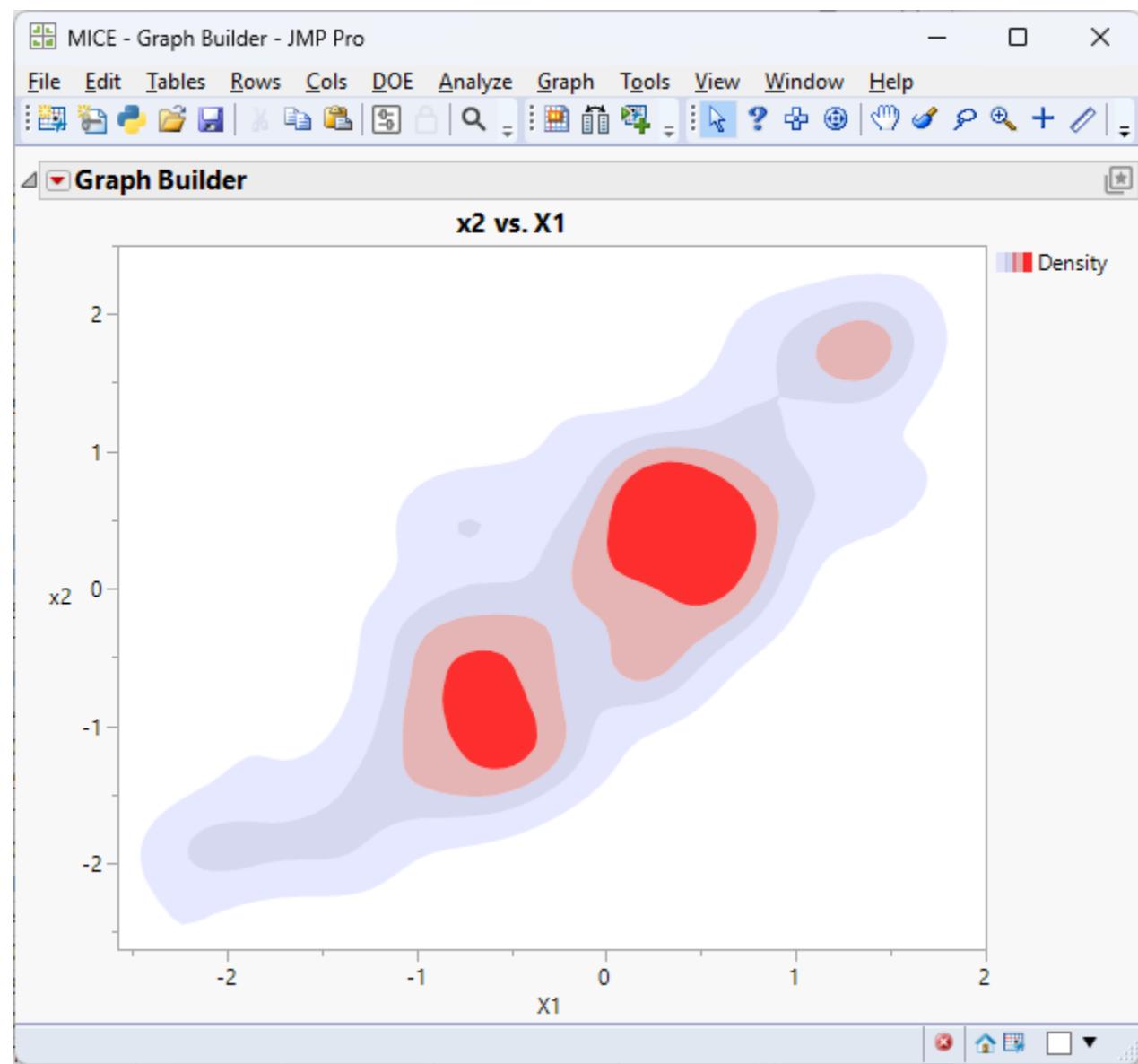
# Relationship of x1 and x2



# 3D Relationship



# 3Dish Relationship

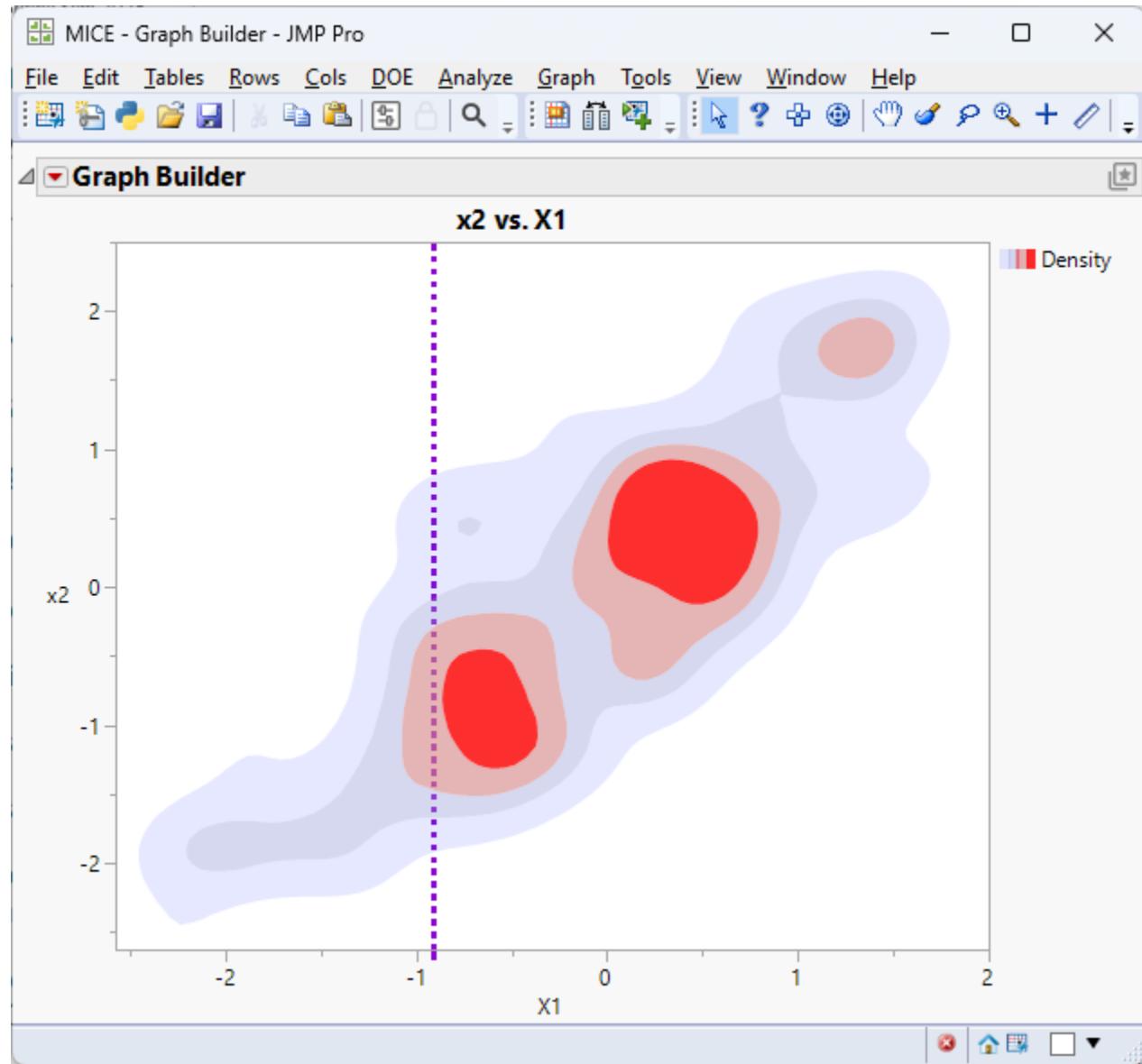


# How Do I Impute This Number?

MICE - JMP Pro

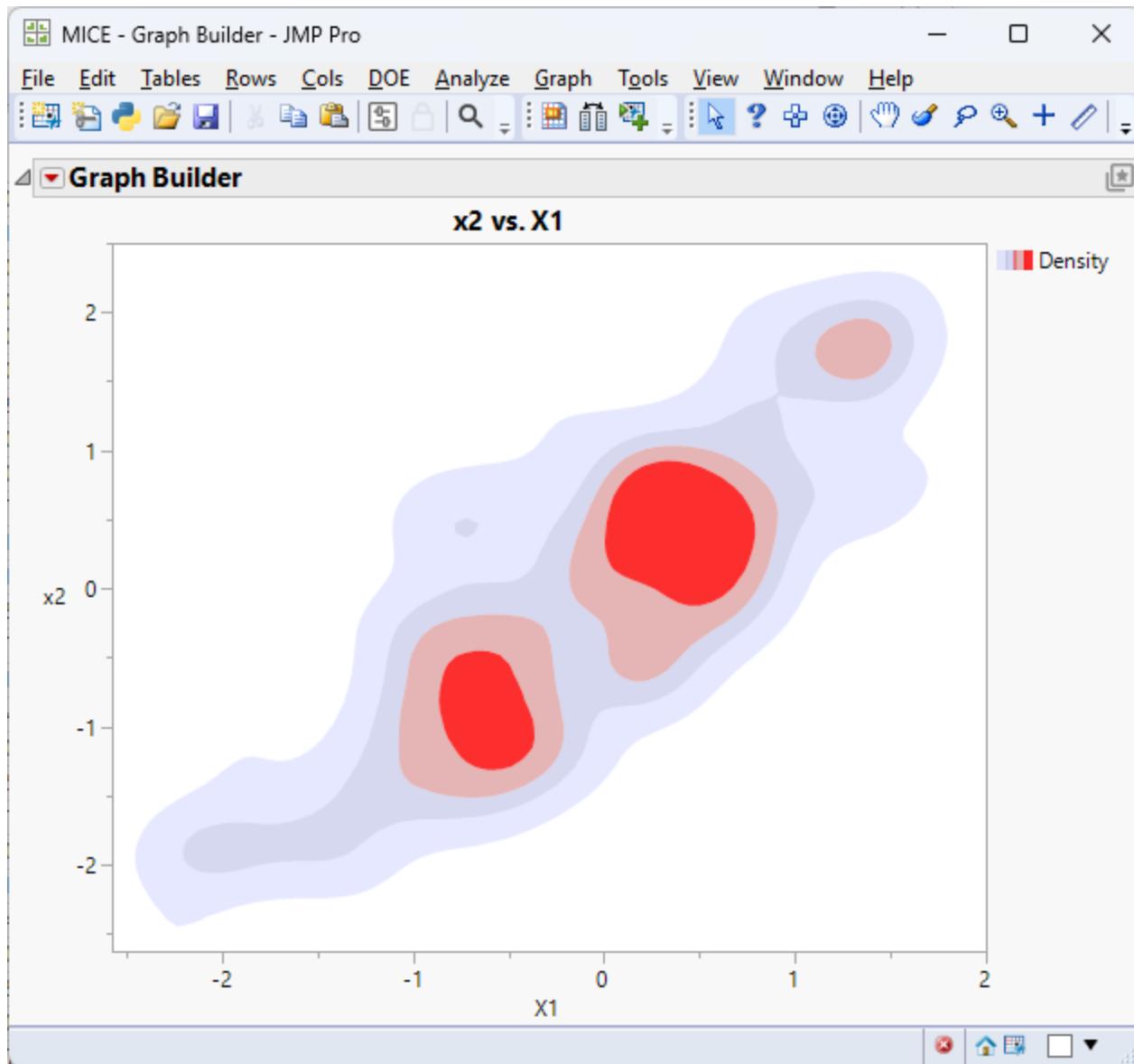
		X1	x2	x3	x4	x1Miss	x2Miss	Pred Formula X1
MICE		Σ						
	7	-1.875771888	-1.844187534	-3.911632873	-1.582168143	-1.875771888	-1.844187534	-1.377351327
	8	0.2940739505	1.0360395225	-2.500799789	-1.562626033	0.2940739505	1.0360395225	0.5825093523
	9	-0.962607894	-0.907714213	1.386936417	0.5842378207	-0.962607894	-0.907714213	-0.540167694
	10	-0.157661354	-1.13407479	3.3617583804	1.2161174692	-0.157661354	-1.13407479	-0.596084233
	11	0.8991945696	1.4935533978	1.2891103198	0.1165326832	0.8991945696	1.4935533978	1.0377493625
