



# Understand Data Quickly & Easily



## Understanding Data Quickly and Easily: Agenda

- Case Study – Understanding Data in ~ 270 seconds
- General Overview of IBM SPSS Statistics Standard
- To Understand Data Quickly and Easily
  - Save Time Getting to the Data
  - Leverage Metadata
  - Start with a Picture
  - Use Words and Code
  - Quick Access to Analytics
  - Automate Tasks
- Case Study – Who Will Churn and When?
- Recap & Questions

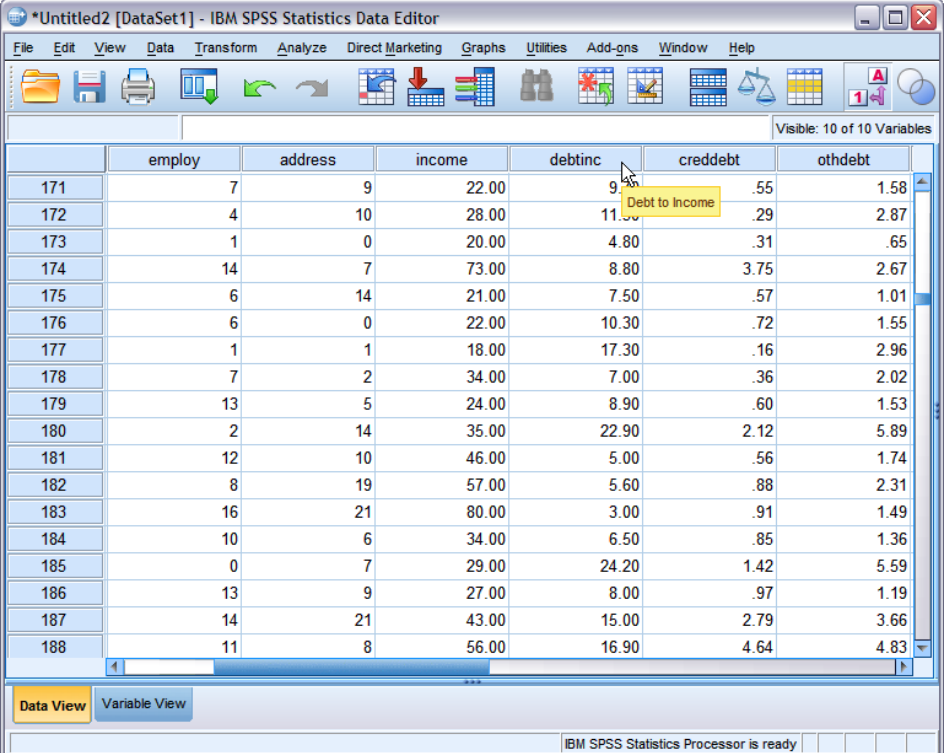
# Understanding Data Quickly and Easily: A Case Study: Analysis in 270 Seconds

## ■ Motivating Question

- Based on the data from my branch, is age or income level a risk indicator at my bank as it pertains to our loans? If not, what are the top two or three attributes about my customers that would help understand (and predict) default on a loan?

## ■ Inputs

- Age
- Level of Education
- Years at Current Employer
- Years at Current Address
- Income
- Debt to Income Ratio
- Credit Card Debt
- Other Debt (e.g. Auto Loan)



\*Untitled2 [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

Visible: 10 of 10 Variables

	employ	address	income	debtinc	creddebt	othdebt
171	7	9	22.00	9.50	.55	1.58
172	4	10	28.00	11.50	.29	2.87
173	1	0	20.00	4.80	.31	.65
174	14	7	73.00	8.80	3.75	2.67
175	6	14	21.00	7.50	.57	1.01
176	6	0	22.00	10.30	.72	1.55
177	1	1	18.00	17.30	.16	2.96
178	7	2	34.00	7.00	.36	2.02
179	13	5	24.00	8.90	.60	1.53
180	2	14	35.00	22.90	2.12	5.89
181	12	10	46.00	5.00	.56	1.74
182	8	19	57.00	5.60	.88	2.31
183	16	21	80.00	3.00	.91	1.49
184	10	6	34.00	6.50	.85	1.36
185	0	7	29.00	24.20	1.42	5.59
186	13	9	27.00	8.00	.97	1.19
187	14	21	43.00	15.00	2.79	3.66
188	11	8	56.00	16.90	4.64	4.83

