Center for Marketing Technology



SPSS

Data Mining/Data Analysis

**Table of Contents**

1. **SPSS Overview** 2

2. **Loading SPSS** 2

3. **Using SPSS** 2

3.1. Start-up Choices 2

3.2. Entering/Modifying Data 3

3.3. Defining Variables 3

3.4. Computing a New Variable 5

3.5. SPSS Windows 5

3.6. Saving 6

3.7. Printing 6

4. **Data Analysis** 6

4.1. Purpose of Analysis 6

4.2. How to Run Statistical Analysis 7

1. Frequencies 7

2. Descriptive 8

3. Crosstabs 9

4. Correlation 11

4.3. Graphs 11

1. Histogram 11

2. Bar Charts 12

5. **Helpful Hints** 13

5.1. Analyzing a subset 13

5.2. Recoding a variable 13

5.3. Importing Excel files 14

5.4. Conclusion 14

6. **Special Topics**

6.1. RFM Analysis

6.2. Logistic Regression Analysis

6.3. Decision Tree Analysis

**1. SPSS Overview**

SPSS (Statistical Package for the Social Sciences) is a data mining tool that covers a broad range of statistical procedures that allow you to summarize data (compute means and standard deviations), determine whether there are significant differences between groups (t-tests, analysis of variance), examine relationships among variables (correlation, multiple regression), and graph results (e.g., bar charts, line graphs). In practice, it offers companies information about the most and least profitable customers, their purchase patterns, buying behaviors and demographic profiles, which are key to developing a successful marketing strategy.

**2. Loading SPSS on your computer**

In order to access SPSS from your computer, you must first load the SPSS shell onto your computer. Once the shell is loaded, you can access SPSS **as long as you are connected to the network**. To get the shell loaded onto your computer do the following:

1. Click on the "Bentley College Virtual Lab Icon"
2. Once in the Virtual Lab, click on the "Tools" icon
3. Once in the Tools window, click on the "SPSS Version11 Base Setup.” Follow the instructions given in the setup

The above procedure should give you an SPSS icon in your start menu. However, if for some reason it does not work try the following:

Click on the “Start” menu, click "Run" then type in:

\\electra\spss$\spss11\Setup\setup.exe

**3. Using SPSS**

**3.1. Startup Choices**

Start SPSS by clicking on Start ⇨ Programs ⇨ SPSS 11.0 for Windows. Once the program launches, you will be presented with a few choices:

1. Run the tutorial
2. Type in data -- use this option to create and enter new data file
3. Run an existing query -- use this option to run syntax
4. Create new query using Database Wizard -- use this option to import data from other formats i.e. excel or access
5. Open an existing data source -- use this option to open a data file you have previously created in SPSS
6. Open another type of file – use this option to open a non-SPSS data file such as Excel

**3.2. Entering Data on SPSS**

1. In the Data Editor Window, each row is a case (or observation) and each column is

a variable.



2. The cells in each row hold the value of a particular case (or observation) for a particular variable. At the top of the every column is the variable name – which can be no more than 8 characters long 3. Before filling in the observation values, you should define the variables you will be

working with – see section below for thorough explanation 4. Once variables have been defined and you are ready to enter values, use the “Enter” key to move to the cell on the row below or the “Tab” key to move to the cell to the right on the same row

**3.3. Defining Variables**

1. Double-click on the first column heading (where it says “var”) or click on the “Variables View” tab at the bottom of the page 2. To define a variable, you must complete the following fields:

1. Variable Name - type the desired variable name, which can be *no more than 8 characters in length*. The first character must be alphabetic; the remaining characters can be alphabetic and/or numeric, and no spaces can appear in the name
2. Type - allows you to define the type of variable, which can which can be:

(Click on the  icon to view the following choices)