	12	-0.57144718	-0.989872076	-1.160148781	-0.72499188	-0.57144718	-0.989872076	-0.692019742
	13	0.0036156936	0.0559100549	4.11119212095	1.1147598139	0.0036156936	0.0559100549	0.2279460342
	14	1.4572282498	1.8654304398	4.3437006548	0.3806511958	1.4572282498	1.8654304398	1.4504325913
	15	-1.511135948	-1.851556401	-3.492771433	1.5369083417	-1.511135948	-1.851556401	-1.482562338
	16	-0.934685402	-0.595085757	-1.943895731	-0.883950235	-0.934685402	-0.595085757	-0.474774177
	17	-0.573303294	-0.77318751	2.9201108419	-0.344438661	-0.573303294	-0.77318751	-0.324399163
	18	-1.404028278	-1.777665327	-1.383141516	-0.257580876	-1.404028278	-1.777665327	-1.237407122
	19	0.196298151	0.6935359169	-4.274107423	-1.995250369	0.196298151	0.6935359169	0.2716270617
	20	-0.655865589	0.4459430932	1.0953033266	-0.372370737	-0.655865589	0.4459430932	0.3631570871
	21	-0.912617587	-1.35291902	-4.019098498	-0.740854368	-0.912617587	*	-1.097326589
	22	-0.57801132	-1.469261722	-4.716527149	-1.476935089	-0.57801132	-1.469261722	1.184907941
	23	1.3283066132	1.6252149812	-1.082249987	-0.033634532	1.3283066132	1.6252149812	0.9891227651
	24	-0.31667272	-0.368323315	-5.53255853	-1.265584296	-0.31667272	-0.368323315	-0.524233303
	25	0.7842094763	0.0655364549	0.3612746405	0.669735087	*	0.0655364549	0.0299269856
	26	-0.574659601	-1.311549415	-2.071828849	-0.723595596	-0.574659601	-1.311549415	-0.955717515
	27	-0.298685352	-1.20458271	-3.322632388	-2.327673521	-0.298685352	-1.20458271	-0.89569632
	28	0.4178181662	0.616409002	-4.420519361	-1.397380378	0.4178181662	0.616409002	0.1886881037
	29	-0.898500438	-0.366645267	4.3723279809	2.2028349436	-0.898500438	-0.366645267	-0.075695854
	30	1.3733509344	1.6339138702	0.1266388403	-0.557532439	1.3733509344	1.6339138702	1.087429408
	31	0.1621934884	-0.677524615	-2.640331735	-0.670505546	0.1621934884	-0.677524615	-0.578320907
	32	1.031396461	0.7310650814	5.0925069798	2.7184256072	1.031396461	0.7310650814	0.6616539548
	33	-0.482411701	-1.282850121	-3.246346887	0.9674722512	-0.482411701	-1.282850121	-1.074472749
	34	-0.7924031	0.5618132313	-1.450183092	-1.168736894	-0.7924031	0.5618132313	0.3198675098
	35	-0.887151427	0.4516594631	-0.624417921	-0.32882548	-0.887151427	0.4516594631	0.2632716363
	36	1.4980775314	1.9961111225	0.8166085684	1.4884632623	1.4980775314	1.9961111225	1.2822073751
	37	1.2860338183	1.7716953969	3.0901277232	0.9534897293	1.2860338183	1.7716953969	1.2920953569