Data View Variable View

IBM SPSS Statistics Processor is ready

# Understanding Data Quickly and Easily: A Case Study: Results

- In 270 seconds (i.e. under five minutes) we were able to:
  - Determine, in order of importance, that: debt to income ratio, years with current employer, credit card debt, and years at current address are the indicators of risk as it relates to default on a loan *and based on our data*

– Generate an analysis that has 80%+ accuracy

- Provide predictive default scores that can be used to manage risk *based on our data*

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 <sup>a</sup>	debtinc	.121	.017	52.676	1	.000	1.129
	Constant	-2.476	.230	116.315	1	.000	.084
Step 2 <sup>b</sup>	employ	-.140	.023	38.158	1	.000	.869
	debtinc	.134	.018	54.659	1	.000	1.143
Step 3 <sup>c</sup>	Constant	-1.621	.259	39.038	1	.000	.198
	employ	-.244	.033	54.676	1	.000	.783
Step 4 <sup>d</sup>	debtinc	.069	.022	9.809	1	.002	1.072
	creddebt	.506	.101	25.127	1	.000	1.658
	Constant	-1.058	.280	14.249	1	.000	.347
	employ	-.247	.034	51.826	1	.000	.781
	address	-.089	.023	15.109	1	.000	.915
	debtinc	.072	.023	10.040	1	.002	1.074
	creddebt	.602	.111	29.606	1	.000	1.826
	Constant	-.605	.301	4.034	1	.045	.546

a. Variable(s) entered on step 1: debtinc.  
 b. Variable(s) entered on step 2: employ.  
 c. Variable(s) entered on step 3: creddebt.  
 d. Variable(s) entered on step 4: address

Classification Table<sup>c</sup>

	Observed	Predicted					
		Selected Cases <sup>a</sup>			Unselected Cases <sup>b</sup>		
		default		Percentage Correct	default		Percentage Correct
0	1	0	1				
Step 1	default 0	361	14	96.3	137	5	96.5
	1	100	24	19.4	45	14	23.7
	Overall Percentage			77.2			75.1
Step 2	default 0	351	24	93.6	136	6	95.8
	1	80	44	35.5	36	23	39.0
	Overall Percentage			79.2			79.1
Step 3	default 0	348	27	92.8	135	7	95.1
	1	72	52	41.9	28	31	52.5
	Overall Percentage			80.2			82.6
Step 4	default 0	352	23	93.9	130	12	91.5
	1	67	57	46.0	27	32	54.2
	Overall Percentage			82.0			80.6

a. Selected cases validate EQ 1  
 b. Unselected cases validate NE 1  
 c. The cut value is .500

