

Marketing Research with IBM® SPSS Statistics

A Practical Guide

KARINE CHARRY
KRISTOF COUSSEMENT
NATHALIE DEMOULIN
NICO HEUVINCK

A Gower Book

Marketing Research with IBM® SPSS Statistics

Marketing researchers, companies and business schools need to be able to use statistical procedures correctly and accurately interpret the outputs, yet generally these people are scared off by the statistics behind the different analyses procedures, thus they often rely on external sources to come up with profound answers to the proposed research questions. In an accessible and step by step approach, the authors show readers which procedures to use in which particular situation and how to practically execute them using IBM® SPSS Statistics. IBM® is one of the largest statistical software providers world-wide and their IBM® SPSS Statistics software offers a very user-friendly environment. The program uses a simple drag-and-drop menu interface, which is also suitable for non-experienced programmers. It is widely employed in companies and many business schools also use this software package. This straightforward, pragmatic reference manual will help: professional marketers who use statistical procedures in IBM® SPSS Statistics; undergraduate and postgraduate students where marketing research and research methodology are taught; all researchers analysing survey-based data in a wide range of frontier domains like psychology, finance, accountancy, negotiation, communication, sociology, criminology, management, information systems, etc. IBM®'s next-generation business analytic solutions help organizations of all sizes make sense of information in the context of their business. You can uncover insights more quickly and easily from all types of data-even big data-and on multiple platforms and devices. And, with self-service and built-in expertise and intelligence, you have the freedom and confidence to make smarter decisions that better address your business imperatives.

Karine Charry is Professor of Marketing at the Catholic University of Louvain-Mons (Belgium). Her research focuses on consumer behaviour – and namely children as consumers – as well as persuasion mechanisms in marketing and health prevention communications.

Kristof Coussement is Professor of Marketing Analytics, Academic Director of the MSc in Big Data Analytics for Business, and Co-director of the research centre for marketing analytics at IÉSEG School of Management (LEM-CNRS) of the Catholic University of Lille in France. Dr Coussement teaches several marketing-related courses including 'Strategic Marketing Research', 'Customer Relationship Management' and 'Database Marketing' in which students are taught the theoretical principles of all aspects of marketing research, operational and analytical CRM and the methodological foundations of predictive marketing modelling.

Nathalie Demoulin is Associate Professor of Marketing at IÉSEG School of Management (LEM-CNRS), the Catholic University of Lille in France. She teaches several courses related to relationship marketing and marketing strategy such as, 'Satisfaction and Loyalty', 'Relationship Management and CRM', 'CRM and Loyalty Programs' and 'Marketing Strategy and Company Observation'.

Nico Heuvinck is Professor of Marketing at IÉSEG School of Management (LEM-CNRS). He teaches several marketing courses including 'Marketing Research', 'Strategic Marketing Research', 'Marketing Research Methodology – Experimental Designs' and 'Neuromarketing'.

Marketing Research with IBM® SPSS Statistics

A Practical Guide

Karine Charry, Kristof Coussement,
Nathalie Demoulin and Nico Heuvinck

First published 2016
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

and by Routledge
711 Third Avenue, New York, NY 10017

Routledge is an imprint of the Taylor & Francis Group, an informa business

© 2016 Karine Charry, Kristof Coussement, Nathalie Demoulin and
Nico Heuvinck

The right of Karine Charry, Kristof Coussement, Nathalie Demoulin and
Nico Heuvinck to be identified as the authors of this work has been
asserted by them in accordance with sections 77 and 78 of the Copyright,
Designs and Patents Act 1988.

All rights reserved. No part of this book may be reprinted or reproduced or
utilised in any form or by any electronic, mechanical, or other means, now
known or hereafter invented, including photocopying and recording, or in any
information storage or retrieval system, without permission in writing from the
publishers.

Created with IBM® Software. © Copyright IBM Corporation 1994, 2015.
All rights reserved.

Trademark notice: Product or corporate names may be trademarks or registered
trademarks, and are used only for identification and explanation without intent
to infringe.

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication Data

A catalog record for this book is available from the Library of Congress

ISBN: 9781472477453 (pbk)

ISBN: 9781315525532 (ebk)

Typeset in Sabon
by Out of House Publishing

*To my lovely wife Ilse for her endless support;
To my little funny rascals, Tobias, Oliver and Bastian;
To my dear parents, Carline and Wim.*

Kristof Coussement

To my close family members for their endless support.

Nathalie Demoulin

*To my soulmate and wife Charlotte for your love, patience and support;
To my lovely girls Noémie and Alizée who spice up my life;
To my family, friends and colleagues who make life exciting.*

Nico Heuvinck

Constantia et labore. Bis vincit qui se vincit.

This page intentionally left blank

Contents

<i>Author Biographies</i>	viii
<i>Foreword</i>	x
<i>Preface</i>	xii
1 Getting Started with IBM SPSS Statistics	1
2 Descriptive Analysis	31
3 Exploratory Factor Analysis	48
4 Cluster Analysis	68
5 Hypothesis Testing	91
6 Correlations	170
7 Regression Analysis	175
8 Moderation and Mediation Analysis	229
<i>Index</i>	248

Author Biographies

Karine Charry (PhD) is Professor of Marketing at the Catholic University of Louvain-Mons (Belgium). Dr Charry's research focuses on consumer behaviour – namely, children as consumers – as well as persuasion mechanisms in marketing and social marketing (health promotion, eco-friendly behaviours, etc.). Her work has been published in French and international peer-reviewed journals (*International Journal of Advertising*, *Journal of Business Ethics*, *Recherche et Applications en Marketing*, etc.). Her ten years of experience in marketing departments of diverse companies and sectors in B-to-B and B-to-C contributes to the pragmatic approach of her publications and teaching ('Consumer Behaviour', 'Persuasion in Marketing Communication' and 'Social Marketing').