1. Numeric - observations are numbers
2. Comma - observations are displayed with commas delimiting every three places, and with the period as a decimal delimiter
3. Dot - observations are displayed with periods delimiting every three places, and with the comma as a decimal delimiter
4. Scientific notation - observations are displayed with an imbedded E and a signed power-of-ten exponent (for example, 123, 1.23E2)
5. Date - observations are displayed in one of several date formats (select one from the list)
6. Dollar - observations are displayed in one of several formats (select one from the list)
7. Custom currency - observations are displayed in one of the custom currency formats that you define by going to the top menu under Edit | Options | Currency tab
8. String - observations are non-numeric, they can contain any characters up to the defined length
9. Width and Decimals – use these fields to limit the number of characters and define how numbers should be displayed d) Label - assign descriptive variable and value labels
10. Label - assign descriptive variable and value labels
11. Variable Label - allows you to list a more comprehensive label for your variable that is not limited to 8 characters and may have spaces
12. Value Label - allows you to provide labels for the various levels of a variable (click on the  icon to proceed)
13. For non-numeric variables, such as gender, it is very useful to set values such as 1=male, 2=female so that they are considered numeric after all and included in the calculations/data analysis
14. Do so if appropriate, by placing the value (e.g., "1") in the Value field and the value label (e.g., "male") in the Value Label field and clicking "Add." You can do this process as many times as necessary for each string variable
15. Missing Values - enables you to designate certain observation scores as missing
16. Columns and Align - these fields allow you to change the maximum number of characters permitted in a column as well as determine the alignment of the text
17. Measurement - allows you to determine the kind of measure for that particular variable. It can be:
18. Scale - numeric data on an interval or ratio scale
19. Ordinal – string and numerical variables with defined value labels
20. Nominal - mostly string variables

A nice shortcut to label several variables at once is to use a global template.

* + - 1. First, in the Data window, highlight the columns that you want to have the same labels. Unfortunately, SPSS only allows you to highlight columns that are next to each other 2.
			2. Click on the "Data" menu. Click on "Templates"
			3. In the "Apply" area, click on "Value Labels"
			4. Click on "Define" In the window that appears, and then follow the same instructions given above

**3.4. Computing a New Variable**

New variables can be created by manipulating the existing data set. For example, you might have weekly data but would like to analyze this same data in monthly time periods.

1. Click "Transform" on the top menu line then click "Compute" from the drop-down menu.
2. In the “Target Variable” dialogue box, enter the name of the new variable (e.g., "Monthly").
3. Select the variable (“Weekly” in this case) you want to compute for by highlighting it and clicking on the arrow to place it into the “Numeric Expression box.”
4. Select the appropriate formula for the computation by using the keypad or predefined functions. Continuing with our example, click "\*" and then "4."
5. Click "OK" and SPSS will compute a new variable called “Monthly” by multiplying each observation under the variable “Weekly” by 4

**3.5. SPSS Windows**

When using SPSS, you will have two types of windows open, the Data Editor window (as illustrated previously), where the data set is displayed and the Output Viewer window.

The Output Viewer window is where you will always find your output (i.e. test results) and any messages or commands from SPSS.



To get back to your Data window, you can click on the menu option titled Window, and then select your data file. Or, you can click on the little button that looks like a data matrix. Whenever you conduct an analysis procedure (e.g. get a frequency distribution), SPSS will automatically bring up the Output Viewer window. You do not need to go back to the Data Editor window to run a different analysis on the same data set.

**3.6. Saving**

You can save any of the windows just like you would on any other program by clicking on the disk icon or going to the file menu. To save a window, make that particular window active by clicking on it. SPSS automatically adds a three-letter suffix to the end of the file name (".sav" for data editor files, ".spo" for output files). Thus, with a particular data set, it is recommended to use a single name for both files. Make sure you save it to a disk and a directory that you will be able to access whenever needed. Save on a regular basis.