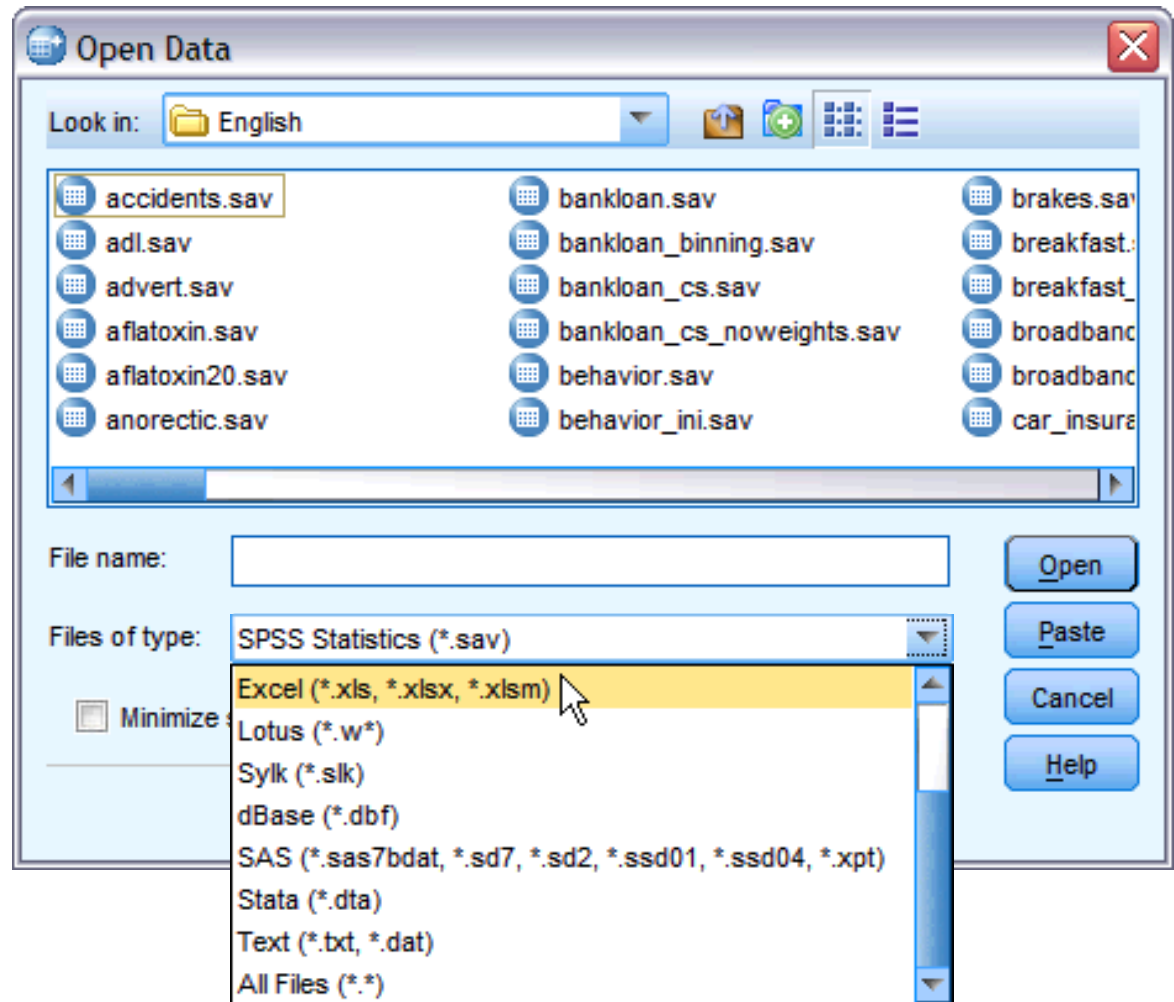
# Understanding Data Quickly and Easily: A General Overview of IBM SPSS Statistics Standard

- Spreadsheet-like look and feel
- General environment for predictive analytics and statistical analysis
- Well-suited for ad-hoc analysis and hypothesis testing
  - Core descriptive statistical capabilities
  - Advanced statistical functions
  - Many types of regression
  - Tabular analysis & output

	patid	physid	age	agecat	gender	active	obesity	diabetes
14	5357069859	822229	81	75+	Male	No	Yes	No
15	5132742071	297378		45-54	Male	Yes	Yes	Yes
16	2660586207	297378	53	45-54	Male	No	Yes	No
17	5408312498	799998	62	55-64	Female	No	No	No
18	9069087682	297378	64	55-64	Male	Yes	No	No
19	8173197592	799998	58	55-64	Female	No	No	No
20	8808732689	822229	83	75+	Male	Yes	No	No
21	5666440246	822229	67	65-74	Female	Yes	Yes	No
22	8483212117	822229	59	55-64	Male	No	No	No
23	0674952107	297378	60	55-64	Male	Yes	No	No
24	9726045305	799998	71	65-74	Female	No	No	No
25	6197593400	297378	52	45-54	Male	Yes	No	No
26	2088521482	799998	52	45-54	Male	No	Yes	No
27	4775117168	822229	54	45-54	Male	Yes	No	No

## Understanding Data Quickly and Easily: Saving Time during Data Access Leads to More Time for Analyses

- IBM SPSS Statistics Standard can open (and export to) a variety of data formats
  - .xls, .csv, .sas, .txt
- Enables analysts to tap directly into databases