Kristof Coussement (PhD) is Professor of Marketing Analytics, Academic Director of the MSc in Big Data Analytics for Business, and Co-director of the research centre for marketing analytics at IÉSEG School of Management (LEM-CNRS) of the Catholic University of Lille in France. Dr Coussement teaches several marketing-related courses including 'Strategic Marketing Research', 'Customer Relationship Management' and 'Database Marketing' in which students are taught the theoretical principles of all aspects of marketing research, operational and analytical CRM and the methodological foundations of predictive marketing modelling. Dr Coussement has had papers published in international peer-reviewed journals and his works have been presented at various conferences around the world. His main research interests are all aspects in customer intelligence, B-to-B intelligence, direct marketing and analytical CRM. While he has improved his 'practical' experience over the years by doing several real-life research projects in a number of industries, his main focus is on doing profound academic research with a high added value to business. More information about his work can be found at www.kristofcoussement.com.

Nathalie Demoulin (PhD) is Associate Professor of Marketing at IÉSEG School of Management (LEM-CNRS), the Catholic University of Lille in France. She teaches several courses related to relationship marketing and marketing strategy such as, 'Satisfaction and Loyalty', 'Relationship Management and CRM', 'CRM and Loyalty Programs' and 'Marketing Strategy and Company Observation'. Her primary research interests were the decision-making process of marketing managers and the impact of marketing decision support systems on managers. She currently conducts research linked to customer loyalty, the store environment and the

adoption of new technologies in the retail sector. She has published in international and French peer-reviewed journals such as *Decision Support Systems*, *Journal of Retailing*, *Journal of Retailing and Consumer Services*, *International Journal of Retail and Distribution Management* and *Systèmes d'information et Management*.

Nico Heuvinck (PhD) is Professor of Marketing at IÉSEG School of Management (LEM-CNRS) of the Catholic University of Lille in France. He teaches several marketing courses including 'Marketing Research', 'Strategic Marketing Research', 'Marketing Research Methodology – Experimental Designs' and 'Neuromarketing'. His research focusses on various consumer psychology-related phenomena such as attitude ambivalence, two-sided messages, goals and motivation, store atmospherics and feelings of nostalgia. He has published his research in international journals such as the *Journal of Consumer Research*.

Foreword

By Cédric Mulier, IBM Business Analytics Solutions Country Manager

Companies continually face the challenge of adapting and modifying their marketing plans according to the dynamics and the diversity of their environments. Furthermore, these surrounding environments have become more data-driven. Many companies do consider data the new oil. To master this new oil, we do consider that you need to set up the right strategy according to various parameters.

Volume is the change of paradigm.

The size of available data is increasing every day and at such a tempo that every day we are producing more than one year of data in the 90s.

Variety is crucial to consider.

Data exists in different formats, all of which can be very useful for marketing research. Data can be structured (e.g. demographic customer data), unstructured (e.g. answers to open-ended questions) or semi-structured (e.g. Web data). Current techniques needed to get access to these data, the ways to extract them, the best channel to refine them and eventually to consume them will be need to be adapted. Thus understanding the specificity of these sources is of utmost interest for marketing research.

Veracity needs to be validated.

Checking that these data are correct, represent a point of view and that it can be verified are musts.

Velocity is the name of the game.

With the Internet of Things era and the ever-increasing adoption of social media by all stakeholders, the speed at which each of us can potentially be confronted with data is exponential.

These are typically the 4 Vs of big data. However, I would like to add one more criterion and this one should always be present regardless of the timeframe, geographies and functional scope of every marketing research project, i.e. *Value*.

Value should be at the cornerstone of any data strategy, regardless of the volume, the variety, the velocity and the veracity of data. Instead of 'Big data', we should all talk about the 'All Data that can bring Value'.

Any data strategy is thus a combination of these 5 Vs with Value being the most searched one in marketing research. Market researchers have a crucial role to make the most of this wealth of data and optimize the value output. They are the ones who will enable this strategy. They are the ones who can provide context to the data. They

are the ones that can be the centre of the expertise making sure that best practices are shared between the various departments of a company.

At IBM, we have decided to lower the barrier of entry to our solutions: For a very large part of the last five years, 16 billion USD has been invested in ensuring this data-overflow barrier to entry is lowered through the right interface, right ergonomics and ease of working with our products amongst IBM SPSS Statistics. The ultimate goal is to ensure a broad adoption of the analytics solutions.

In the same spirit, we have decided to collaborate extensively with universities and business schools worldwide to ensure that our youth, our next-to-come marketing analysts have access to the best-performing IBM technologies. They can train themselves on the latest solutions and bring value directly to companies and organizations as they are entering into their professional life. Our collaboration with IÉSEG School of Management is a nice example of how we at IBM want to speed up the analytics journey for our society.

