Introduction into SPSS

- **The objective of this deck is to provide you with a how-to-guide about the most common analyses you will likely conduct with SPSS.**

What this deck is	What this deck is NOT
An introduction to SPSS	An exhaustive guide to SPSS
A source of initial exploration for SPSS	A guide on advanced SPSS analyses
A how to guide to the most common analyses	A statistics-heavy deck
	An obsolete deck (please contribute your knowledge to the guide!)*

In addition you should use the SPSS online help function

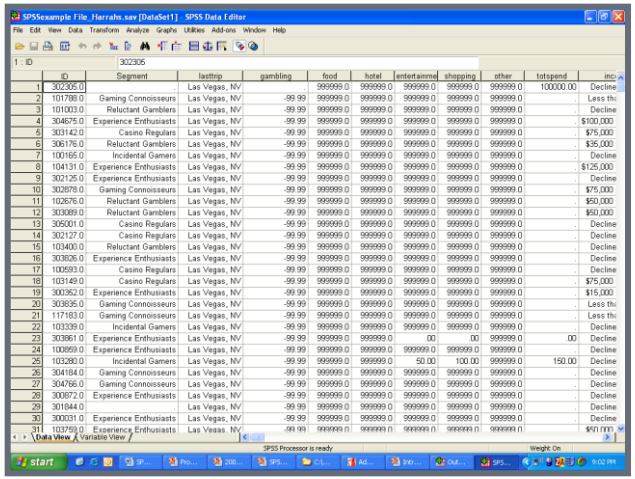
What is SPSS?

- **SPSS is a computer application that provides statistical analysis of data.**
  - It allows for in-depth data access and preparation, analytical reporting, graphics and modeling.
  - Its statistical capabilities range from simple percentages to complex analyses of variance, multiple regressions, and general linear models.
- The many features of SPSS are accessible via pull-down menus or can be programmed with a proprietary syntax language.
  - Syntax programming has the benefits of reproducibility and handling complex data manipulations and analyses.
  - Although Syntax language can sound intimidating, it is very easy to use.
- SPSS datasets always have 2-dimensional table structure where the rows typically represent cases (such as individuals or households) and the columns or variables represent measurements (such as age, sex or household income).
- SPSS can read and write data from ASCII text files, other statistics packages, spreadsheets and databases.

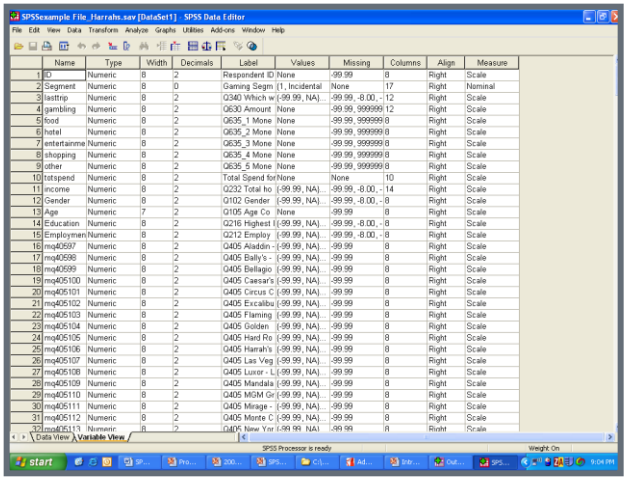
What does data in SPSS look like?

- **There are four main views in SPSS.**

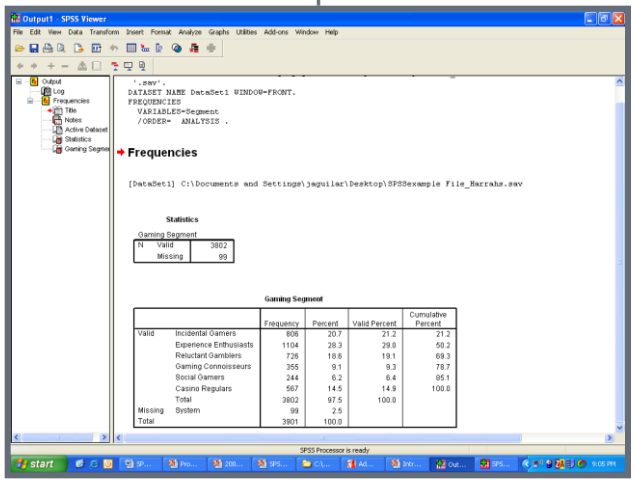
The Data View



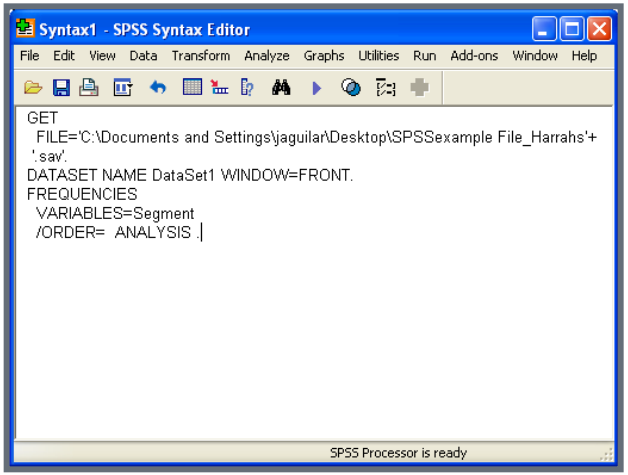
The Variable View



The Output View

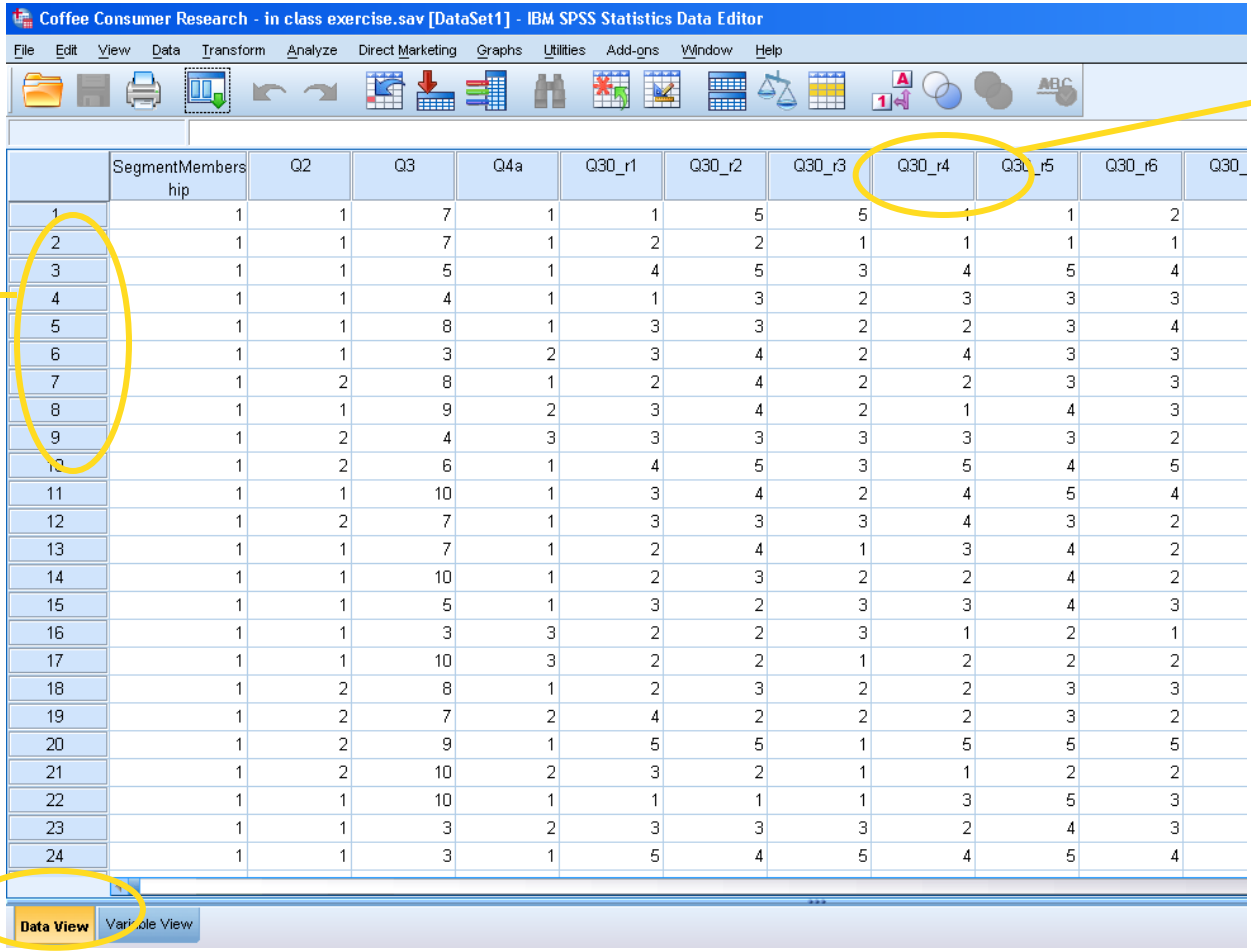


The Syntax View



What is the Data View?