## Understanding Data Quickly and Easily: Leveraging Metadata Enables More Extensive Analytics

- The IBM SPSS Statistics Standard interface provides easy access to metadata – information about the data that accelerates and sharpens analyses

bankloan.sav [DataSet2] - IBM SPSS Statistics Data Editor

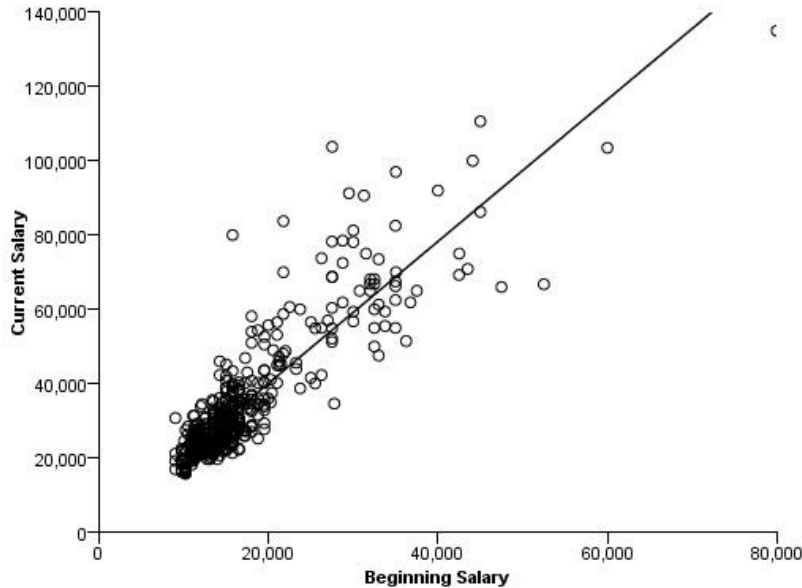
	Name	Type	Width	Decim...	Label	Values	Missing	Columns	Align	Measure	Role
1	age	Numeric	4	0	Age in years	None	None	4	Right	Scale	Input
2	ed	Numeric	4	0	Level of education	{1, Did not c...	None	15	Right	Ordinal	Input
3	employ	Numeric	4	0	Years with current employer	None	None	6	Right	Scale	Input
4	address	Numeric	4	0	Years at current address	None	None	7	Right	Scale	Input
5	income	Numeric	8	2	Household income in thousands	None	None	8	Right	Scale	Input
6	debtinc	Numeric	8	2	Debt to income ratio (x100)	None	None	8	Right	Scale	Input
7	creddebt	Numeric	8	2	Credit card debt in thousands	None	None	8	Right	Scale	Input
8	othdebt	Numeric	8	2	Other debt in thousands	None	None	8	Right	Scale	Input
9	default	Numeric	4	0	Previously defaulted	{0, No}...	None	7	Right	Nominal	Target
10	preddef1	Numeric	11	5	Predicted default, model 1	None	None	11	Right	Scale	None
11	preddef2	Numeric	11	5	Predicted default, model 2	None	None	11	Right	Scale	None
12	preddef3	Numeric	11	5	Predicted default, model 3	None	None	11	Right	Scale	None
13											

Data View Variable View

IBM SPSS Statistics Processor is ready

# Understanding Data Quickly and Easily: Starting with a Picture Provides “Holistic” Data Understanding

- The IBM SPSS Statistics Standard interface enables a quick dragging and dropping with palettes, boxes, menus, and variable names



The screenshot shows the 'Chart Builder' window in IBM SPSS. The 'Variables' list on the left includes 'Age in years [age]', 'Level of education [...]', 'Years with current ...', 'Years at current ad...', 'Household income i...', 'Debt to income ratio...', 'Credit card debt in t...', 'Other debt in thous...', 'Previously defaulte...', 'Predicted default, m...', and 'Predicted default, m...'. The 'Previously defaulte...' variable is selected and placed in the chart area. The chart preview shows a stacked bar chart with two categories: 'No' and 'Yes'. The Y-axis is labeled 'Count'. The bars are stacked with three colors: tan at the bottom, green in the middle, and blue at the top. A dashed box highlights the 'Stack: set color' option. Below the chart, there are buttons for 'Element Properties...' and 'Options...'. At the bottom of the window, there are buttons for 'OK', 'Paste', 'Reset', 'Cancel', and 'Help'.