With this marketing research book, the authors are enabling and helping the whole marketing research community to better grasp this wealth of value. Allow me here to thank them for this contribution and wish them the best success that this tutorial-based book deserves.

Preface

Nowadays a lot of researchers involved in marketing face the problem of correctly using statistical procedures and accurately interpreting the outputs. Usually these people are scared off by the statistics behind the different analyses procedures and thus often rely on external sources to come up with sound answers to the proposed research questions. This book intends to show its readers how to select the right statistical procedures and how to put these methods into practice by always starting from a real managerial problem. It shows its readers, through a step-by-step approach, which procedures to use in which particular situation and how to practically execute them in an IBM SPSS Statistics environment. It offers a very user-friendly environment for executing marketing research projects. This software uses a simple drag-and-drop menu interface, also suitable for non-experienced users. It is widely employed by companies, universities and business schools.

The purpose of this book is straightforward: it offers a pragmatic approach based on real-life marketing research examples to help the reader solve their day-to-day (business) problems. Furthermore, a complete section is dedicated to the managerial interpretation of the results.

This book is aimed at several target audiences, all of whom need robust answers to existing business problems.

- This book intends to be a reference manual for all professional marketers who would like to use statistical procedures in IBM SPSS Statistics. Consequently, the manual does not only give an overview of the basic options for the statistical tests used, but it also digs deeper into more specific and detailed options.
- This book is suitable for all undergraduate and postgraduate academic programmes in which *Marketing Research* and *Research Methodology* are taught.
- This book is also suitable for all researchers analysing survey-based data in a wide range of frontier domains such as psychology, finance, accountancy, negotiation, communication, sociology, criminology, management, management information systems and so on.

The statistical procedures considered in this book refer to the most common marketing issues encountered by the various target audiences listed above. But in order to enable novice users of IBM SPSS Statistics to feel empowered, Chapter 1 is devoted to a thorough description of the software. First, we propose a tour of the environment and different data preparation steps. Chapter 2 is devoted to descriptive statistics and their

usefulness. Chapter 3 and Chapter 4 consider exploratory procedures: Exploratory Factor Analysis and Cluster Analysis, respectively. The next chapters discuss the confirmative statistical tests. Chapter 5 is devoted to hypothesis testing for parametric and non-parametric data, Chapter 6 explains the relevancy of correlations, Chapter 7 shows how to run Regression Analyses (linear and logistic). Chapter 8, the last chapter, digs into the frequently used Moderation and Mediation Analysis.

Chapters referring to the various statistical procedures (Chapters 4, 5, 6, 7 and 8) could be read independently of each other. According to the type of analysis one has to consider, one may limit the reading to the relevant chapter. However, the first two chapters are recommended for readers new to IBM SPSS Statistics and/or statistics.

All the statistical procedures mentioned above are explained using the same pedagogical scheme. As such, the reader will become familiar with the methodological and mental process of solving a particular marketing research problem. For most analyses, the following structure is used within the book.

Fundamentals

Managerial Problem

Translation of the Managerial Problem into Statistical Notions

Hypotheses

Dataset Description

Data Analysis

Interpretation

Managerial Recommendations

In the *Fundamentals* section, the objective of the statistical procedure is explained. The managerial situations in which the analysis can be used are presented. This section also communicates the important steps required to successfully lead the analysis. The section *Managerial Problem* describes a real-life managerial problem with which every researcher or manager could be confronted. The reader is guided on how to solve the problem him/herself. Once the managerial problem is clearly defined, it is translated into the description of the statistical purpose of the analysis in the section *Translation of the Managerial Problem into Statistical Notions* without formally using symbols or statistical formulas. In other words, the research question is translated using statistical terminology. This is a necessary step that enables choosing the appropriate statistical procedure, but it is kept very simple to facilitate understanding. Providing in-depth theoretical explanations of statistical issues is beyond the scope of this book (nevertheless, a few references are provided at the end of each chapter for the diggers). Furthermore, a statistical representation of the statistical problem is proposed by translating it into a null and an alternative hypothesis in the section *Hypotheses*. In the section *Dataset Description*, a detailed overview of the data delivered is given the name of the file, number of observations, descriptions of the variables and the measurement scale of the variables used. All datasets may be obtained by contacting one of the authors. At this stage, the readers should completely comprehend all elements, including the business problem and the way to solve it. The next step is data analysis. The section *Data Analysis* will show through a step-by-step approach how to perform the statistical procedure in IBM SPSS Statistics. Readers of this book will be taken by the hand and shown how to run a procedure using multiple

screenshots of the software environment. This will drastically enhance the readability of the book. Based on the outputs of the statistical program, guidance is proposed to facilitate interpretation of the results. The interpretation is done in a fragmented approach in which the title of each table or the header of each figure is stated, followed by the corresponding IBM SPSS Statistics output in the section *Interpretation*. In the last stage, the statistical output is converted into a detailed answer on the managerial problem in the section *Managerial Recommendations*.

We would like to thank the various people and institutions that directly or indirectly contributed to this book project and our students, whom we constantly had in mind while working on this book.