**4. SPSS and DATA ANALYSIS**

**4.1. Purpose of Statistical Analysis**

To calculate statistics or create graphs, you must first select the appropriate procedure from the "Analyze" menu on the top toolbar. The drop-down menu will give you many options (some of which are outlined below). **Please note**: SPSS does a lot more than what's described here. Use the SPSS help menus at any time to find out more.

|  |  |  |
| --- | --- | --- |
| **Statistical Analysis** | **Procedures** **Available** | **Description** |
| Descriptive Statistics | Frequencies | Gives you frequency counts. Useful for summarizing the number of occurrences and percentages of observations under each variable. |
| Descriptives | Gives you various descriptive statistics such as means, standard deviations, minimum and maximum values. |
| Crosstabs | Calculates frequency tables combining two variables. A Chisquare test may be added to identify whether or not the difference in frequency counts are statistically significant or not. |
| Compare Means | Means | Get means (and other statistics) for the entire data set or for subsets of the data (e.g., an average GPA for males and for females, as opposed to the overall average). |
| One-Sample T-Test | Test if the mean of some variable is significantly different from some hypothesized number. |
| Independent Samples T-Test | Test if the means of two groups are significantly different. |
| One-way ANOVA | Determine if categorical independent variables have an effect on a dependent variable. |
| Correlate | Bivariate | Determine if two variables are positively or negatively correlated such that they "move together” or “in opposite directions” |
| Regression | Linear | Calculate a "line" that will use one or more independent variables to predict the value of a dependent variable. |

The choice of which procedure to run really depends on the type of data you are analyzing and what exactly you are trying to find out.

There are two main types of data: categorical and continuous. Categorical data represents types of data that can be divided into groups. Examples include race, sex, age group, and educational level. Continuous data represents types of data that are associated with some sort of measurement. Examples include dollar amounts, time and units. Briefly, frequency tables and crosstabs are indicated for categorical data and descriptives and correlations are indicated for continuous data.

**4.2. How to run a statistical analysis**

The following analyses are based on customer surveys conducted by “PC’s Unlimited” Catalog Company. This dataset is available at the CMT for practice purposes. To open this file, click on the “My Computer” icon on the desktop, then click on C Drive (C:) / Data/ mk 721/ PC’sUnlimited.sav

1. Frequencies

The “Frequencies” procedure is useful for summarizing the number and percent of observations that have each value of a categorical variable. For example, the PCsUnlimited’s dataset includes a variable called REGION indicating which REGION each customer lives in.

To determine out how many customers are from each REGION:

1. From the pull-down menus, select **Analyze / Descriptive Statistics / Frequencies**
2. In the dialog box that appears, select REGION for the variable list (click on REGION, then click on the arrow to move it to the Variable List)
3. Make sure that the “Display Frequency Tables” box contains a check
4. Click on the **Charts** button. Select “None”
5. Click “Continue”
6. The **Statistics** and **Format** buttons allow you to change the output produced. For this example, no changes from the default values are required
7. Click “OK”. The Output window will display both the number and percentage of customers with different values for the REGION variable



The Table above is an output of the Frequency run on the Region variable of PCs Unlimited, it clearly shows the customer distribution of the company. The highest being in Midwest region, followed by Northeast and so on.

**2. Descriptives**

The “Descriptive” procedure produces summary statistics such as mean, standard deviation, min and max values, sum and range. These statistics are useful for variables measured on a continuous scale – such as dollars, years, units, pounds, or meters.

Suppose you want to know the total dollars spent by all customers and the average dollars spent per customer as well as the largest and smallest amounts spent by any customer. You also want the same summary statistics for number of purchases

1. From the pull-down menus, select **Analyze / Descriptive Statistics / Descriptive**
2. In the dialog box that will appear, select TOTDOL and NUMPURCH for the variable list
3. Click on the **Options** button. Check the “Mean,” “Sum,” “Minimum,” and “Maximum” boxes, and remove the checks from the remaining boxes
4. Click “Continue”
5. Click “OK”. The Output window will display following result:



The Table above is the result after a Descriptive Statistics is run; it shows the Number of Observations (N) as 1000. Since we are measuring the $ spent, we have the result showing that the Minimum $ spent by a customer is $50 and Maximum is $8763, the Average (Mean) spending by a customer is $335.23. Similar interpretation is applicable for Number of Purchases.