- **Data View is arranged in a spreadsheet format that contains variables in columns and cases in rows.**



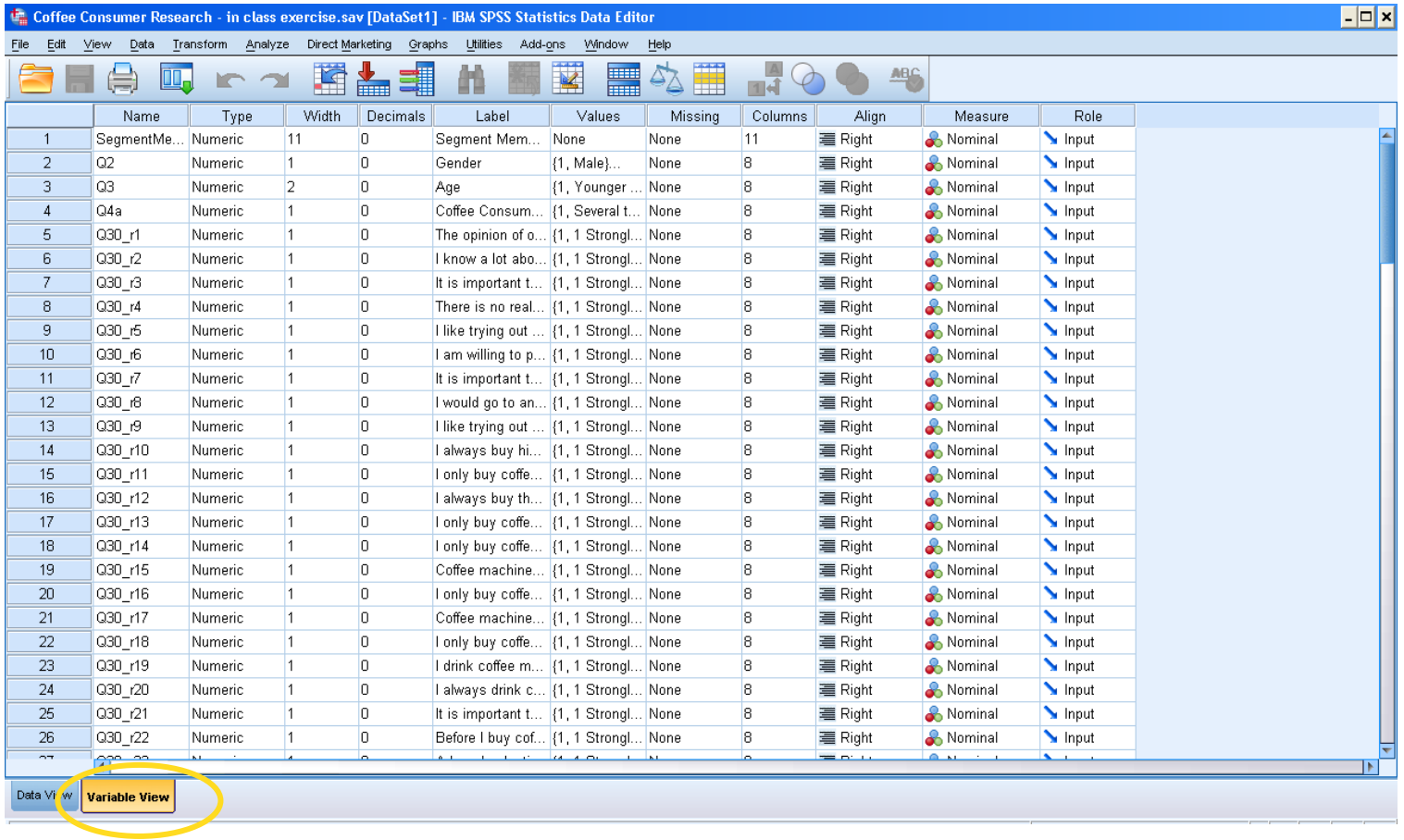
Variables

Cases or Records

Click on “Data View” to access this screen

What is the Variable View?

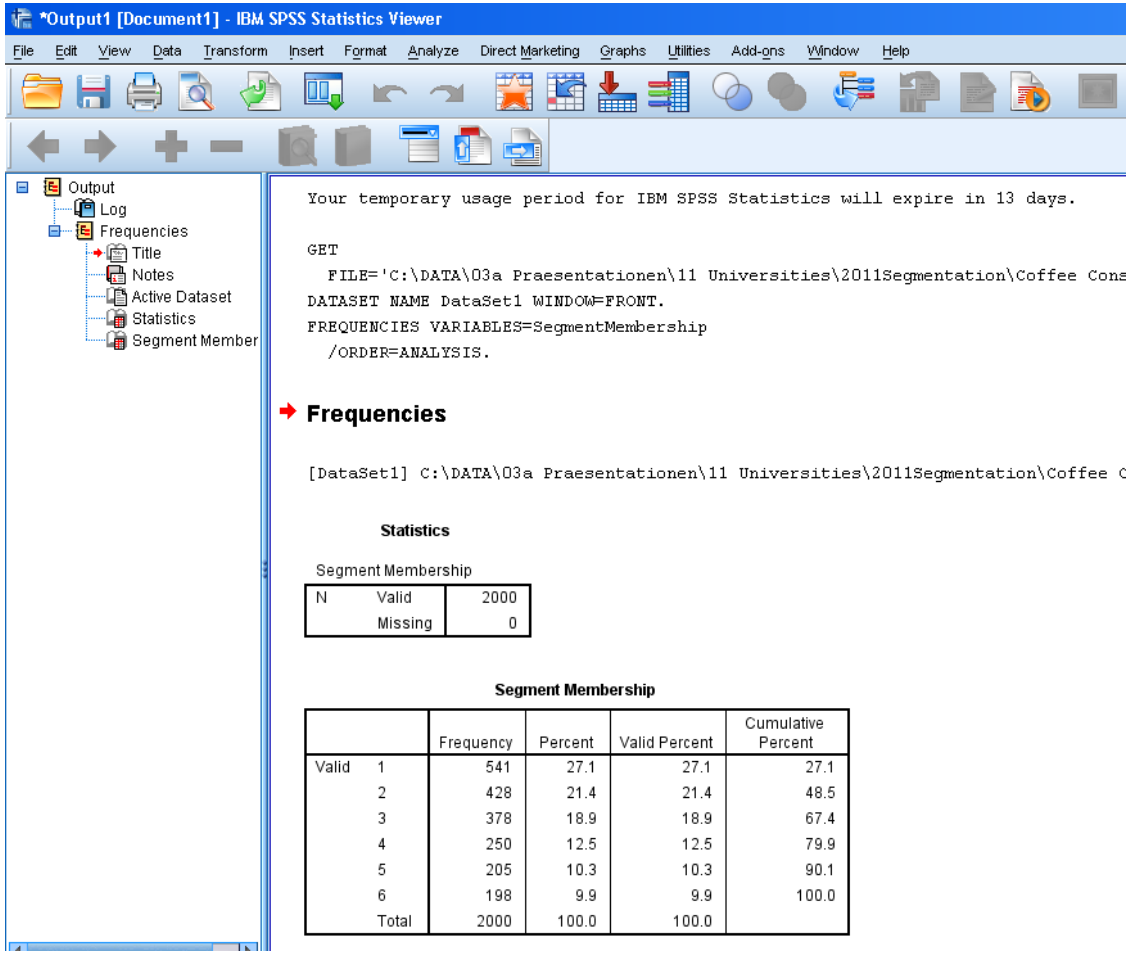
- **The Variable View window contains the definitions of each variable in your data set.**



Click on “Variable View” to access this screen

What is the Output View?