Dr Karine Charry
Dr Kristof Coussement
Dr Nathalie Demoulin
Dr Nico Heuvinck

Getting Started with IBM SPSS Statistics

Objectives

1. Introduce the IBM SPSS Statistics environment.
2. Understand the structure of the IBM SPSS Statistics work space.
3. Learn how to create and import an IBM SPSS Statistics data file.
4. Learn basic data manipulation tasks like dealing with missing or invalid data, selecting data, splitting data, sorting data, recoding variables and calculating summated scales.

1.1. What Is IBM SPSS Statistics?

IBM SPSS Statistics provides a graphical interface that helps the marketer exploit the power of a marketing research tool and publish dynamic results in a Microsoft Windows client application. This solution is the preferred interface for business analysts, programmers and statisticians and it is the key marketing research application in business.

The benefits of using IBM SPSS Statistics are threefold:

1. *Provide a self-service environment for analysts and statisticians.* IBM SPSS Statistics integrates the extensive array of analytics with an efficient, friendly graphical user interface application. Business analysts can produce the analyses and distribute reports they need, freeing IT to focus on other strategic projects.
2. *Provide easy access to data sources through a graphical interface.* IBM SPSS Statistics is the only front end that provides a guided mechanism to access data across multiple platforms, operating systems and databases. A centralized system for managing access to corporate data ensures that users have appropriate access privileges while empowering them to react quickly to evolving business conditions.
3. *Make reporting and analytics available to everyone.* The ability to develop and deploy customized tasks enables users to extend the core functionality to create custom wizards, which can be distributed easily as needed. Information can be delivered through an established publishing framework with the ability to publish dynamic, interactive content to Microsoft Office and Web users.

1.2. Software Requirements

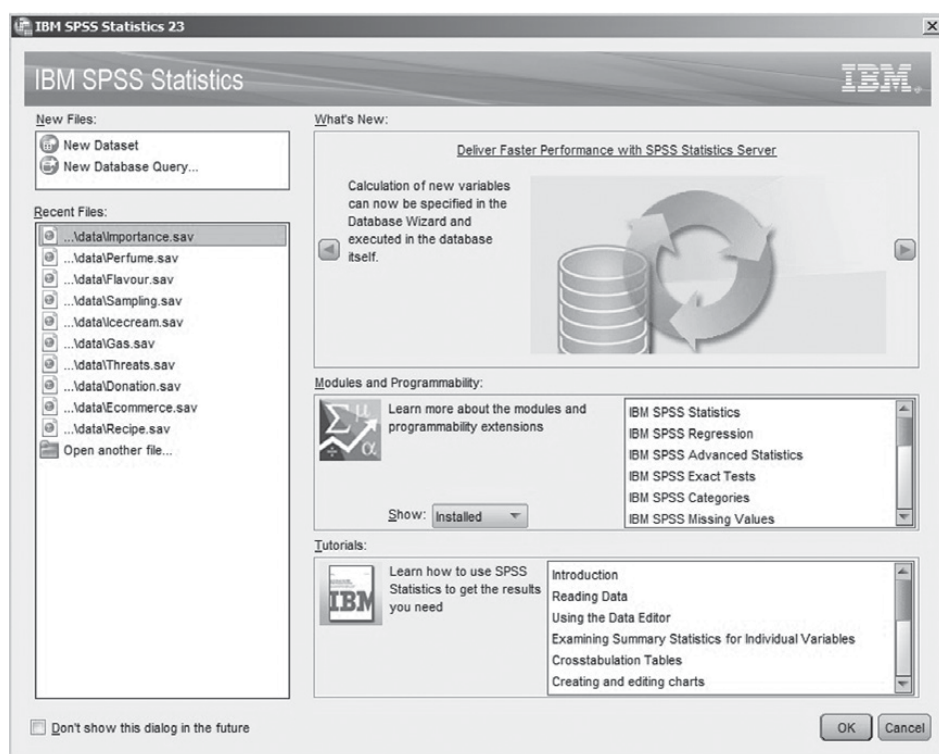
This book assumes that the technical environment has already been set up. Information about the IT requirements necessary to install IBM SPSS Statistics 23 is given here. This software can run on different operating systems. More detailed information can be found at www-01.ibm.com/software/analytics/spss/products/statistics/.

The examples in this book are shown using IBM SPSS Statistics 23, running under Windows 7 Professional. Slight differences may be observed compared to previous and future releases of IBM SPSS Statistics.

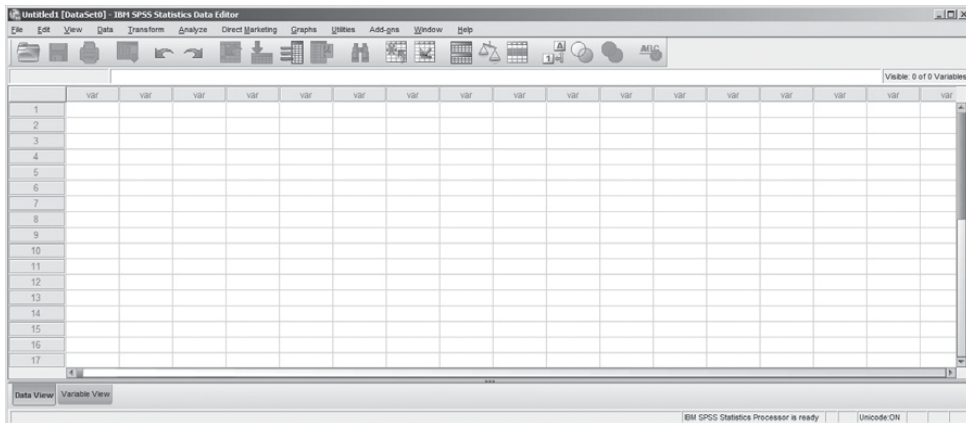
If you have any additional questions or remarks concerning the technical requirements of running IBM SPSS Statistics on your computer, we advise you to contact your local IBM SPSS representative or to visit www-01.ibm.com/software/analytics/subscriptionandsupport/spss.html.