# Conditional Distribution



# What if Both Numbers are Missing

	X1	x2	x3	x4	x1Miss	x2Miss	P
Σ							
52	-0.368649873	-0.151443341	-0.456782379	0.491750146	-0.3686498...	-0.151443341	
53	0.1917255859	0.577267611	4.9099022224	0.2172522187	0.19172558...	0.577267611	
54	-0.732347358	0.040813987	-2.49870863	-2.654524439	-0.7323473...	0.040813987	
55	-2.22052147	-1.988689458	-1.699800468	-1.846142745	-2.22052147	-1.988689458	
56	-0.771056244	-1.526287378	-1.765647961	-2.115228462	-0.7710562...	-1.526287378	
57	-1.884214243	-1.204406795	-2.027651774	0.2337723375		• -1.204406795	
58	-1.044849697	-1.026338461	2.3764153315	0.6180031411	-1.0448496...	-1.026338461	
59	0.3910286738	0.6142195417	2.7774536408	0.9993584634	0.39102867...	0.6142195417	
60	0.1917306275	-0.569918934	0.9629400972	0.046644882	0.19173062...	-0.569918934	
61	0.7272017785	1.1573709454	0.6679870034	-0.540779478	0.72720177...	1.1573709454	
62	1.873062272	2.6671957865	5.534167169	-1.476006661	1.873062272	2.6671957865	
63	0.362187858	0.5530526317	1.9404688946	-0.438578612	0.362187858	0.5530526317	
64	0.7079389715	0.4950434086	0.1688987228	-0.212359795	0.70793897...	0.4950434086	
65	0.6911339283	0.1872359357	0.8398205008	-0.804899535	0.69113392...		•
66	1.6123075691	0.8616579201	4.6400347092	-0.022844443	1.61230756...	0.8616579201	
67	-0.26451626	-0.736471388	1.9942436839	0.9615207666	-0.26451626	-0.736471388	
68	0.4414984327	-0.168029756	-1.620869131	-0.055144926	0.44149843...	-0.168029756	
69	0.0347608012	0.2310397684	0.2361824848	-0.297433886	0.03476080...	0.2310397684	
70	0.215246471	-0.589439247	0.3641124231	0.7865467001		•	•
71	0.0345768541	-0.044854159	-1.479975537	-1.283946011	0.03457685...	-0.044854159	
72	-1.278169474	-1.446657655	-4.546570219	-1.712052972	-1.2781694...	-1.446657655	
73	0.2136152269	0.4741926496	-1.711273961	-2.073650463	0.21361522...	0.4741926496	
74	-2.13399653	-1.894927954	-7.569148448	0.7937281643		• -1.894927954	
75	0.252645175	0.6967548423	-0.75301626	-1.500994385	0.252645175	0.6967548423	
76	-1.304739384	-1.962194742	-2.009920975	2.4230867402	-1.3047393...	-1.962194742	
77	0.5963860424	0.0879066476	2.5171918432	-0.428373401	0.59638604...	0.0879066476	
78	1.0690242008	1.8623950506	4.9447965689	4.5744941425	1.06902420...	1.8623950506	
79	0.720939421	0.4076529632	4.3316385568	1.1262439694	0.720939421	0.4076529632	
80	0.4787819311	0.5625894477	0.0570461815	-0.104178261	0.47878193...	0.5625894477	
81	1.0135998514	0.6563030981	1.7133748357	-1.738420238		• 0.6563030981	



# How about just a conditional model?

The screenshot shows the SPSS MICE interface. In the center, there is a text input field containing a mathematical expression representing a conditional model:

$$-0.007295732 + 0.6517120247 \cdot x_2 + 0.0592321033 \cdot x_3 + -0.040146074 \cdot x_4$$

This expression is highlighted with a blue border. Below the input field, there is a "Preview" button. At the bottom of the dialog, there are "OK", "Cancel", "Apply", and "Help" buttons.

x1Miss	x2Miss	Pred Formula X1
-1.875771888	-1.844187534	-1.377351327
0.2940739505	1.0360395225	0.5825093523
-0.962607894	-0.907714213	-0.540167694
-0.157661354	-1.13407479	-0.596084233
0.8991945696	1.4935533978	1.0377493625
-0.57144718	-0.989872076	-0.692019742
0.0036156936	0.0559100549	0.2279460342
1.4572282498	1.8654304398	1.4504325913
-1.511135948	-1.851556401	-1.482562338
-0.934685402	-0.595085757	-0.474774177
-0.573303294	-0.77318751	-0.324399163
-1.404028278	-1.777665327	-1.237407122
0.196298151	0.6935359169	0.2716270617
-0.655865589	0.4459430932	0.3631570871
-0.912617587	•	-1.097326589
-0.57801132	-1.469261722	-1.184907941
1.3283066132	1.6252149812	0.9891227651
-0.31667272	-0.368323315	-0.524233303
	• 0.0655364549	0.0299269856
-0.574659601	-1.311549415	-0.955717515
-0.298685352	-1.20458271	-0.89569632
0.4178181662	0.616409002	0.1886881037
-0.898500438	-0.366645267	-0.075695854
1.3733509344	1.6339138702	1.087429408
0.1621024824	-0.677524615	-0.578320007