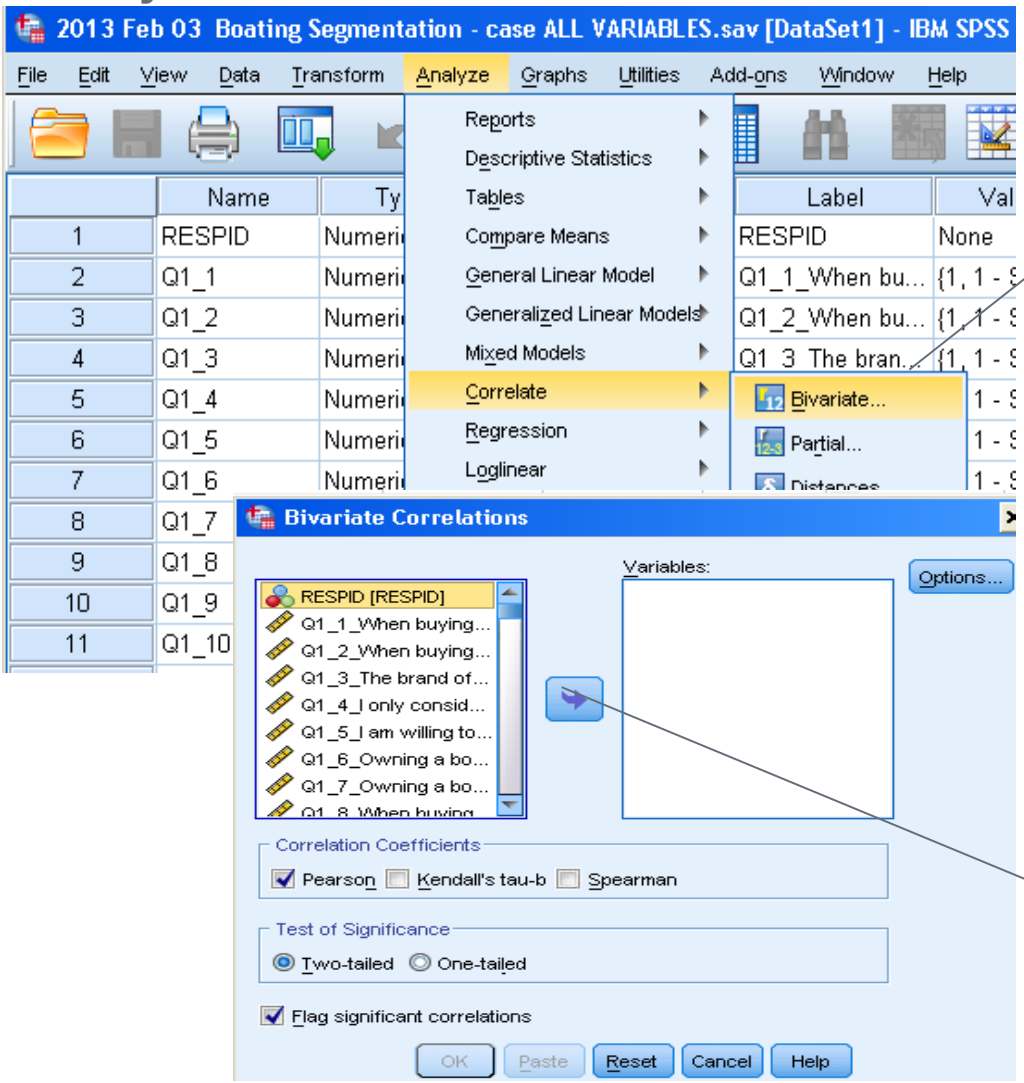
- **The Output View is where you see the results of your various queries such as frequency distributions, cross-tabs, statistical tests, and charts.**



- If you've worked with Excel, you're probably used to seeing all your work on one page, charts, data, and calculations.
- In SPSS, each window handles a separate task. The output window is where you see your results.

How do I run a correlation?

- **Analyze → Correlate → Bivariate…**



The **bivariate correlation** is for situations where you are interested in the relationship between different variables (e.g.could be run before a factor analysis)

To obtain correlations, click on the variable names in the variable list on the left side of the dialog box. Each variable listed in the Variables box will be correlated with every other variable in the box.

How do I interpret the standard correlation analysis output?

- **A correlation coefficient has a value ranging from -1 to 1. Values that are closer to the absolute value of 1 (or -1) indicate that there is a strong positive (or negative) relationship between the variables.**

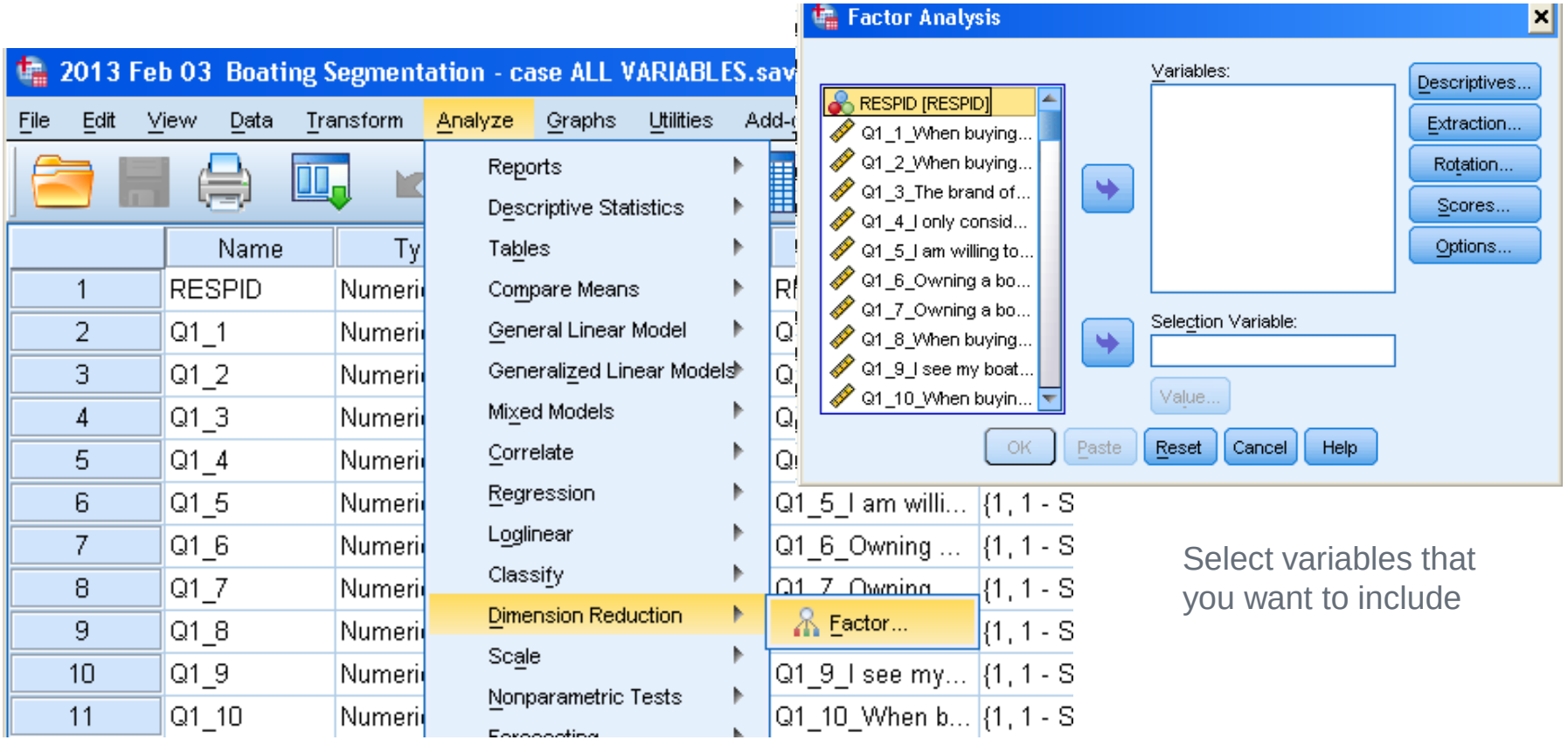
Correlations

		Q1_1_When buying a boat, I do a lot of shopping around and visit multiple dealers	Q1_2_When buying a boat, getting the lowest price is more important than the boat brand	Q1_3_The brand of boat I buy says a lot about who I am	Q1_4_I only consider buying a boat from a reputable brand	Q1_5_I am willing to pay a premium for a brand with a reputation for high quality
Q1_1_When buying a boat, I do a lot of shopping around and visit multiple dealers	Pearson Correlation	1	.009	.107**	.198**	.184**
	Sig. (2-tailed)		.634	.000	.000	.000
	N	2813	2813	2813	2813	2813
Q1_2_When buying a boat, getting the lowest price is more important than the boat brand	Pearson Correlation	.009	1	-.030	-.210**	-.206**
	Sig. (2-tailed)	.634		.117	.000	.000
	N	2813	2813	2813	2813	2813
Q1_3_The brand of boat I buy says a lot about who I am	Pearson Correlation	.107**	-.030	1	.264**	.400**
	Sig. (2-tailed)	.000	.117		.000	.000
	N	2813	2813	2813	2813	2813
Q1_4_I only consider buying a boat from a reputable brand	Pearson Correlation	.198**	-.210**	.264**	1	.367**
	Sig. (2-tailed)	.000	.000	.000		.000
	N	2813	2813	2813	2813	2813
Q1_5_I am willing to pay a premium for a brand with a reputation for high quality	Pearson Correlation	.184**	-.206**	.400**	.367**	1
	Sig. (2-tailed)	.000	.000	.000	.000	
	N	2813	2813	2813	2813	2813

This positive correlation coefficient (.400) indicates that there is a statistically significant ($p < .001$) linear relationship between these two variables

How do I run a factor analysis?