1.3. Touring the Environment

When one launches IBM SPSS Statistics, the software opens with a dialog window as shown below.



At this step, the user has the choice between opening recently used files on the computer or creating a new file. When the user selects **New Dataset** in the **New Files** box, and then clicks **OK**, a data input screen is launched as shown below. The latter can also be achieved by simply clicking **Cancel**.



The default IBM SPSS Statistics interface consists of two tabs, **Data View** and **Variable View**.

- The **Data View** tab is used to enter and explore the data into the data editor.
- The **Variable View** tab describes the various characteristics of the variables used in the dataset. The **Variable View** tab is opened by default.

The active tab is indicated by an orange tab label at the bottom of the screen. To open a tab, the user just has to click the right tab label.

1.4. Entering Data in IBM SPSS Statistics

The IBM SPSS Statistics environment facilitates the creation of SPSS data files (i.e. .sav format) by *encoding manually* or by *importing data* from existing data files.

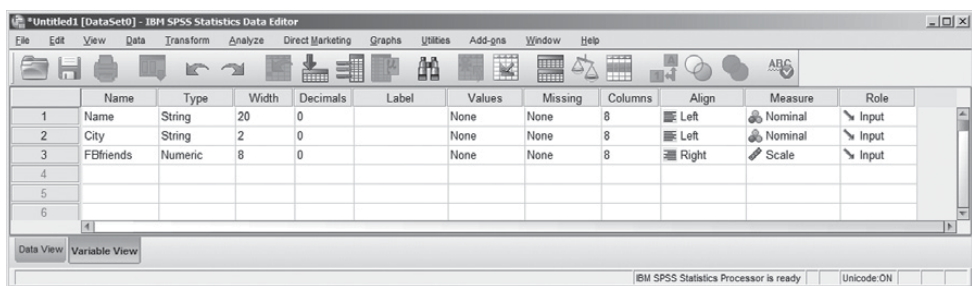
Manually entering data is very useful, and the encoding of a survey could be performed directly in the IBM SPSS Statistics environment. More specifically, it is possible to manually input different values for different variables that correspond, for instance, to the answers of respondents on survey questions. Suppose that a researcher collected information on the name of the respondent (*Name*), the place where they live (*City*) and the number of Facebook friends they have (*FBfriends*) as shown in the table below. After collecting these data, the researcher wants to get these data into the IBM SPSS Statistics environment to get them analysed.

<i>Name</i>	<i>City</i>	<i>FBfriends</i>
Melanie	NY	326
Tiffany	LA	221
Mary	LA	758
Susan	NY	444
Lisa	LA	658
Hellen	NY	238
Betty	NY	112

Name	City	FBfriends
Ed	LA	221
James	NY	587
Joe	LA	1165
John	NY	931

The following procedure is used for manually entering the data into the IBM SPSS Statistics environment.

Make sure that the **Variable View** tab is active by clicking on it. Here the user is able to enter all characteristics of the different variables included in the dataset. The rows in the **Variable View** tab represent the variables, while the variable attributes for each variable are summarized in the columns.



Each variable included in a dataset needs to get a name. If not, IBM SPSS Statistics uses default names (*VAR00001*, *VAR00002*, etc.) which make the execution and interpretation of statistical analyses afterwards much harder. In the first column (**Name**) of the **Variable View** tab, the user may enter or modify the name of the variables. In each dataset, one is free to choose the names of the variables. However, there are some restrictions:



- The name of the variable has to begin with a letter and not a number.
- Do not use special characters or symbols (such as *, °, -, etc.) within the variable name, with the exceptions of €, \$, @, _ and #.
- Do not use blank spaces.
- Every variable name should be unique. Duplication is not allowed.

The second column (**Type**) represents the variable type. In the example, the variables *Name* and *City* are given a string format as they include text, whereas a numerical format has been assigned to the number of Facebook friends someone has, *FBfriends*. Indicating the correct variable type can be done by clicking the relevant cell, and then by clicking on the ... that will appear at the right-hand side of the cell.

The third column (**Width**) can be used to change the number of characters that will be shown for the relevant variable in the **Data View** tab. Default width is 8 characters but this can be changed for string variables or numerical variables that include more than 8 characters.

The fourth column (**Decimals**) indicates the number of decimals that are shown for the values in the **Data View** tab.