**3. Crosstabs**

By forming two-way (or multi-way) tables, the “Crosstabs” procedure count records for two (or more) categorical variables. In addition, statistical tests (including the commonly used chi-square test) can be used to test for significant associations between categorical variables.

For example, if you want to know what percentage of females vs. males responded to the latest offering from PCsUnlimited, we need to calculate four percentages: (1) % of females who responded; (2) % of females who did not respond; (3) % of males who responded; and (4) % of males who did not respond.

(1) % Of females who responded

(2) % Of females who did not respond

(3) % Of males who responded and

(4) % Of males who did not respond

1. From the pull-down menus, select **Analyze / Descriptive Statistics / Crosstabs**
2. In the dialog box that appears, select GENDER for the Row variable and REPOND for the column variable
3. Click “Cells.” Click on “**Row percentages**”. This will compute the percentage of males and females who responded
4. Click “Continue” and then click “OK” to produce the crosstabs listing



The Table above gives us the % of responders and non-responders by **Rows** that is why we selected the **Row** % option. The table can be interpreted as out of 275 females only 43 (15.6%) responded and 232 (84.4%) did not; where as out of 725 males 93 (12.8%) responded and the rest 632 (87.2%) did not.

Now, suppose we want to know whether the higher percentage of females who responded to an offer (15.6%) than males (12.8%) is statistically significant. We can repeat the analysis adding the following:

* Click on the “Statistics” box at the bottom of the Cross-tabs dialog box
* In the dialog box that appears, click on Chi-square to produce the following:



The p-value for the **Pearson Chi-square1 is .247**, which is well above the **typical .05** cutoff, indicating that the difference in response rate between males and females is **not statistically significant.**

Finally, suppose that instead of finding the % of females vs. males who responded, we want to know the % of responders and non-responders who were male or female. In other words, we want to know:

(1) % Of responders who are female

(2) % Of responders who are male

(3) % Of non-responders who are female and

(4) % Of non-responders who are male

To calculate these values, use the above procedure, with one small change: Instead of clicking on “Row percentages,” click on “Column percentages”. You can also create a crosstabs listing with **both** row and column percentages, though some find this confusing!



The Table above shows that there are 275 (27.5%) females in the database. Out of 136 responses 43 were female (31.6%) and 93 (68.4%) were males and similar interpretation for the ‘no’

It doesn’t matter which variable is the row variable and which is the column variable – although you will need to think through whether you want to request row or column percentages depending on which is your row and which is your column variable.

Hint: if you are cross-tabbing two variables and one has many categories – it is best to make that one the row variable. For example, consider a cross-tabs of gender by profession (and you have 25 professions). That will either produce a table with 2 columns and 25 rows or one with 25 columns and 2 rows – and the first (2 columns and 25 rows) will fit on a sheet of paper much better!

**3.Bivariate correlation**

One measure of the relationship between continuous variables is their correlation. To compute the correlation between total dollar purchases and number of purchases:

1. From the pull-down menus, select Analyze / Correlate / Bivariate
2. Select NUMPURCH and TOTDOL for the Variables list
3. Under “Correlation Coefficients,” select “Pearson”
4. If you want statistically significant correlations marked with an asterisk, select “Flag significant correlations
5. Click “OK” to produce the output



**4.3. Graphs**

1. Histogram

The “Descriptive” procedure gives us information about the distribution of the values of a variable. As we saw, we can find the minimum and maximum values and the mean; we can also display the variance and standard deviation, to give us some idea of how tightly values are clustered around the mean.