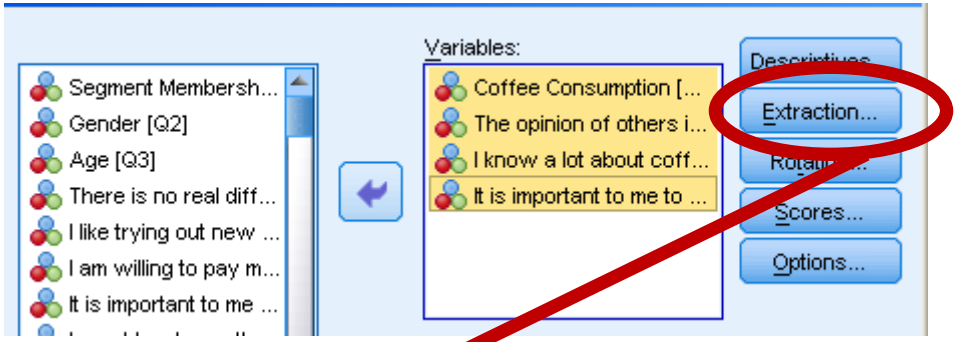
- **To run the factor analysis you must first select the variables you want to analyze.**
  - Analyze → Dimension Reduction → Factor…

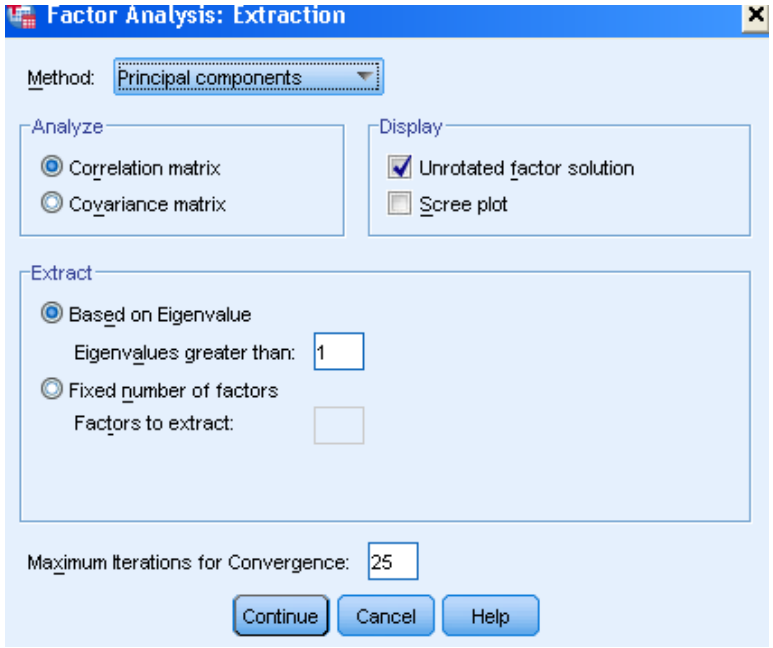


Select variables that you want to include

How do I run a factor analysis?

- **There are different methods of extracting the factors from a set of data. The following screenshots will show you the most common settings***

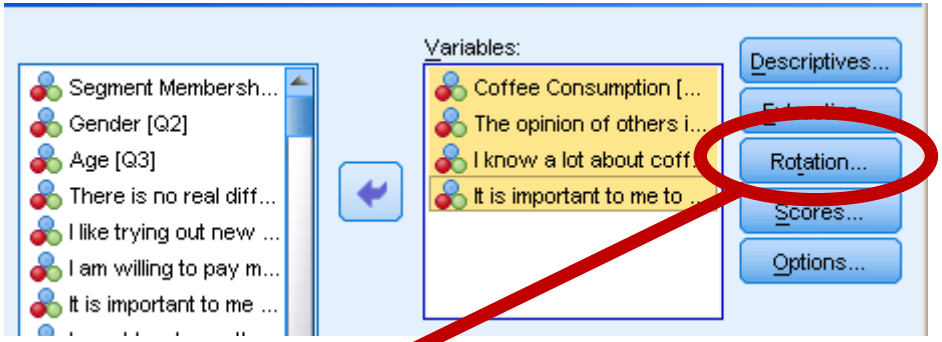


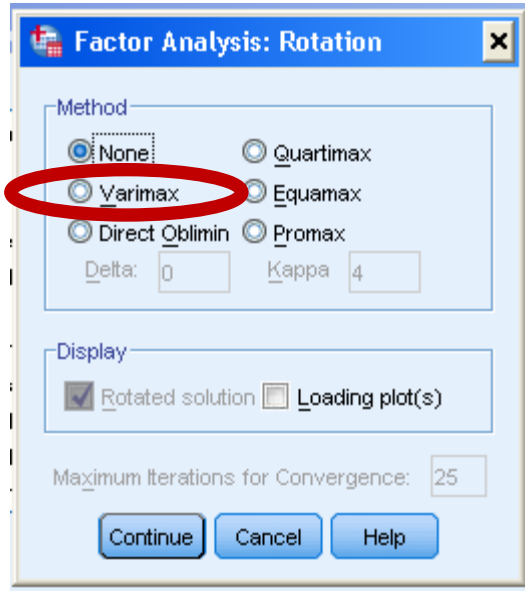


- There are several different types of factor analysis, with the most common being **principal components** analysis (PCA). Other available methods are unweighted least squares, generalized least squares, maximum likelihood, principal axis factoring, alpha factoring, and image factoring.
- You can either retain all factors whose eigenvalues exceed a specified value (e.g. commonly used value is 1), or you can retain a specific number of factors. The eigenvalue for a given factor measures the variance in all the data which is accounted for by that factor.

How do I run a factor analysis?

- **There are different methods of extracting the factors from a set of data. The following screenshots will show you the most common settings***



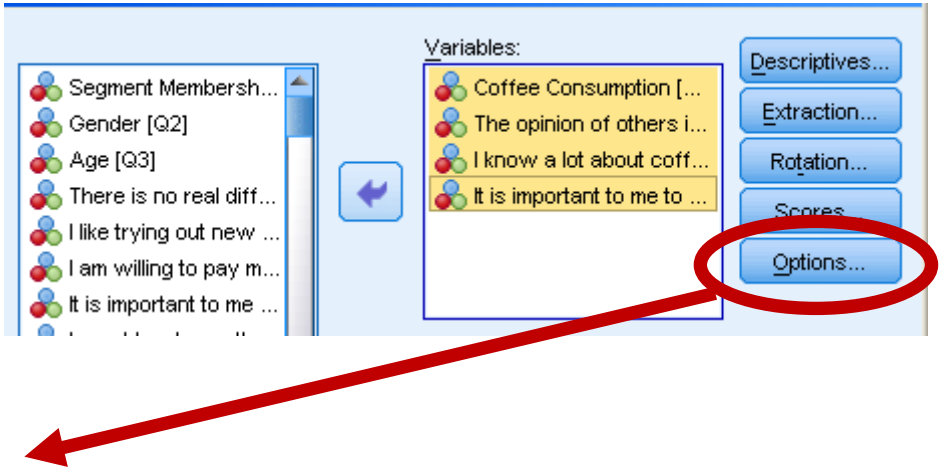


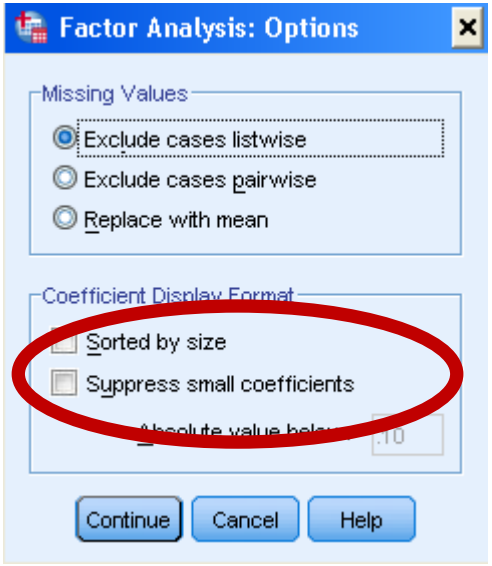
- Rotation serves to make the output more understandable and is usually necessary to facilitate the interpretation of factors.
- A **varimax solution** yields results which make it easier to identify each variable with a single factor. This is a common rotation option.

How do I run a factor analysis?

- **How can I get a more readable output?**

- By default SPSS will list the variables in the order in which they are entered into the data editor. Although this format is often convenient, when interpreting factors in large outputs it can be useful to list variables by size or suppress values less than a specific value.





- The Coefficient Display Format allows you to control aspects of the output matrices. You sort coefficients by size and suppress coefficients with absolute values that are less than the specified value.
- This makes the output easier to read by removing the clutter of low correlations that are probably not meaningful anyway. A common value is 0.4 or less.

How do I run a factor analysis?