## Understanding Data Quickly and Easily: Using Words and Code Help Analysts Identify Trends & KPIs

- IBM SPSS Statistics Standard easily supports the creation of new variables with words and variable names
  - Quickly introduce new variables and metrics using variable names and labels
- Variable names & labels also speed the sorting and filtering processes

The screenshot displays the IBM SPSS Statistics Data Editor interface. The main window shows a dataset named 'bankloan.sav [DataSet2]' with variables including 'age', 'debtinc', 'creddebt', and 'othdebt'. The 'Compute Variable...' menu option is highlighted in the 'Analyze' menu. A blue box highlights the 'Compute Variable' dialog box, which is used for creating new variables. The dialog box includes fields for 'Target Variable' and 'Numeric Expression', a list of variables to choose from, a function group list, and an 'If...' option for case selection.

The 'Compute Variable' dialog box is shown in the foreground, with the following fields and options:

- Target Variable:** (Empty field)
- Numeric Expression:** (Empty field)
- Type & Label:** (Buttons for Type and Label)
- Function group:** (List of function groups: All, Arithmetic, CDF & Noncentral CDF, Conversion, Current Date/Time, Date Arithmetic, Date Creation)
- Functions and Special Variables:** (Empty list)
- If... (optional case selection condition):** (Empty field)
- Buttons:** OK, Paste, Reset, Cancel, Help

# Understanding Data Quickly and Easily: Quick Access to Algorithms & Procedures Improves Analysis

- Menus provide quick access to and a logical organization of analytics
- Dialog boxes provide an intuitive set-up of analysis
- Both are configurable and both improve analysis by offering a choice of approach (and extensive help with that approach)

The screenshot displays the IBM SPSS Statistics Data Editor interface. The main window shows a data table with 12 rows and 4 columns: 'age', 'ed', 'income', and 'debtinc'. The 'Analyze' menu is open, and the 'Regression' submenu is selected, showing options like 'Automatic Linear Modeling...', 'Linear...', 'Curve Estimation...', 'Partial Least Squares...', 'Binary Logistic...', 'Multinomial Logistic...', and 'Ordinal...'. The 'Binary Logistic...' option is highlighted. Below the main window, the 'Logistic Regression' dialog box is open. It shows the 'Dependent' variable as 'Previously defaulted [default]', the 'Covariates' list containing 'income', 'debtinc', and 'Credit card debt in thousands...', and the 'Method' set to 'Forward: LR'. The 'Selection Variable' field is empty.

age	ed	income	debtinc	creddebt	othdebt
1	41	Som			
2	27	Did not comple			
3	40	Did not comple			
4	41	Colle			
5	24	High scho			
6	41	High scho			
7	39	Did not comple			
8	43	Did not comple			
9	24	Did not comple			
10	36	Did not comple			
11	27	Did not comple			
12	25	Did not comple			

## Understanding Data Quickly and Easily: Automating Tasks Helps Streamline Analysis

- Command Syntax, IBM SPSS Statistic Standard's scripting language, can be generated from every dialog box
- Enables the analytic professional (e.g. analyst, statistician) to focus on analysis and to automate routine tasks

The image illustrates the process of generating command syntax from SPSS dialog boxes. The main dialog box, 'Logistic Regression', shows the following settings:

- Dependent:** Previously defaulted [default]
- Covariates:** age, ed(Cat), employ, address, income, debtinc, creddebt, othdebt
- Method:** Forward: LR
- Selection Variable:** validate=1

The sub-dialogs shown are:

- Logistic Regression: Define Categorical Variables:** Shows 'ed(Indicator)' as a categorical covariate.
- Logistic Regression: Save:** Shows 'Probabilities' checked under 'Predicted Values' and 'Studentized' checked under 'Residuals'.
- Logistic Regression: Options:** Shows 'Classification plots' checked, 'Iteration history' checked, and 'Include constant in model' checked.