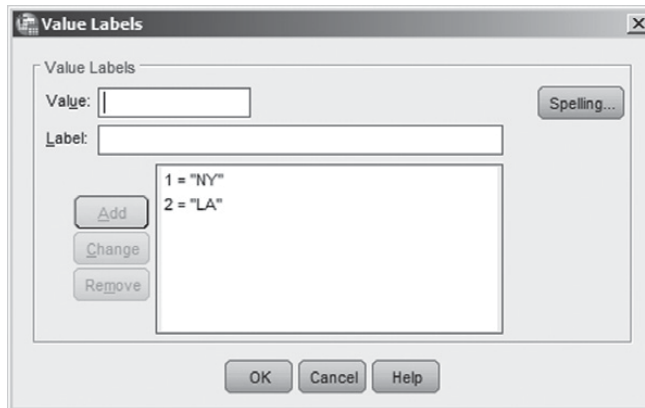


One important remark is that instead of inputting data with words, one has to use numbers in IBM SPSS Statistics. Therefore, the user is advised to use as many numbers as possible in the dataset as the software only treats numbers for further statistical analyses. In the example above, for the variable *City*, only two possible answers exist: LA or NY. It is a good idea to assign, e.g. a value 1 to the group NY and a value of 2 to the group LA. This will enable the user to use this variable afterwards to split up the data file per region, run separate analyses per region, select cases for a specific region, compare the two regions, etc. Three steps are needed to convert the region abbreviations to numbers. The first step will be to replace the words NY and LA in the **Data View** tab manually with the assigned numbers 1 and 2 as shown below. Another option is to use the **Recode** function in IBM SPSS Statistics (cf. below).

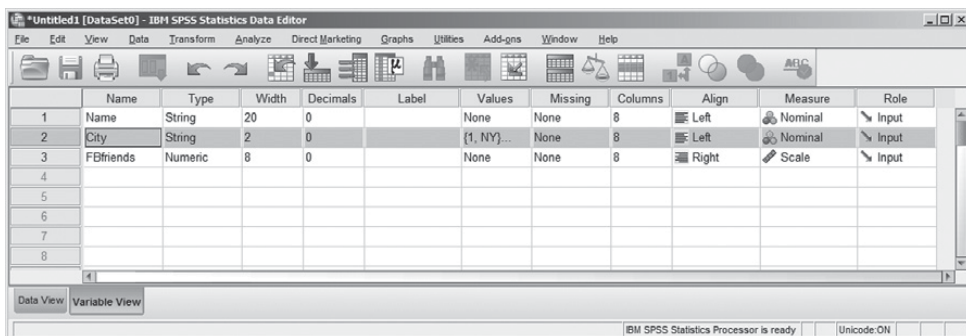
	Name	City	FBfriends	var	var	var	var	var	var	var	var
1	Melanie	1	326								
2	Tiffany	2	221								
3	Mary	2	758								
4	Susan	1	444								
5	Lisa	2	658								
6	Hellen	1	238								
7	Betty	1	112								
8	Ed	2	221								
9	James	1	587								
10	Joe	2	1165								
11	John	1	931								
12											

In the second step, the user needs to correct in the **Variable View** tab the type of data in the **Data** column. For the variable *City*, the user needs to change the format from *String* to *Numeric*.

Finally, in the third step, it is advised to label the new values in order to trace back which number represents which label or group. This facilitates the interpretation of the analyses, especially when one works with a lot of different variables in the dataset. In this example, the output of the statistical analyses needs to show the labels NY and LA instead of 1 and 2. Practically, in the **Variable View** tab, click the **Values** cell for the variable one would like to add labels. In this case, one labels the variable *City*, and thus one clicks on the ... in the **Values** column of the variable *City*. The **Value Labels** dialog window will appear. Start with typing in 1 in the value cell and NY for the label cell and then click **Add**. Repeat this procedure for 2 and LA.



Finally, click **OK** in the **Value Labels** screen and one notices in the **Variable View** tab the changes made for the variable *City* in the columns **Type** and **Values** as shown below.



*When should one use labels? The suggestion is that the user should assign a label to every category for categorical variables such as colour with possible answers 1=red, 2=blue and 3=yellow. For interval scales like seven-point Likert-scales, labels should be assigned at least to the end points of the scale (i.e. 1=totally disagree and 7=totally agree). For ratio scales, i.e. the time people spent in the store, it is not really necessary to assign labels, because many possible answers exist. Note that if there are variables that share identical value labels, labels need not be assigned for each variable separately. The labels of a certain variable in the **Variable View** tab may be copied by right-clicking the values cell of the desired variable, and pasting it into the values cell(s) for the desired other variable(s).*

Data can be manually entered into the IBM SPSS Statistics environment, but it can be also *imported* from a number of different sources such as IBM SPSS Statistics, Microsoft Excel, SAS, Stata, delimited text files, amongst others. This can easily be done via **File** → **Open** → **Data...**. A new pop-up window appears that shows the data file to import.

1.5. Preparing Data in IBM SPSS Statistics

In this section an explanation is given of the techniques that may help the researcher in preparing the data before starting the statistical analyses. The following data preparation methods are illustrated:

- Dealing with missing and invalid data.
- Selecting data.
- Sorting data.
- Splitting data.
- Recoding variables.
- Summated scales with computing new variables.

These dataset preparation tools are often required in traditional marketing research projects. In order to explain these different techniques, please open the dataset *Cleaning.sav*. The following information was requested from 11 respondents:

- Name of the respondent (*Name*).
- City of residence with 1 equal to NY and 2 equal to LA (*City*).
- Number of Facebook friends (*FBfriends*).
- Gender with 1 equal to Male and 2 equal to Female (*Sex*).
- Body weight expressed in kilograms (*Weight*).
- Body length expressed in centimetres (*Length*).
- Attitude towards candy measured by three questions or items on a seven-point semantic differential scale (*AttCandy1*, *AttCandy2*, *AttCandy3*).