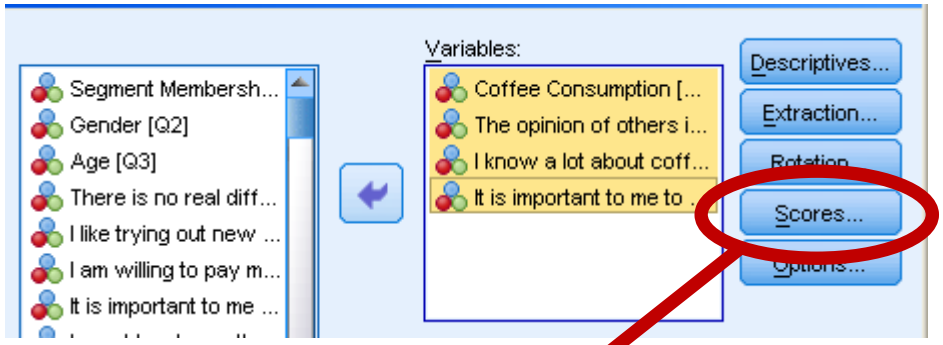
- **Rotated component matrix solutions are clear and simple to understand.**

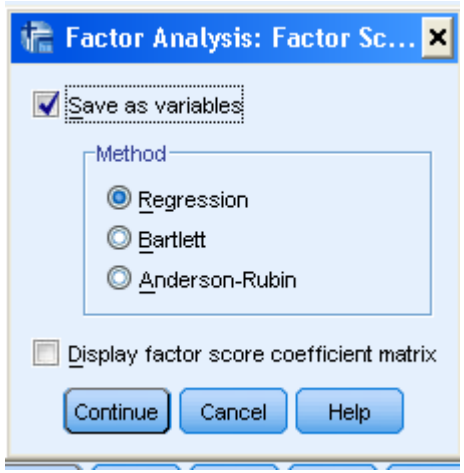
Rotated Component Matrix[a]

	Compone					
	1	2	3	4	5	6
I like trying out new things (i.e. new flavors) of the brands that I already use	.774					
I like trying out new coffee product brands	.635					
A broad selection of flavors and degrees of boldness is important to me	.603					
I am curious, like making new discoveries and learning new things.	.570					
I drink coffee to indulge myself	.410					
I drink different coffee products at different times of the day (e.g. sometimes with and sometimes without milk)						
Style, design and fashion are important to me.		.772				
I am a trend setter		.713				
I like to see and be seen.		.703				
Coffee machines must match the rest of my kitchen (e.g. design, color,		.617				

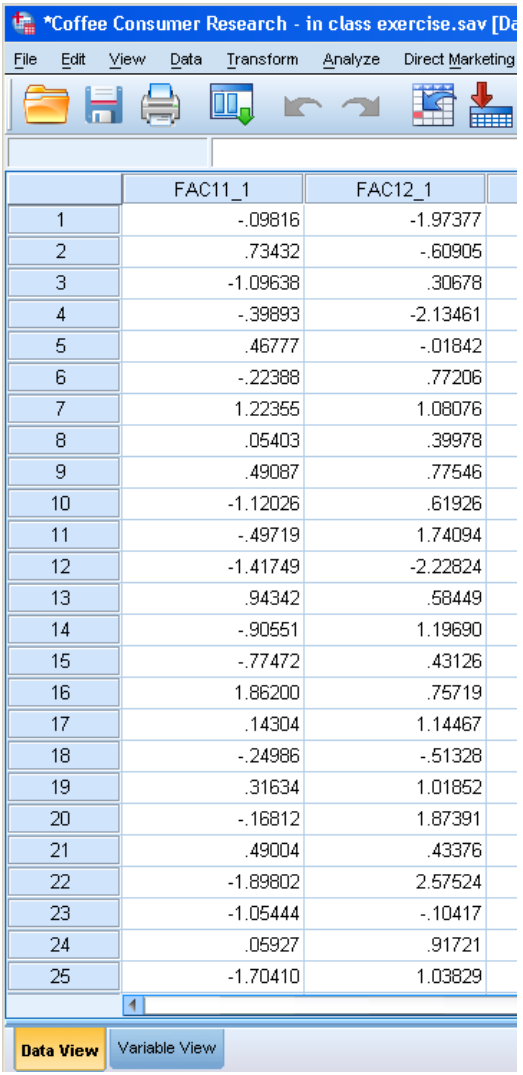
- The high loadings indicate that these variables are most associated with Factor 1
- Naming the factors help interpretation and explanation of output (e.g. factor 1 = “gambling excitement”)
- Values range from -1 to +1
- Magnitude is the key to understanding the factors, so .703 is as important as -.703
- Loadings below .4 are often ruled out for analysis

How do I save factor scores?





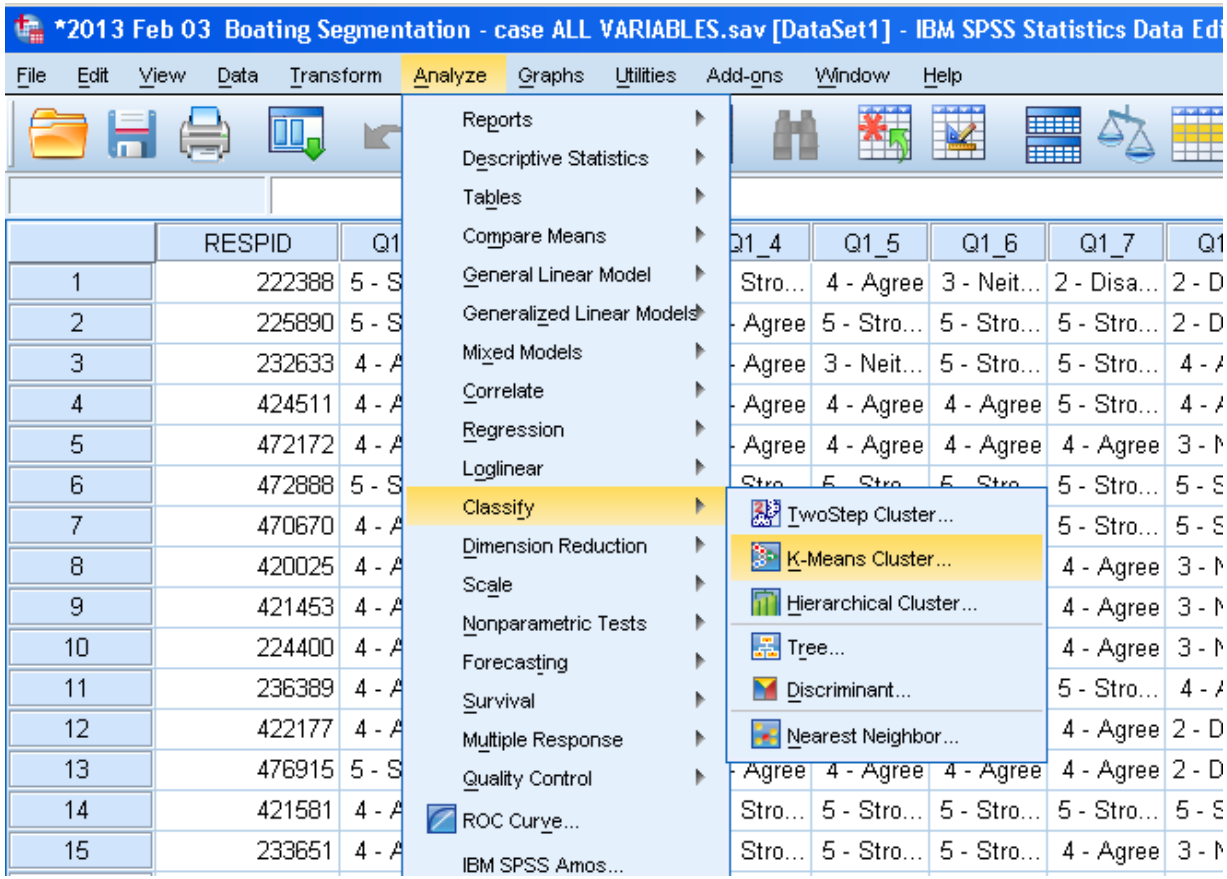
- The factor scores can be saved so that they can be used for future analysis
- Factor scores are saved as variables in the data file, **at the very last columns** of the data so far



	FAC11_1	FAC12_1
1	-.09816	-1.97377
2	.73432	-.60905
3	-1.09638	.30678
4	-.39893	-2.13461
5	.46777	-.01842
6	-.22388	.77206
7	1.22355	1.08076
8	.05403	.39978
9	.49087	.77546
10	-1.12026	.61926
11	-.49719	1.74094
12	-1.41749	-2.22824
13	.94342	.58449
14	-.90551	1.19690
15	-.77472	.43126
16	1.86200	.75719
17	.14304	1.14467
18	-.24986	-.51328
19	.31634	1.01852
20	-.16812	1.87391
21	.49004	.43376
22	-1.89802	2.57524
23	-1.05444	-.10417
24	.05927	.91721
25	-1.70410	1.03829