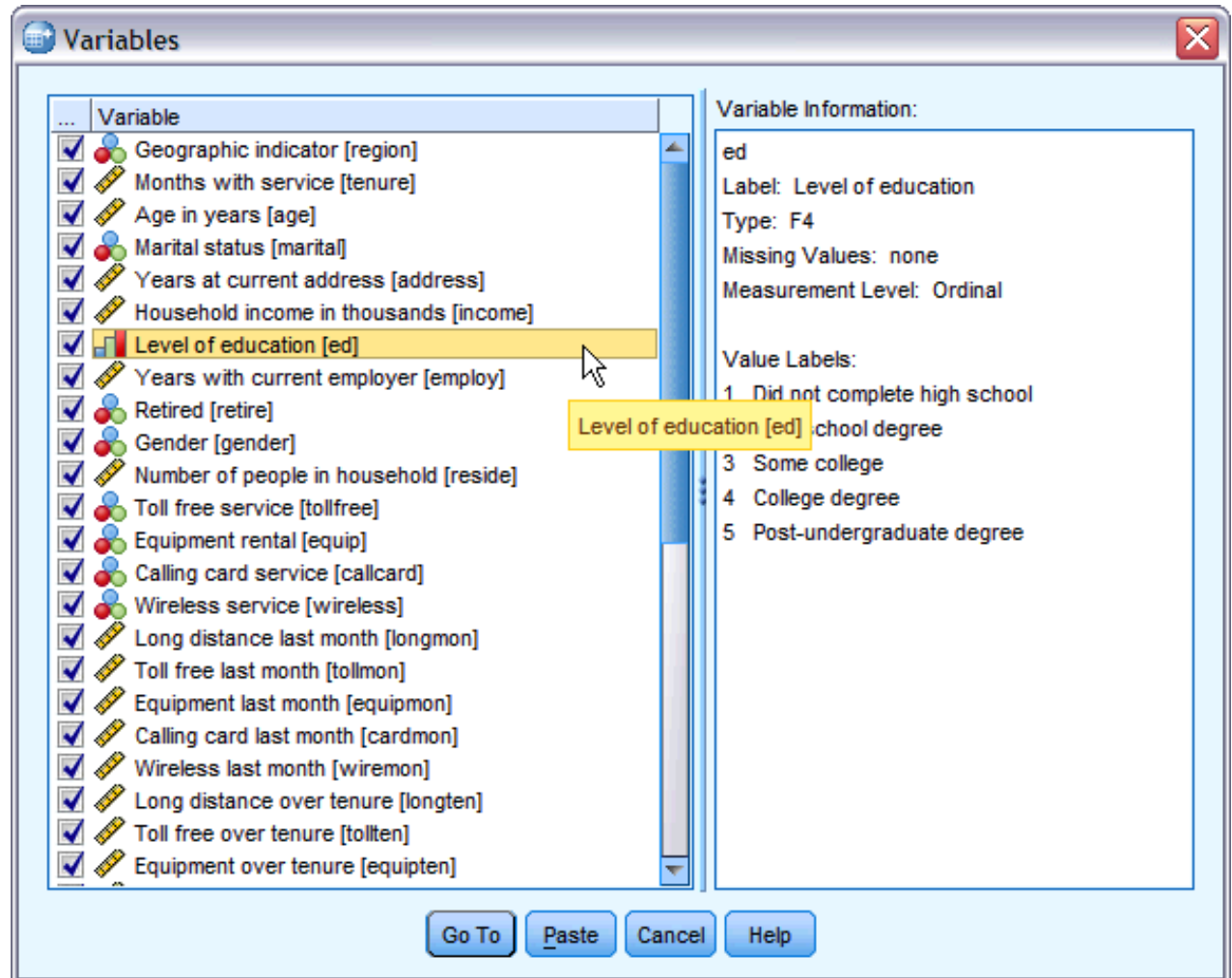
The command syntax window at the bottom right shows the following code:

```

1 DATASET ACTIVATE DataSet2.
2 LOGISTIC REGRESSION VARIABLES default
3 /SELECT=validate EQ 1
4 /METHOD=FSTEP(LR) age ed employ address income debtinc creddebt othdebt
5 /CONTRAST (ed)=Indicator
6 /SAVE=PRED COOK SRESID
7 /PRINT=GOODFIT
8 /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
9
10
  
