CHANDIGARH UNIVERSITY
Discover. Learn. Empower.

UNIVERSITY
SCHOOL OF BUSINESS

# LAB. MANUAL

# PREDICTIVE ANALYSIS

## BAP-703

## (2018-2020)

## Dr. Shalini Aggarwal

# Vision and Mission of the Department

## Vision of the Department

To create excellence in business management for nurturing value driven business leaders with analytical and entrepreneurial mindset to foster innovative ideas in order to transform the world and serve the society.

## Mission Statements of the Department

M1 : Design a unique competency directed and industry relevant curriculum with outcome oriented teaching learning process facilitated by world class infrastructure.

M2 : Enhance students' cognitive, research, analytical, ethical and behavioral competencies through programs that equip them to meet global business challenges in the professional world.

M3 : Facilitate student centric sound academic environment with co-curricular and extra-curricular activities to groom and develop future ready business professionals.

M4 : Design a transparent evaluation system for objective assessment of the program learning.

M5 : Align meaningful interactions with the academia, industry and community to facilitate value driven holistic development of the students.

M6 : Develop ethical and socially responsible entrepreneurial attitude for harnessing the environmental opportunities through creativity and innovation for a vibrant and sustainable society.

## Program Educational Objectives (PEOs)

Program Educational Objectives (PEOs) are the broad statements that describe career and professional accomplishments that graduates will attain within a few years of graduation. After successful completion of MBA program from Chandigarh University, the graduates will:

**PEO1**: Make significant impact as successful management professionals with a sound business and entrepreneurial acumen leading to a promising career in the various management domains.

**PEO 2**: Develop the professional competence for astute decision making, organization skills, planning and its efficient implementation, research, data analysis and interpretation with a solution finding approach.

**PEO 3**: Be known for their team player qualities to handle diversity and the leadership skills to make sound decisions while working with peers in an inter-disciplinary environment with people of cross-cultural attributes

**PEO 4:** Be adaptable to new technology, innovations and changes in world economy that positively impacts and contributes towards industry, academia and the community at large.

**PEO 5**: Be responsible citizens with high ethical conduct that will empower the business organizations with high integrity, moral values, social effectiveness and legal business intelligence.

## Program Outcomes (POs)

| Program Outcome | After completing the program, the students will be able to: |
|---|---|
| PO1 | Apply knowledge of management theories and practices to solve business problems. |
| PO2 | Foster Analytical and critical thinking abilities for data-based decision making |
| PO3 | Ability to develop Value based Leadership ability |
| PO4 | Ability to understand, analyze and communicate global, economic, legal, and ethical aspects of business. |
| PO5 | Ability to lead themselves and others in the achievement of organizational goals, contributing effectively to a team environment. |
| PO6 | Ability to develop innovative and entrepreneurial mindset. |

**(Batch 2018-2020)**

**Version-2018**

| University School Business,<br>CHANDIGARH UNIVERSITY, GHARUAN | | | | | |
|---|---|---|---|---|---|
| | **Scheme of MBA BATCH (2018-2020)** | | **Total Credits=106** | | |
| **Ist Semester** | | | | | |
| **Subject Code** | **Subjects** | **L** | **T** | **P** | **Cr** |
| BAT- 601 | Accounting for Managers | 4 | 0 | 0 | 4 |
| BAT- 602 | Fundamentals of Management and Organizational Behaviour | 4 | 0 | 0 | 4 |
| BAT- 603 | Managerial Economics | 4 | 0 | 0 | 4 |
| BAT- 604 | Quantitative Techniques for Managers | 4 | 0 | 0 | 4 |
| BAT- 605 | Marketing Management | 3 | 0 | 0 | 3 |
| PCT-610 | Professional Business Communication | 3 | 0 | 0 | 3 |
| PCP-611 | Professional Business Communication (LAB) | 0 | 0 | 2 | 1 |
| BAT- 608 | Computer Applications for Business | 2 | 0 | 2 | 3 |
| BAT- 609 | Supply Chain Management | 3 | 0 | 0 | 3 |
| | **TOTAL CREDITS IN SEMESTER** | | | | **29** |
| | Mentoring Lectures | 2 | 0 | 0 | 0 |
| | **TOTAL NO. OF SESSIONS** | | | | **33** |
| **IInd Semester** | | | | | |
| BAT- 660 | Legal and Business Environment | 3 | 0 | 0 | 3 |
| BAT- 661 | Corporate Finance | 4 | 0 | 0 | 4 |
| BAT- 662 | Operations Management and Research | 4 | 0 | 0 | 4 |
| BAT- 654 | Research Methodology | 4 | 0 | 0 | 4 |
| BAT- 655 | Social Media and Digital Marketing | 3 | 0 | 0 | 3 |
| BAT- 656 | Human Resource Management | 3 | 0 | 0 | 3 |
| BAP- 657 | Introduction to SPSS | 0 | 0 | 4 | 2 |
| | **TOTAL CREDITS IN SEMESTER** | | | | **23** |
| | **TOTAL NO. OF SESSIONS** | | | | **25** |
| | **TOTAL NO. OF SESSIONS (8 SESSIONS/WEEK TO TPP + 2 sessions of Mentoring)** | | | | **35** |
| **IN ADDITION TO COMPOULSORY SUBJECTS, A STUDENT HAS TO CHOOSE TWO SUBJECTS FROM EACH OPTED SPECIALISATION** | | | | | |
| **IIIrd Semester** | | | | | |
| BAT- 731 | Retail Management | 3 | 0 | 0 | 3 |
| BAT- 732 | Corporate Strategy | 3 | 0 | 0 | 3 |