How do I identify like groups of respondents (Clustering)? K-means approach

- **K-means approach is often used for clustering data**



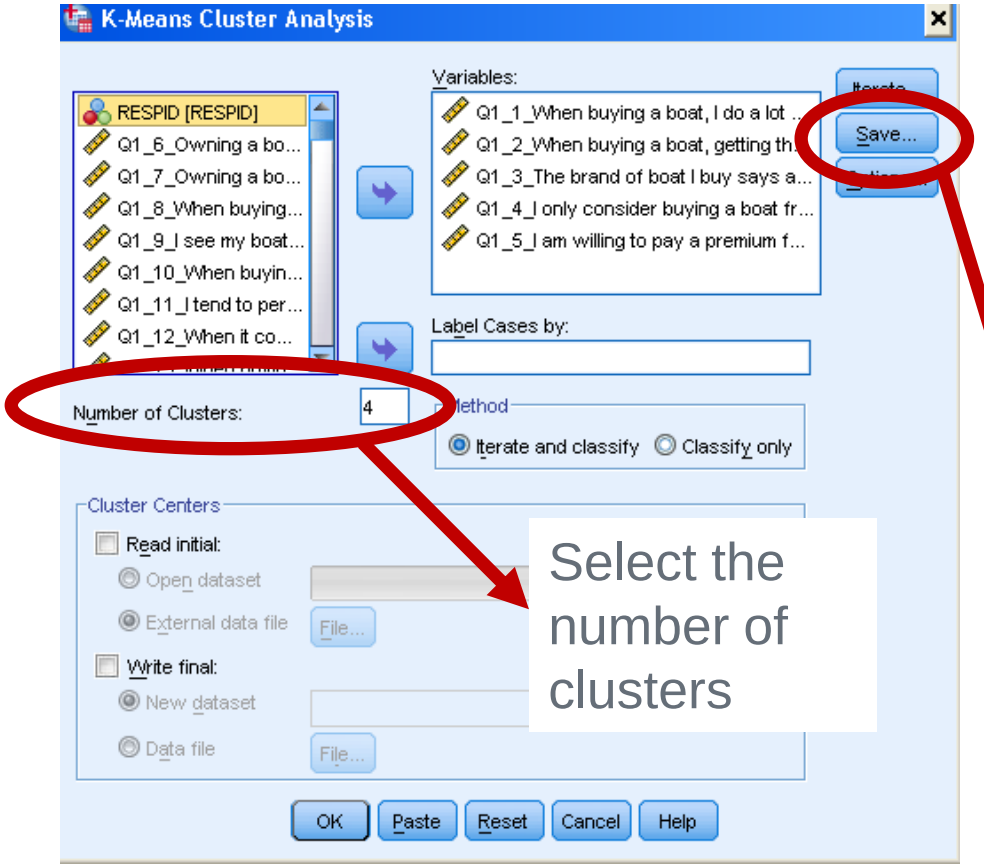
With kmeans Clustering we can:

- Identify the appropriate number of clusters/segments
- Assign respondents to cluster/segment
- Describe each cluster/segment

If you have many variables k-means may not be appropriate: hence factor analysis can be used before k-means to decrease the "dimensionality" of the data

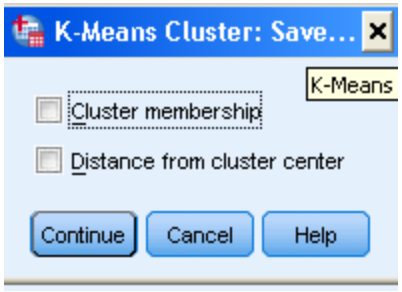
How do I identify like groups of respondents (Clustering)? K-means approach

- **To run the cluster analysis you must first select the variables you will use to group respondents**



Select the number of clusters

- SPSS uses an algorithm to group respondents based on their answers to the questions you selected.
- You can also ‘save’ the cluster membership so you can tell which respondents fall into which group and you can do further analyses. The cluster membership will be added again after the **last column of your data.**



How do I identify like groups of respondents?
K-means approach

- **The Cluster output will tell us how similar our respondents are if we need more or less clusters based on the distance from the cluster center and the number of respondents in each group.**

Final Cluster Centers

	Cluster			
	1	2	3	4
Q505_1 How important: Feels exciting, high-energy	6.50	7.96	4.72	6.91
Q505_2 How important: Feels lucky, like people are winning	6.42	8.01	4.68	7.50
Q505_3 How important: Feels comfortable	7.67	8.90	7.05	8.63
Q505_4 How important: Is a place I would take my friends	7.23	8.55	6.27	7.78
Q505_5 How important: I would describe as upscale	5.97	7.59	4.43	6.51
Q505_6 How important: Is in a convenient location	7.13	8.06	6.59	7.49
Q505_7 How important: Has casinos in multiple cities				
Q505_8 How important: Appeals to people like me				
Q505_9 How important: Appeals to experienced gamblers				
Q505_10 How important: Appeals to blue collar/working-class people				
Q505_11 How important: Appeals to a younger crowd				

Number of Cases in each Cluster

		Unweighted	Weighted
Cluster	1	313.000	1153.130
	2	296.000	1199.164
	3	211.000	627.604
	4	140.000	451.462
Valid		960.000	3431.360
Missing		120.000	469.495

- Run several analyses with a different number of clusters then choose the best solution
- Look at the cluster centers to see if you have unique segments
  - You want to see differences across the segments on at least several questions so you know there is a difference across segments
- Look at the number of respondents in each cluster to see if they are distributed across segments
  - If your clusters are heavily skewed towards one or two groups you may not have enough clusters

Which Segmentation Solution is Better?

- **Compare these two cluster analyses and decide which solution provides the most difference across segments.**

4 Cluster Solution

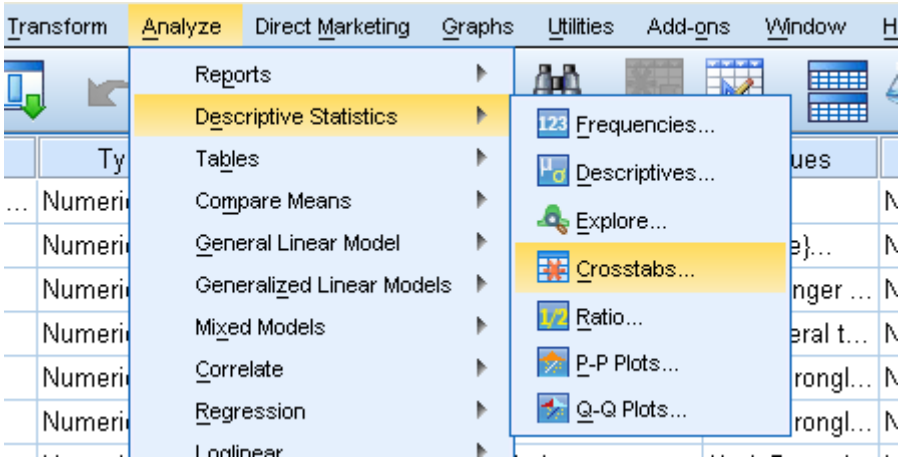
	Cluster			
	1	2	3	4
Q505_1 How important: Feels exciting, high-energy	6.50	7.96	4.72	6.91
Q505_2 How important: Feels lucky, like people are winning	6.42	8.01	4.68	7.50
Q505_3 How important: Feels comfortable	7.67	8.90	7.05	8.63
Q505_4 How important: Is a place I would take my friends	7.23	8.55	6.27	7.78
Q505_5 How important: I would describe as upscale	5.97	7.59	4.43	6.51
Q505_6 How important: Is in a convenient location	7.13	8.06	6.59	7.49
Q505_7 How important: Has casinos in multiple cities	3.14	4.48	2.45	2.89
Q505_8 How important: Appeals to people like me	7.31	8.53	5.82	7.87
Q505_9 How important: Appeals to experienced gamblers	4.87	6.37	3.84	5.88
Q505_10 How important: Appeals to blue collar/working-class people	4.81	6.49	3.46	4.67
Q505_11 How important: Appeals to a younger crowd	4.86	5.68	3.74	4.57

5 Cluster Solution

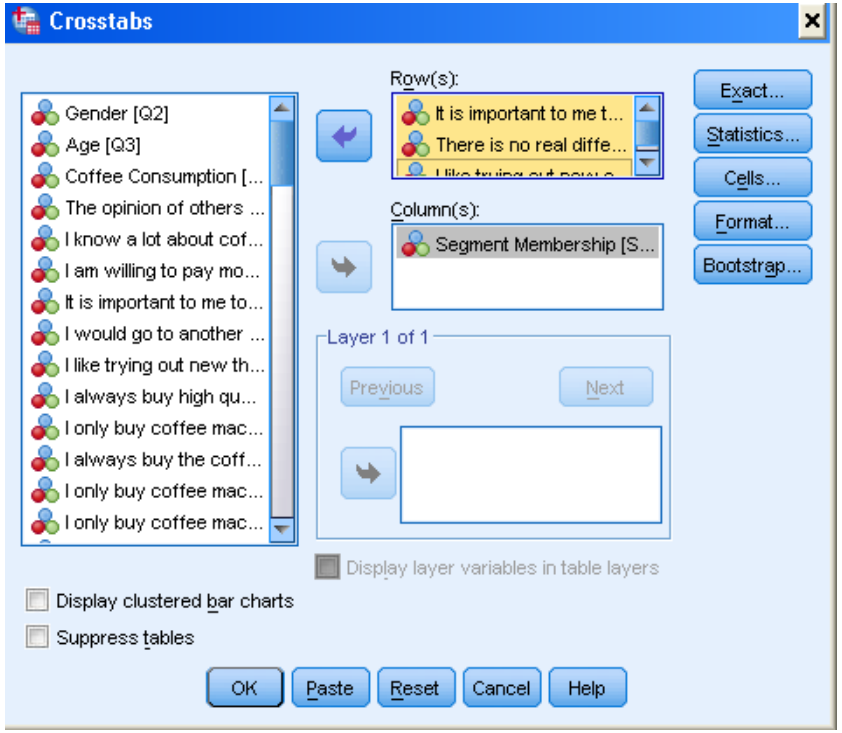
	Cluster				
	1	2	3	4	5
Q505_1 How important: Feels exciting, high-energy	5.72	6.68	6.66	3.43	8.26
Q505_2 How important: Feels lucky, like people are winning	5.75	6.29	7.14	3.46	8.23
Q505_3 How important: Feels comfortable	7.48	7.62	8.26	6.51	9.06
Q505_4 How important: Is a place I would take my friends	6.78	7.23	7.82	4.36	8.82
Q505_5 How important: I would describe as upscale	5.27	6.71	5.80	3.82	7.87
Q505_6 How important: Is in a convenient location	7.11	6.93	7.44	4.98	8.32
Q505_7 How important: Has casinos in multiple cities	2.63	2.94	3.24	1.89	4.86
Q505_8 How important: Appeals to people like me	6.55	7.25	7.85	4.35	8.78
Q505_9 How important: Appeals to experienced gamblers	4.50	4.88	5.39	2.78	6.69
Q505_10 How important: Appeals to blue collar/working-class people	3.95	4.57	5.40	2.34	6.65
Q505_11 How important: Appeals to a younger crowd	3.99	5.12	4.67	3.02	6.06

How do I profile clusters?