1.5.1. Dealing with Missing and Invalid Data

Missing values are often present in marketing research datasets. Respondents do not always answer all the questions because they miss a few, they do not know what to answer, or they simply refuse to give their opinion. When this occurs, the researcher cannot fill in a value in the **Data View** tab for the respective question. As a result, the variable cell remains empty. IBM SPSS Statistics automatically inserts a single period for numerical variables and a blank space for string variables, while it treats this cell as a *System Missing*. The problem with these missing data is that they could affect the statistical results. Therefore, these missing values need to be assigned explicitly to the IBM SPSS Statistics environment.

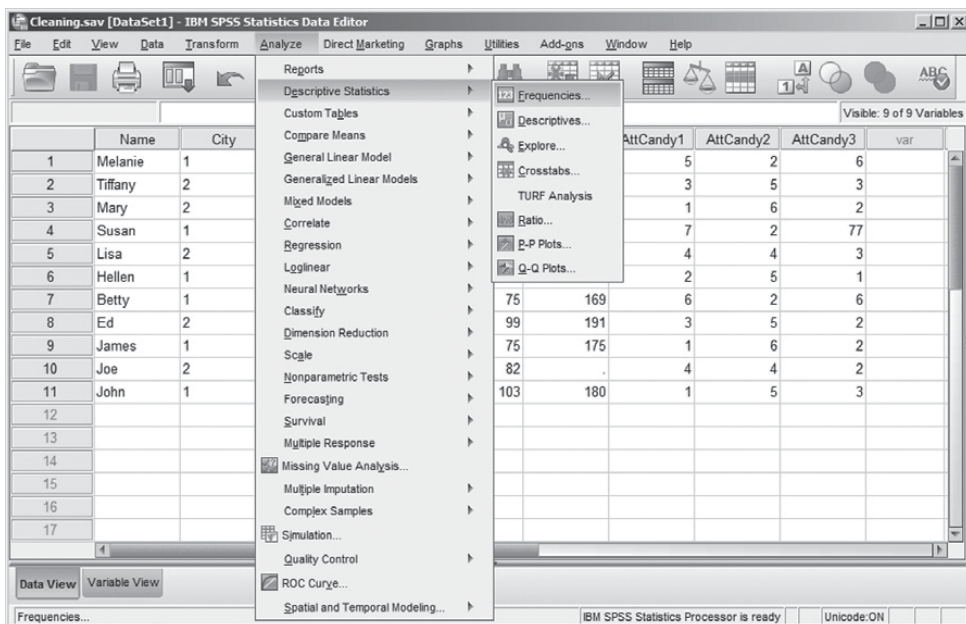
DETECTING MISSING VALUES

In the *Cleaning.sav* dataset, one immediately notices that there is a missing value through a *System Missing* for the respondent named Joe (*Name*) for the variable *Length* as shown below.


	Name	City	FBfriends	Sex	Weight	Length	AttCandy1	AttCandy2	AttCandy3	var	v2
1	Melanie	1	326	2	65	175	5	2	6		
2	Tiffany	2	221	2	60	180	3	5	3		
3	Mary	2	758	2	72	153	1	6	2		
4	Susan	1	444	2	48	164	7	2	77		
5	Lisa	2	658	2	88	178	4	4	3		
6	Hellen	1	238	2	69	164	2	5	1		
7	Betty	1	112	2	75	169	6	2	6		
8	Ed	2	221	1	99	191	3	5	2		
9	James	1	587	1	75	175	1	6	2		
10	Joe	2	1165	1	82	.	4	4	2		
11	John	1	931	1	103	180	1	5	3		
12											

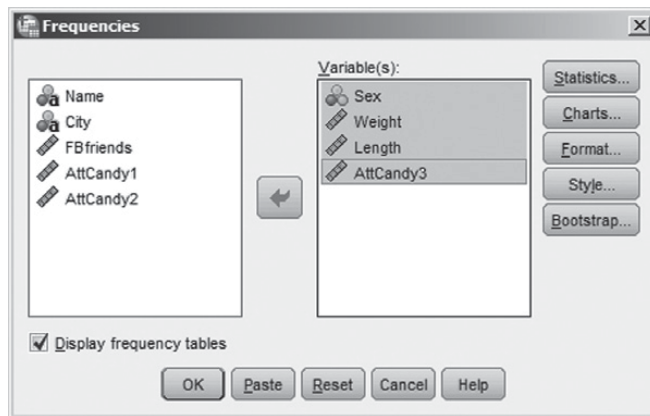
However, in a large dataset, it is not always possible to directly identify missing values at first sight. Therefore it is always a good idea to verify the frequency of missing values for the variables *Sex*, *Weight*, *Length* and *AttCandy3* in the dataset. This can easily be done in IBM SPSS Statistics using the **Frequencies...** task.

1. Open *Cleaning.sav* in the IBM SPSS Statistics environment.
2. Open the **Frequencies...** task via **Analyze** → **Descriptive Statistics** → **Frequencies...** as shown below.