| | | | | | |
|---|---|---|---|---|---|
| BAP- 703 | Predictive Analysis | 0 | 0 | 4 | 2 |
| **SPECIALISATION (MARKETING, HR, FINANCE & INTERNATIONAL BUSINESS)** | | | | | |
| | Specialisation Group A/B/C/D/E: Elective I | 4 | 0 | 0 | 4 |
| | Specialisation Group A/B/C/D/E: Elective II | 4 | 0 | 0 | 4 |
| | Specialisation Group A/B/C/D/E: Elective III | 4 | 0 | 0 | 4 |
| | Specialisation Group A/B/C/D/E: Elective IV | 4 | 0 | 0 | 4 |
| BAI- 705 | Summer Training Report | 0 | 0 | 0 | 4 |
| | **TOTAL CREDITS IN SEMESTER** | | | | **28** |
| | **TOTAL NO. OF SESSIONS** | | | | **26** |
| | **TOTAL NO. OF SESSIONS (8 SESSIONS/WEEK TO TPP + 1 sessions of Mentoring)** | | | | **35** |
| **AFTER THIRD SEMESTER, A STUDENT HAS TO CHOOSE TWO SUBJECTS FROM ONE SPECIALISATION** | | | | | |
| **IVth Semester** | | | | | |
| BAT- 780 | Corporate Social Responsibility and Sustainability | 3 | 0 | 0 | 3 |
| BAT- 781 | Indian Ethos and Business Ethics | 3 | 0 | 0 | 3 |
| **SPECIALISATION (ENTREPRENEURSHIP & INTERNATIONAL BUSINESS)** | | | | | |
| | Specialisation Group A/B/C/D/E: Elective I | 4 | 0 | 0 | 4 |
| | Specialisation Group A/B/C/D/E: Elective II | 4 | 0 | 0 | 4 |
| | Specialisation Group A/B/C/D/E: Elective III | 4 | 0 | 0 | 4 |
| | Specialisation Group A/B/C/D/E: Elective IV | 4 | 0 | 0 | 4 |
| BAR- 752 | Final Research Project and Publications | 0 | 0 | 0 | 4 |
| | **TOTAL CREDITS IN SEMESTER** | | | | **26** |
| | **TOTAL NO. OF SESSIONS** | | | | **22** |