- **Once you have finalized the clusters you can easily profile them with some descriptive statistics. It may be better to do so instead of using the output of cluster analysis indicated in the previous slide**
  - From the menu, select Analyze → Descriptive Statistics → Crosstabs



  - Click once on Segment, then click the right arrow next to Column(s) to move the variable to the Columns pane.
  - Now move the statements to the Row(s) pane. Click OK.

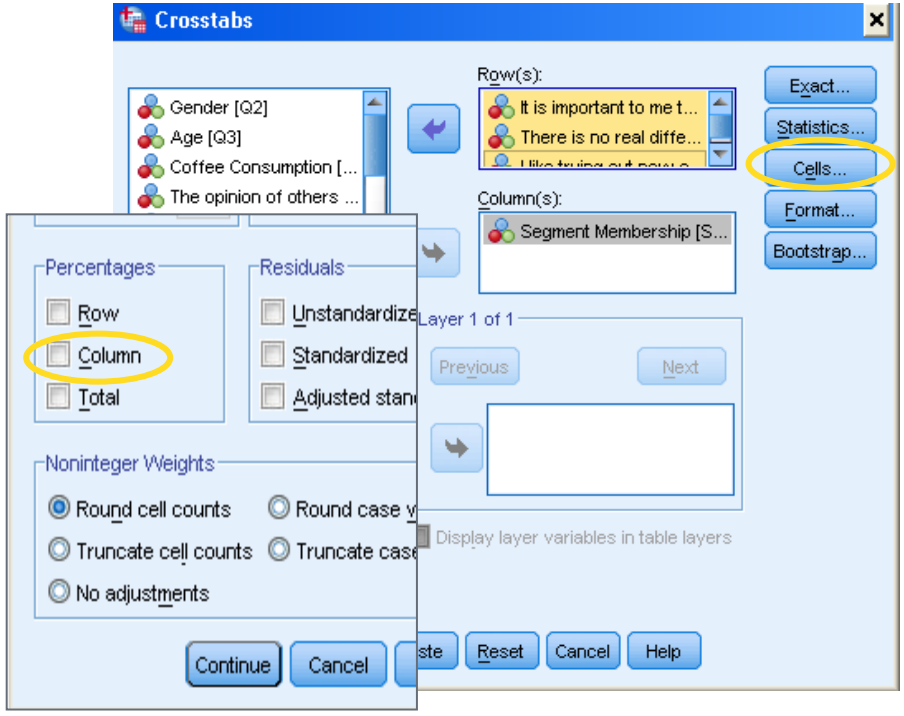


  - **Remember that cluster membership should had been saved, and it should appear at the end of the data (last column)**

How do I profile clusters?

- **You can also use percentages in crosstabs to understand the frequency distribution or e.g. Top 2 box percentages by segment.**

  - After selecting the variables that will go into Rows and Columns, click on the Cells button.
  - From there, check the Column box under Percentages. Click on Continue



  - This will result in the following table that also shows you the percentages WITHIN a segment:

always buy the coffee machines and coffee products with the lowest price * Segment Membersh

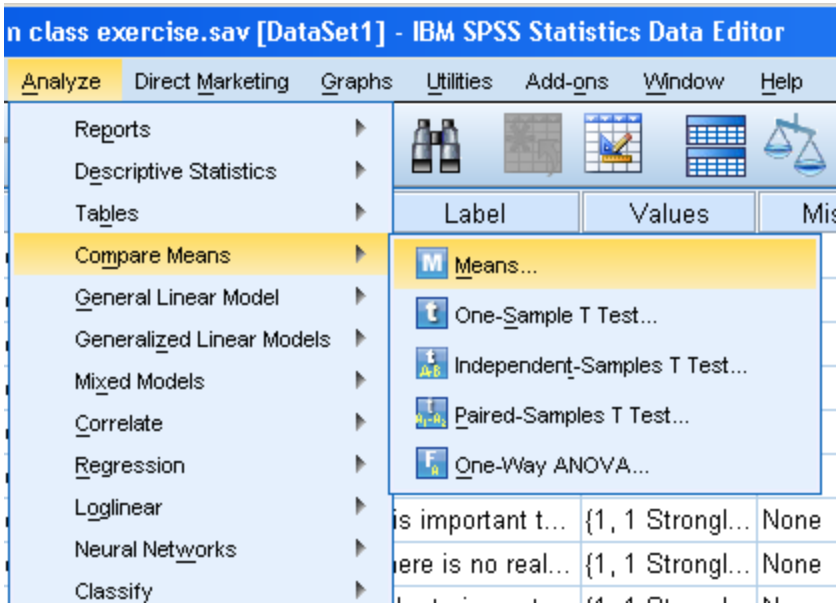
		Segment Membership			
		1	2	3	4
1 Strongly disagree	Count	51	167	75	8
	% within Segment Membership	9.4%	39.0%	19.8%	3.2%
2 Disagree	Count	156	198	143	67
	% within Segment Membership	28.8%	46.3%	37.8%	26.8%
3 Neutral	Count	196	62	118	79
	% within Segment Membership	36.2%	14.5%	31.2%	31.6%
4 Agree	Count	108	1	37	85
	% within Segment Membership	20.0%	.2%	9.8%	34.0%
5 Strongly agree	Count	30	0	5	11
	% within Segment Membership	5.5%	.0%	1.3%	4.4%
	Count	541	428	378	250
	% within Segment Membership	100.0%	100.0%	100.0%	100.0%

Top 2 Box: 25.5% vs. 38.4%

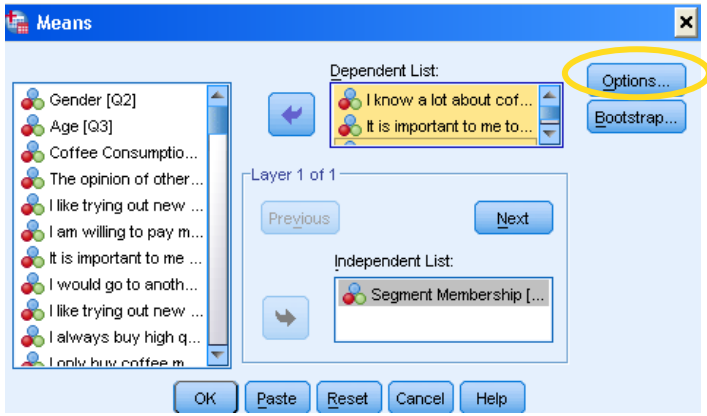
How do I profile clusters?

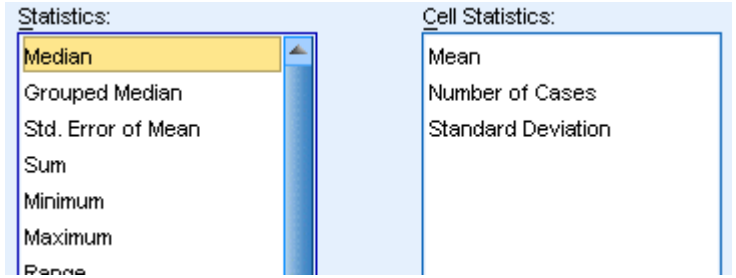
- **Another useful tool to help understand a segment solution is the Compare Means procedure. In this analysis you also can use standard deviation, Sum, Range and other statistics.**

  - From the menu, select Analyze → Compare Means → Means



  - Select the segment variable as the Independent variable and the brand attributes as the dependent variables.
  - Under Options you can choose from a list of different statistics. From there, click OK.





How do I profile clusters?