It is often more useful, though, to see a histogram showing the distribution of the values. There are two ways to create histograms in SPSS:

(1) Using the histogram option under the Charts menu, or

(2) With the “Frequencies” procedure under the Analyze menu

The following examples show how to create a histogram showing the distribution of customer ages.

(1) Using histogram option:

1. Open the dataset in SPSS

2. From the pull-down menus, select Graphs / Histogram

3. In the dialog box that will appear, select AGE for the variable

4. Click on the Titles button if you want to add a title

5. Click “OK”. A histogram showing the distribution of AGE will appear in the Output window. The mean and standard deviation for AGE is also printed:



(2) Using the frequencies procedure (this option allows you to get frequency tables as well as a histogram):

1. From the pull-down menus, select Analyze / Descriptive Statistics / Frequencies
2. In the dialog box that appears, select AGE for the variable list
3. Click on the Charts button. Select “Histograms”
4. Click “Continue”
5. The “Display Frequency Tables” box is probably checked. Click this box to “un-select” this option (if you forget to do this you will get a long – and probably useless – frequency table)
6. The Statistics and Format buttons allow you to change the output produced.
7. Click “OK”. A frequency histogram showing the distribution of AGE will appear in the Output window. As before, the mean and standard deviation for AGE will also be printed
* Bar Charts

Bar charts are commonly used to provide a visual summary for categorical data. For example, to create a bar chart showing the number of customers in each region:



**5. Helpful Hints**

**5.1. Analyzing a Subset**

In some cases, you may only want to analyze a subset of the total data collected. For example, you may only be interested in women's opinions on a certain matter. You can tell SPSS to only analyze that group.

* 1. Click "Data" on the top menu line, then click "Select Cases" from the drop down menu.
	2. In the resulting dialogue box, click the "if condition is satisfied" check box.
	3. Click the "If' button. In the resulting dialogue box, highlight the variable that you want to split into a subset and click the arrow.

o Complete the If Statement. For example, if females are coded as "1 " type or click "= 1” Therefore you should have a statement that reads, gender (if that’s how you named the variable) = 1.

o Click "Continue."

o Click "OK."

o **VERY IMPORTANT** - this restriction will remain in place until you go through this process again and click the "All Cases" check box.

**5.2. Recoding a variable into a new variable**

Suppose you want to divide your customers into two groups: one-time purchases and repeat customers or those who have purchased more than once.

An easy way to do this in SPSS is to use the **Transform / Recode/ Into Different Variables** to create a new Yes/No variable:

1. Make sure you are in the Data Editor window.
2. From the pull-down menus, select Transform / Recode / Into Different Variables.
3. Select NUMPURCH for “Input Variable -> Output Variable.”
4. Under ‘Output variable,’ type a name for the new variable (perhaps REPEAT) and a brief description. Click “Change.”
5. The “Numeric Variable -> Output Variable” box should show:

NUMPURCH -> REPEAT

1. Click the “Old and New Values” box.
2. Click the “Output Variables are Strings” box on the lower right-hand side of the dialog box.
3. Under “Old Value,” click “Value” and type a 1 in the box.
4. In the “New Value” box, type No.
5. Click “Add.” This specification will set REPEAT to ‘No’ whenever NUMPURCH is equal to one (i.e., the customer made only one purchase
6. Under “Old Value” click “Range: \_\_\_ through highest” and type a 2 in the box.
7. In the “New Value” box, type Yes.
8. Click “Add.” This specification will set REPEAT to ‘Yes’ whenever NUMPURCH is 2 or greater (i.e., the customer has made 2 or more purchases). (14)
9. The “OldÆNew” box should show two lines:
	1. 1->’No’
	2. 2 through Highest->’Yes’
10. Click “Continue.”
11. Click “OK.” The new variables will be added. Depending on the file size, this can take from a few seconds to several minutes.

**5.3. Transferring Data from Excel into SPSS**

To open a file in SPSS from Excel:

1. In Excel, make sure the first row includes the names of each variable (e.g., Name,

Gender, etc..).Format the first row as text.