**Final Resreach Project & Publications- Students are required to conduct Research Project Survey under the supervision of Assigned Supervisor (Faculty). In this regard, each faculty member will be assigned 4 groups (each group contains 5 students) and each Group has to publish two Research Paper from their research project work in UGC listed Journal. Each published Research Paper will carry 2 credits.**

| **FINANCE (A)** | | | | | |
|---|---|---|---|---|---|
| **SPECIALIZATIONS (Semester-3)** | | | | | |
| BAA-735 | Investment Analysis and Portfolio Management | 4 | 0 | 0 | 4 |
| BAA-736 | Managing Banks and Financial Institutions | 4 | 0 | 0 | 4 |
| **SPECIALIZATIONS (Semester-4)** | | | | | |
| BAA-785 | Financial Markets and Services | 4 | 0 | 0 | 4 |
| BAA-786 | Taxation | 4 | 0 | 0 | 4 |
| **MARKETING (B)** | | | | | |
| **SPECIALIZATIONS (Semester-3)** | | | | | |
| BAB-711 | Consumer Behaviour | 4 | 0 | 0 | 4 |
| BAB-749 | Rural Marketing | 4 | 0 | 0 | 4 |

| | | | | | |
|---|---|---|---|---|---|
| **SPECIALIZATIONS (Semester-4)** | | | | | |
| BAB-787 | Sales and Distribution Management | 4 | 0 | 0 | 4 |
| BAB-788 | Services Marketing | 4 | 0 | 0 | 4 |
| | | | | | |
| **HUMAN RESOURCE MANAGEMENT (C)** | | | | | |
| **SPECIALIZATIONS (Semester-3)** | | | | | |
| BAC-741 | Compensation and Benefits Management | 4 | 0 | 0 | 4 |
| BAC-742 | Strategic HRM | 4 | 0 | 0 | 4 |
| **SPECIALIZATIONS (Semester-4)** | | | | | |
| BAC-789 | Cross Cultural Management | 4 | 0 | 0 | 4 |
| BAC-790 | Employee Relations | 4 | 0 | 0 | 4 |
| | | | | | |
| **INTERNATIONAL BUSINESS (D)** | | | | | |
| **SPECIALIZATIONS (Semester-3)** | | | | | |
| BAD-743 | Export Import Documentation | 4 | 0 | 0 | 4 |
| BAD-745 | International Trade and Laws | 4 | 0 | 0 | 4 |
| **SPECIALIZATIONS (Semester-4)** | | | | | |
| BAD-791 | Globalisation and Indian Multinational Companies | 4 | 0 | 0 | 4 |
| BAD-792 | International Marketing | 4 | 0 | 0 | 4 |
| | | | | | |
| **ENTREPRENEURSHIP (E)** | | | | | |
| **SPECIALIZATIONS (Semester-3)** | | | | | |
| BAE-746 | Entrepreneurial Strategies-I | 4 | 0 | 0 | 4 |
| BAE-747 | Social Entrepreneurship | 4 | 0 | 0 | 4 |
| **SPECIALIZATIONS (Semester-4)** | | | | | |
| BAE-793 | Entrepreneurial Strategies-II | 4 | 0 | 0 | 4 |
| BAE-794 | Entrepreneurial Marketing | 4 | 0 | 0 | 4 |

# CHANDIGARH UNIVERSITY
# Gharuan, Mohali

**Department: University School of Business**
**Division: MBA**
**Subject Name: Predictive Analytics – LAB MANUALS**
**Subject Code: BAP – 703**

<table>
<tr><td rowspan="3"><b>BAP-703</b></td><td colspan="2"><b>Predictive Analysis</b></td><td><b>L</b></td><td><b>T</b></td><td><b>P</b></td><td><b>C</b></td></tr>
<tr><td colspan="2"><b>Total Contact Hours 60</b></td><td><b>0</b></td><td><b>0</b></td><td><b>4</b></td><td><b>2</b></td></tr>
<tr><td colspan="6"><b>MBA</b></td></tr>
<tr><td></td><td colspan="6"><b>Prerequisite: Logical reasoning and aptitude</b></td></tr>
<tr><td></td><td colspan="6"></td></tr>
<tr><td colspan="7"><b>Marks 100</b></td></tr>
<tr><td colspan="2"><b>Internal : 100</b></td><td colspan="5"><b>External : 0</b></td></tr>
<tr><td colspan="7"><b>Course Objective</b></td></tr>
<tr><td colspan="7">To develop the fundamental understanding and application of Mathematics and Statistics in business organizations</td></tr>
<tr><td><b>Unit</b></td><td colspan="6"><b>Course Outcome</b></td></tr>
<tr><td>1</td><td colspan="6">Students will be able to initiate effective use of SPSS in business problem</td></tr>
<tr><td>2</td><td colspan="6">Student will learn to evaluate and solve the business problems logically</td></tr>
<tr><td>3</td><td colspan="6">Students will be able to avoid risks and spot opportunities.</td></tr>
</table>

## Content of the Syllabus

### Unit-1

Introduction to IBM SPSS Statistics: Course Introduction, Introducing IBM SPSS Statistics, Reading Data, Variable Properties, Working with the Data editor, Summarizing Individual Variables, Modifying Data Values: Recode, Modifying Data Values: Compute, Describing Relationship between Variables, Selecting Cases, Creating and Editing Charts, Output in the Viewer, Syntax Basics, Course Summary, Menus and the Help System.