```

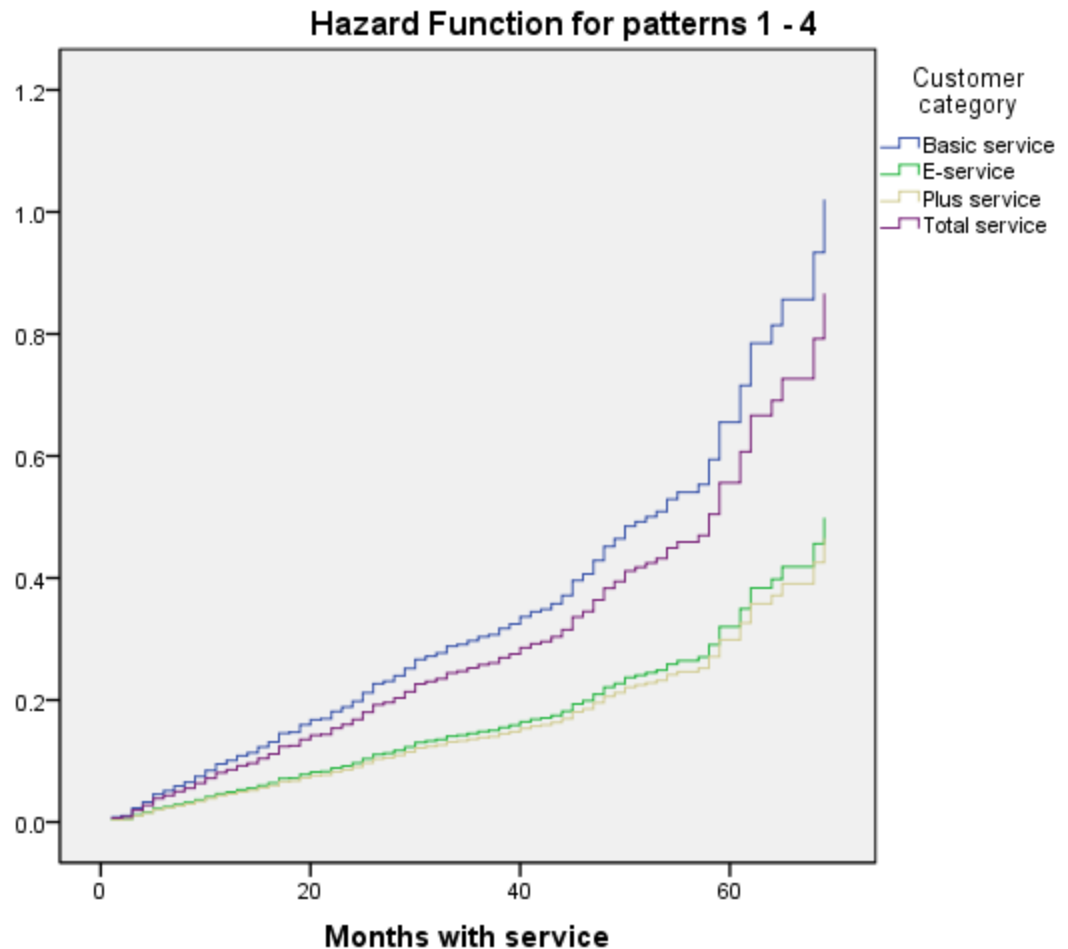
## Understanding Data Quickly and Easily: Who is Likely to Churn and When?

- Data from Telecommunications Service Provider
- Goal: Understand the Who and When of Customer Churn



## Understanding Data Quickly and Easily: Who is Likely to Churn and When – Results

- Customers with Basic Service and Total Service packages are more likely to defect and at a quicker pace than those with E-service and Plus Service packages
- The gap widens between these service levels as the number of months past, and the likelihood of churn appears to double at the two year mark *based on our data*



## Recap & Questions – NEXT STEPS

- Case Study – Understanding Data in ~ 270 seconds
- General Overview of IBM SPSS Statistics Standard
- To Understand Data Quickly and Easily
  - Save Time Getting to the Data
  - Leverage Metadata
  - Start with a Picture
  - Use Words and Code
  - Quick Access to Analytics
  - Automate Tasks
- Case Study – Who Will Churn and When?
- Recap & Questions
- For more information call (800) 543-2185