2. In Excel, format all the numbers as numbers and any text as text.

3. Save and close the file.

4. In SPSS, go to File⇨Open. Select the folder to “Look in” and choose File Type -

Excel (\*.xls).

5. Choose your file to open. Click Open.

6. A window will open. Click on "Read variable names"

7. Click OK and the file will open on SPSS

**5.4. Conclusion**

Understanding SPSS is generally a matter of just sitting down and using it for awhile. Yes, it may be a little painful at first, but after a short time, you will be able to use the program fairly well. *Remember*, the help menus in SPSS are extremely useful

**6.Special Topics**

**6.1 RFM Analysis**

RFM analysis is a customer segmentation method commonly used in the CRM (Customer Relationship Marketing) field to identify potential profitable customer segment.

In, RFM analysis, R stands for Recency, F stands for frequency, M stands for monetary. So RFM analysis uses previous customer purchase recency, purchase frequency and total purchase amount to identify which segment is profitable to market to. Here is how RFM analysis is done is SPSS.

(Here we will be using the Tuscan Lifestyle dataset from MK726 course and we will be using quantile in the RFM analysis, so the customer will be split into 5 category)

**Quintile Analysis for RFM**

1. To group observation by Recency, Go **Transform** - **Rank Cases**



1. Place the variable that is related to R,F,M (Days since prior purchase, Total life-to-date # orders, Total life-to-date dollars) under the Variables box
2. Assign rank
	1. For R Check Assign Rank 1 to Largest Value to rank customer with most recent purchase as highest rank
	2. For F and M Check Assign Rank 1 to Smallest Value to rank customer with the least purchase amount/ dollar
3. Click Rank Type



1. Check **Ntiles** and put in 5 for deciles and click ok to generate new variable and rename the new variables as R, F, M
2. Create RFM variable by Click **Transform**-> **Compute variable**
	1. Create a new Variable name RFM and enter R\*100+F\*10+M in the **Numeric Expression box**

**6.2 Logistic Regression**

Logistic Regression is the counter part of linear regression. However, logistic regression is used to predict a dependent variable that is categorical(binomial) Here is how RFM analysis is done is SPSS. (Here we will be using the Customer dataset from MK726 course as an example)

1. Go to **Analyze->Regression->Binary Logistic**
2. Place intended dependent variable (*BUYER*) under **Dependent**



1. Place the intended independent variables under **Covariates**
2. Press the **Save** in the window



1. Check the **Probabilities Box** in under **Predicated Values**
2. Press **OK** to run the regression
3. Charts should be outputted in the SPSS output and a new variable PRE\_1 should be generated as the predicated probability.

**6.2 Decision(classification) Tree Analysis**

Decision tree is an advance analytic topic that classified as supervised machine learning method. The decision tree algorism is to form a tree like structure for prediction that rather than a linear structure in linear and logistic regression and the analysis will rely on tree punning to prevent model overfitting. Here is how to do Decision tree in SPSS. (In order to do decision tree, you have to have the **SPSS Premium Grad Pack** installed, and I will be using Customer dataset from MK726 course as an example)

**How to grow decision tree**

1. First go to **Analyze- Classify-Tree**

2. Place intended dependent variable (BUYER) in the Dependent Variable Box

3. Place the intended variable in the Independent Variables Box

4. Select intended Growing Method CRT under Growing Method

5. Click Categories and check the box next to yes to select Yes as the targeted category



6. Click Continue to exist to the Decision tree tab and click Criteria



1. In the **Growth Limits** Tab, under **Maximum Tree Depth** Make sure **Automatic** is selected
2. Under Minimum Number of Cases box Make sure that Parent Node is limited to 100 and Child Node is limited to 50
3. **Go to Pruning** make sure to check the box for: Prune tree to avoid overfitting



1. Click continue to generate the decision tree