### Unit-II

Data Management and Manipulation with IBM SPSS Statistics, Helpful Data Management Features, Transformations: Computing with Numeric Data, Transformations: Computing with Date and Time Variables, Transformations: Computing with Alphanumeric Data. Additional Data Transformations, Identifying Duplicates and Restructuring Data, Aggregating Data, Merging Files – Adding Cases Adding Variables, Analyzing Multiple Response Questions, Working with Pivot Tables , Working with Charts , Exporting Tables and Charts, An Introduction to Output Management System, Automating IBM SPSS Statistics, Controlling the IBM SPSS Statistics Environment

## Unit-III

Introduction to Statistical Analysis Using IBM SPSS Statistics, Introduction to Statistical Analysis, Understanding Data Distribution Theory, Data Distribution for Categorical Variables, Regression Analysis and Multiple regression analysis

Multidimensional scaling, Factor analysis and Cluster analysis, Concepts of Logistic Regression, Comparison of Several Populations (One way analysis of Variance and Analysis of Variance of Ranks)

**Recommended Books:**

1. Business Research Methods, Cooper, Schindler, TMH
2. Management Research Methodology, Krishnaswamy, Sir Kumar, Pearson
3. Research Methodology, C. R. Kothari, Newage Publication
4. Research Methodology, Zeikmund, Cengage
5. Research Methodology, Paneer Selvam, PHI
6. Research Methodology, Prasanta Sarangi, Taxmann
7. A Text Book of Research Methodology, AKPC Swain, Kalyani
8. SPSS for Windows, Step; George and Mallery,

| **CODE: BAP 703** | **Name: Predictive Analysis** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **University School of Business** | | | | | | | | | | | |
| Program Outcome | | PO1 | PO2 | PO3 | PO4 | PO5 | PO6 | PO7 | PO8 | PO9 | PO10 |
| Mapping of Course outcome with Program outcome | CO1 | **2** | **3** | **2** | **2** | **1** | **2** | **2** | **3** | **2** | **1** |
| | CO2 | **2** | **2** | **1** | **2** | **2** | **3** | **3** | **2** | **2** | **2** |
| | CO3 | **2** | **3** | **2** | **2** | **1** | **2** | **2** | **2** | **1** | **1** |

**NOTE FOR THE PAPER SETTER**

*The syllabus has been divided into three units. Paper setter will set 3 questions from each unit and 1 compulsory question spread over the whole syllabus consisting of 5 short answer questions. Compulsory question will be placed at number one. Candidate shall be required to attempt 6 questions in all including compulsory question and selecting not more than 2 questions from each unit. All questions carry equal marks.*

| Course Name | Course Code | Description of CO | PO1 | PO2 | PO3 | PO45 | PO5 | PO6 |
|---|---|---|---|---|---|---|---|---|
| **Predictive Analysis**<br>**(2018-2020 Batch)** | **BAP 703** | To understand the nature of various data sets and types | 3 | | | | | |
| | | Enabling students with application of advance excel, SPSS and E-views. | | 3 | | | | |
| | | To analyze the different sets of data with the help of different Statistical software's | 2 | 2 | | | | |
| | | To select the appropriate software for analyzing the different set of data | | | | | | |
| | | To create the hypothesis for various business problems | | 3 | | | | |