3. The **Frequencies** pane opens. Drag and drop the corresponding variables to test for missing values to the **Variable(s)** section as shown below. Alternatively, select

all the relevant variables in the left box and by clicking the right arrow sign afterwards .



4. Click **OK** to finish the **Frequencies...** task and an output window opens.

The results are shown in the table below.

Statistics					
		Sex	Weight	Length	AttCandy3
N	Valid	11	11	10	11
	Missing	0	0	1	0

In the **Statistics** table, one finds the number of missing values per variable in the row *Missing*. Here the dataset contains one missing value for the variable *Length*.

RESOLVING MISSING VALUES

Once one has discovered variables with missing values, one has to give a clear indication to the IBM SPSS Statistics environment that variables contain missing values. One does this by replacing the missing values in the variable cell with non-representative values, i.e. values that do not occur amongst the possible answers of the variable value range. Commonly used non-representative values are 99, 999, and -1. For instance, the dataset *Cleaning.sav* misses the value for the variable *Length* for the respondent Joe. Here the missing value is replaced by -1 to indicate the presence of a missing value as shown below.

The screenshot shows the IBM SPSS Statistics Data Editor window with the file name '*Cleaning.sav [DataSet1]'. The 'Data View' tab is active, displaying a dataset with 12 cases and 11 variables. The variable 'Length' has a value of -1 for the 10th case (Joe).

	Name	City	FBfriends	Sex	Weight	Length	AttCandy1	AttCandy2	AttCandy3	var	va
1	Melanie	1	326	2	65	175	5	2	6		
2	Tiffany	2	221	2	60	180	3	5	3		
3	Mary	2	758	2	72	153	1	6	2		
4	Susan	1	444	2	48	164	7	2	77		
5	Lisa	2	658	2	88	178	4	4	3		
6	Hellen	1	238	2	69	164	2	5	1		
7	Betty	1	112	2	75	169	6	2	6		
8	Ed	2	221	1	99	191	3	5	2		
9	James	1	587	1	75	175	1	6	2		
10	Joe	2	1165	1	82	-1	4	4	2		
11	John	1	931	1	103	180	1	5	3		
12											

Furthermore, the researcher should explicitly indicate in the **Variable View** tab that -1 concerns a missing value for the variable *Length*. If not, IBM SPSS Statistics will treat the -1 as a regular data point. To indicate that -1 is a code for missing values for the variable *Length*, the user should go to the **Variable View** tab and click ... in the **Missing** column for the *Length* variable. This opens the **Missing Values** pane as shown below. One clicks the **Discrete missing values** option, and fills in -1. One clicks **OK** to finish the task. The IBM SPSS Statistics environment considers now the value -1 for the variable *Length* as a missing value indicator.

The screenshot shows the 'Missing Values' dialog box. The 'Discrete missing values' option is selected. The value -1 is entered in the first input field. The 'OK' button is highlighted.

Note that two possibilities are given to indicate missing values (*System Missings*):

- **Discrete missing values**, i.e. the possibility to indicate three distinct discrete missing value indicators (e.g. -1, 99 and 999).

- **Range plus one optional discrete missing value**, i.e. the possibility to indicate one discrete missing value indicator in combination with a range of missing value indicators. For instance, one could decide that -1 and all values between 980 and 999 should be considered as missing.

In this example, the first option has been chosen and from now on -1 will be considered as a missing value, but only for the variable *Length*. To specify this for other variables as well, copy this in the **Variable View** tab by right-clicking the missing cell of the desired variable and paste it into the missing cell(s) for other desired variable(s).

INVALID DATA

There are many errors possible during the data entry phase that could result in invalid data points. For instance, a researcher could have typed 66 instead of 6 for a given variable. Two strategies exist to detect invalid data points. First, one could manually explore the **Data View** tab trying to detect invalid variable values.

The second option is to calculate a frequency table using the **Frequencies...** task (cf. above), and explore the different data values of the variables. For instance, by investigating the output of the **Frequencies...** task for the variable *AttCandy3*, one notices that although the variable is measured on a seven-point scale, and thus the variable values could only range from 1 till 7, one cell contains a value of 77 as shown in the table below.

<i>AttCandy3</i>		<i>Frequency</i>	<i>%</i>	<i>Valid %</i>	<i>Cumulative %</i>
Valid	Bad	1	9.1	9.1	9.1
	2	4	36.4	36.4	45.5
	3	3	27.3	27.3	72.7
	6	2	18.2	18.2	90.9
	77	1	9.1	9.1	100.0
	Total	11	100.0	100.0	

A possible solution to the problem is searching for the corresponding cell in the **Data View** tab by scrolling down the data points of the variable *AttCandy3*. One could also use the tasks **Select Cases...** or **Sort Cases...** (cf. below how to run these tasks) to find the erroneous cell. Imagine that it concerns a typo, i.e. in the original setup a 7 is indicated, while a value 77 is present in the dataset. The researcher may want to edit the corresponding cell. However, when the original data is not accessible, and it is not clear what went wrong during the data entry stage, it is always a good idea to treat the value of this cell as missing.

1.5.2. Selecting Data

Creating subsets of data is a crucial element in traditional marketing research projects. It is the process of selecting the observations that satisfy one or more conditions. Suppose that one would like to run a statistical analysis on the males, without permanently deleting the data on the females. To temporarily (de)activate observations,