- **This will result in the following table:**

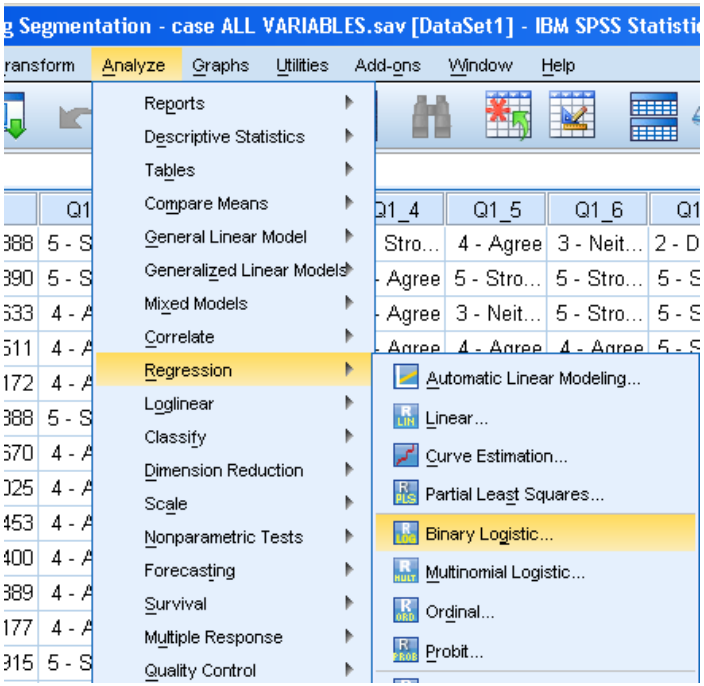
Report

Segment Membership		The opinion of others is important to me when it deals with the purchase of coffee machines	I know a lot about coffee products and make my decisions based on my own knowledge	It is important to me to purchase the coffee products that I grew up with	There is no real difference between the various brands for coffee products	I like trying out new coffee product brands	I am willing pay more f coffee bran that I have emotiona relationshi with
1	Mean	2.91	2.93	2.33	2.39	3.37	2.
	N	541	541	541	541	541	5
	Std. Deviation	1.053	.979	.978	1.049	.976	1.0
2	Mean	2.78	3.54	2.14	1.61	3.79	3.
	N	428	428	428	428	428	4
	Std. Deviation	1.051	.824	.917	.877	.926	1.1
3	Mean	2.93	2.91	2.22	2.00	3.44	3.
	N	378	378	378	378	378	3
	Std. Deviation	.980	.904	.935	1.000	.857	1.0
4	Mean	3.18	2.97	2.10	1.98	3.70	2.
	N	250	250	250	250	250	2
	Std. Deviation	.955	.873	.821	.986	.848	1.1
5	Mean	2.58	3.39	1.95	1.85	3.59	2.
	N	205	205	205	205	205	2
	Std. Deviation	1.107	.882	.833	1.088	.964	1.0
6	Mean	2.52	3.04	1.93	1.77	3.17	2.
	N	198	198	198	198	198	1
	Std. Deviation	1.116	.854	.887	.947	1.112	1.1
Total	Mean	2.85	3.12	2.16	1.98	3.52	2.
	N	2000	2000	2000	2000	2000	20
	Std. Deviation	1.055	.934	.924	1.031	.961	1.1

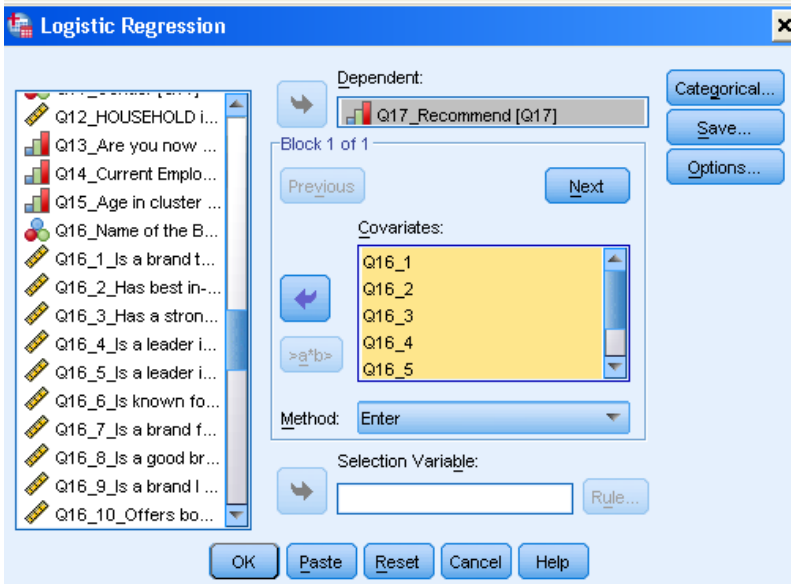
How do I run a binary logistic regression?

- **Let's understand which variables are good predictors of "Purchase"**

- From the menu, select Analyze → Regression → Binary Logistic



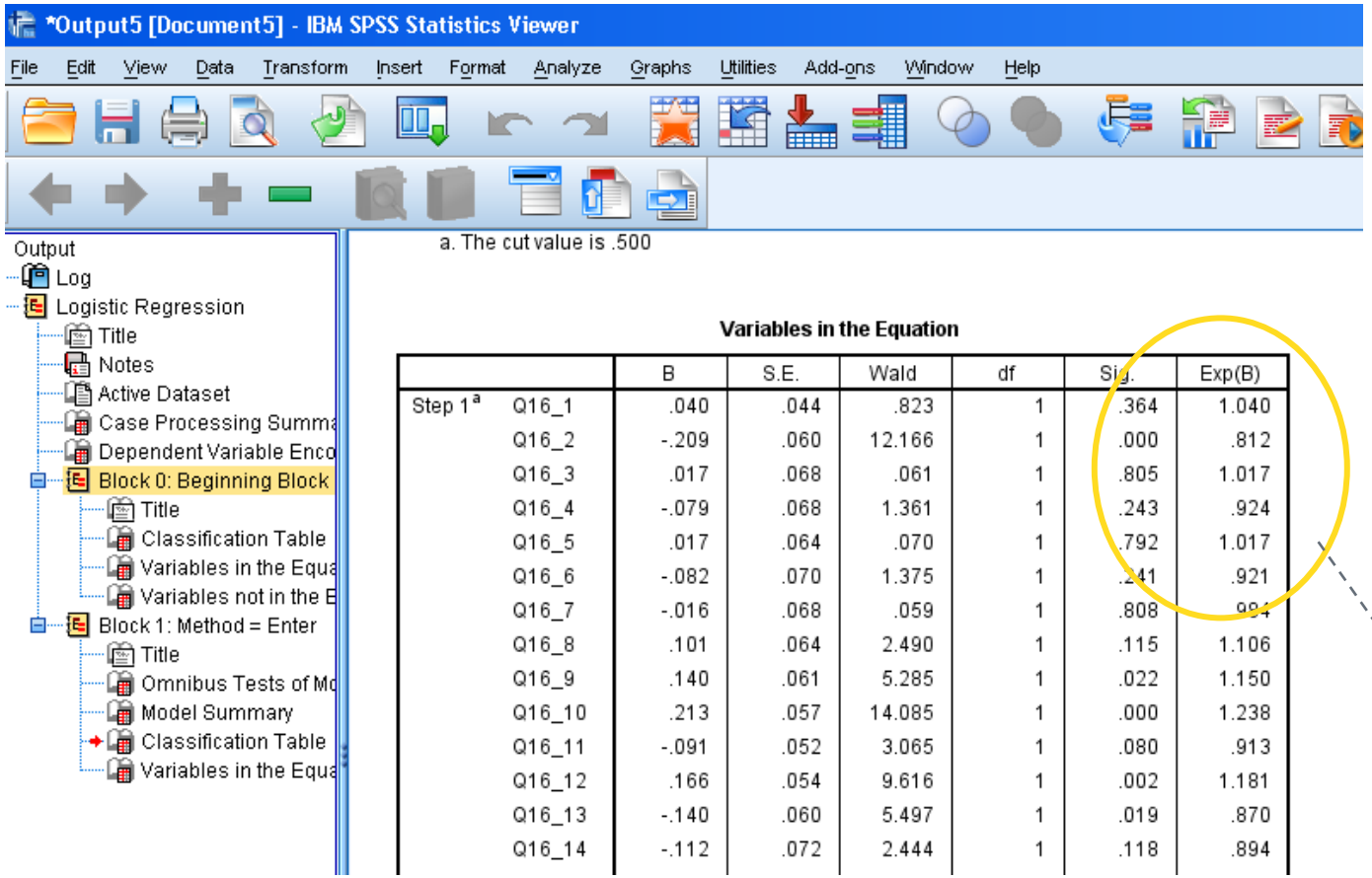
- Chose the variable "Recommend" as the dependent variable.
- In the Independent(s) box, several predictor variables have been entered, e.gQ16



- The box labeled Method allows you to select from one of five methods: "Enter" is the standard approach in regression models.

How do I run a binary logistic regression?

- **How do I interpret the output?**



a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Q16_1	.040	.044	.823	1	.364	1.040
	Q16_2	-.209	.060	12.166	1	.000	.812
	Q16_3	.017	.068	.061	1	.805	1.017
	Q16_4	-.079	.068	1.361	1	.243	.924
	Q16_5	.017	.064	.070	1	.792	1.017
	Q16_6	-.082	.070	1.375	1	.241	.921
	Q16_7	-.016	.068	.059	1	.808	.984
	Q16_8	.101	.064	2.490	1	.115	1.106
	Q16_9	.140	.061	5.285	1	.022	1.150
	Q16_10	.213	.057	14.085	1	.000	1.238
	Q16_11	-.091	.052	3.065	1	.080	.913
	Q16_12	.166	.054	9.616	1	.002	1.181
	Q16_13	-.140	.060	5.497	1	.019	.870
	Q16_14	-.112	.072	2.444	1	.118	.894

- Exp(B) shows you the importance of each variable