## List of Experiments – Predictive Analytics

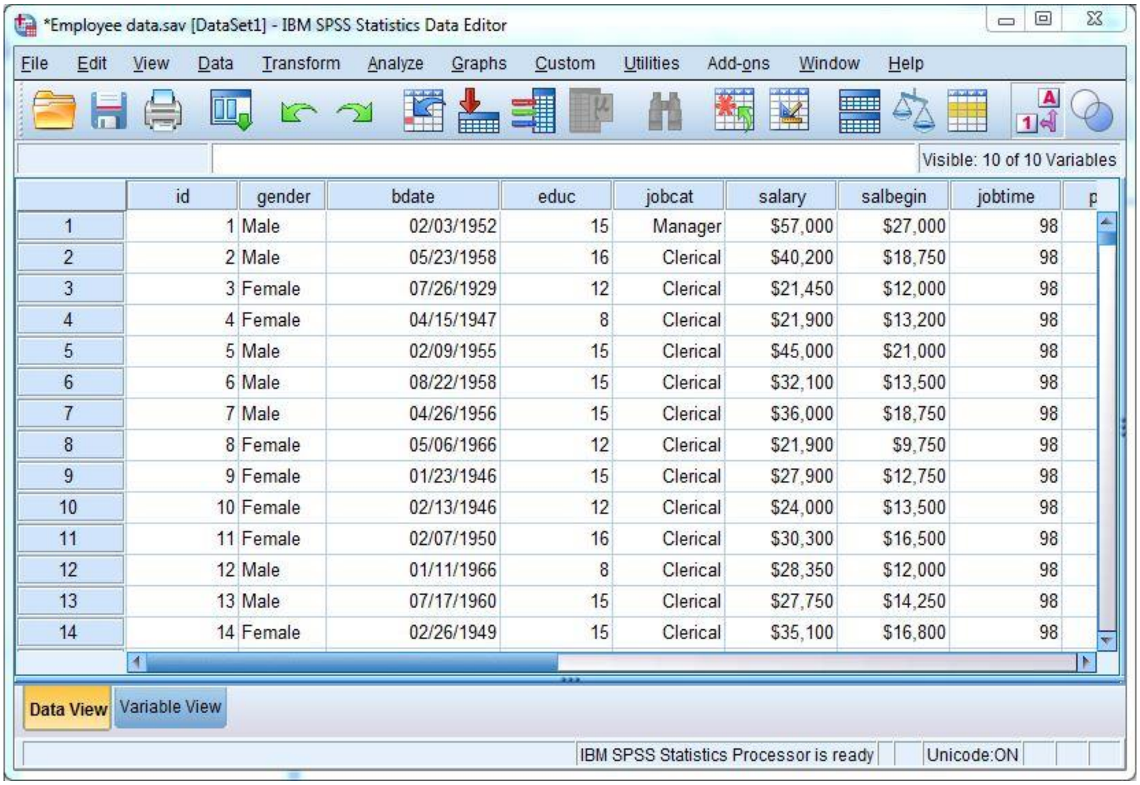
| Experiment Number | Name of Experiment |
|---|---|
| 1 | Introduction to SPSS, Sorting File, Split File, Compute File, Recode File and Select Cases |
| 2 | Chi- Square Test (Parametric and Non-Parametric Test) |
| 3 | Exploratory Factor Analysis |
| 4 | Cluster Analysis |
| 5 | Logistic Regression |
| 6 | Discriminant Analysis |
| 7 | Confirmatory Factor Analysis |
| 8 | Conjoint Analysis |
| 9 | Time Series |
| 10 | MANOVA |
| Additional 11 | Decision Tree Analysis |

# Experiment No. 1

Q. 1. Introduction to SPSS.

SPSS (Statistical Package for Social Sciences) is a versatile and responsive program designed to undertake a range of statistical procedures. SPSS software is widely used in a range of disciplines and is available from all computer pools within the University of South Australia.

It's important to note that SPSS is not only statistical software – there are many others that you may come across if you pursue a career that requires you to work with data. Some of the common statistical packages include Stata and SAS (and there are many others).



| | id | gender | bdate | educ | jobcat | salary | salbegin | jobtime | p |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | Male | 02/03/1952 | 15 | Manager | $57,000 | $27,000 | 98 | |
| 2 | 2 | Male | 05/23/1958 | 16 | Clerical | $40,200 | $18,750 | 98 | |
| 3 | 3 | Female | 07/26/1929 | 12 | Clerical | $21,450 | $12,000 | 98 | |
| 4 | 4 | Female | 04/15/1947 | 8 | Clerical | $21,900 | $13,200 | 98 | |
| 5 | 5 | Male | 02/09/1955 | 15 | Clerical | $45,000 | $21,000 | 98 | |
| 6 | 6 | Male | 08/22/1958 | 15 | Clerical | $32,100 | $13,500 | 98 | |
| 7 | 7 | Male | 04/26/1956 | 15 | Clerical | $36,000 | $18,750 | 98 | |
| 8 | 8 | Female | 05/06/1966 | 12 | Clerical | $21,900 | $9,750 | 98 | |
| 9 | 9 | Female | 01/23/1946 | 15 | Clerical | $27,900 | $12,750 | 98 | |
| 10 | 10 | Female | 02/13/1946 | 12 | Clerical | $24,000 | $13,500 | 98 | |
| 11 | 11 | Female | 02/07/1950 | 16 | Clerical | $30,300 | $16,500 | 98 | |
| 12 | 12 | Male | 01/11/1966 | 8 | Clerical | $28,350 | $12,000 | 98 | |
| 13 | 13 | Male | 07/17/1960 | 15 | Clerical | $27,750 | $14,250 | 98 | |
| 14 | 14 | Female | 02/26/1949 | 15 | Clerical | $35,100 | $16,800 | 98 | |

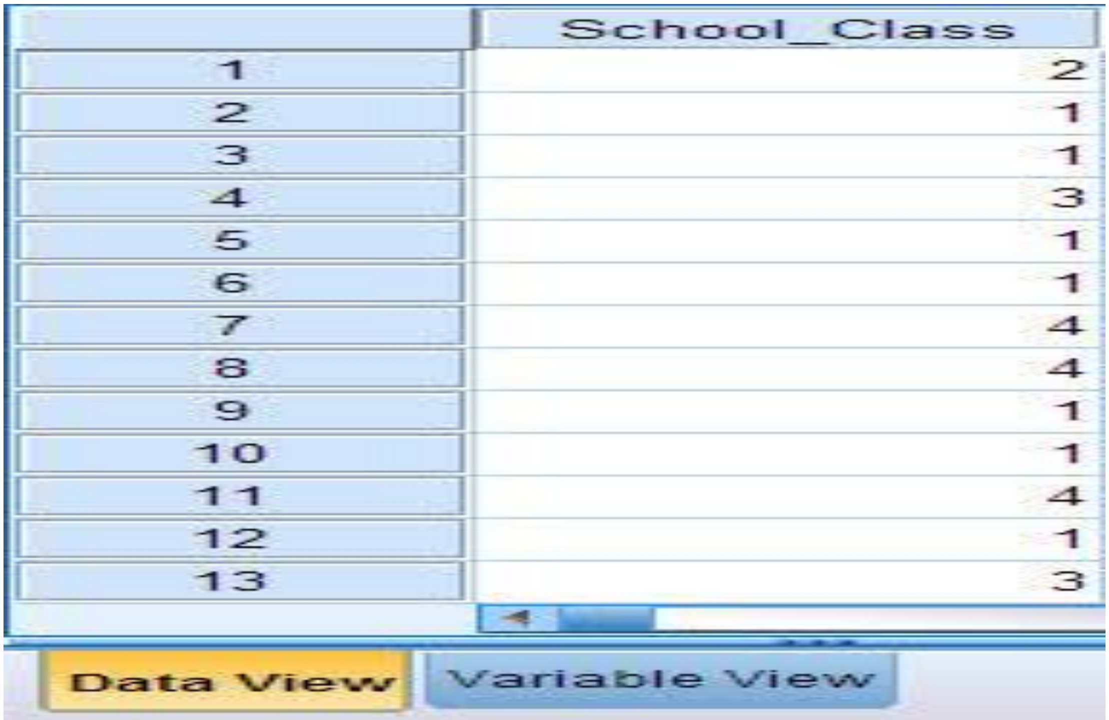
IBM SPSS Statistics Processor is ready

Q.2. Four Windows of SPSS.

1. Input window : It is the first page that gets displayed when we open SPSS software. Here we enter the data primarily.

- Data view : The data view is used to store and show your data. It is much like an ordinary spreadsheet although in general the data is structured so that rows are cases and the columns are for the different variables that relate to each case.

- Variable view : The variable view contains the variables on your data set , so it defines the properties of your dataset. Each row will define all of the various variables . The variables includes – name, type, width, decimals, label, values, missing, column, align and measure.

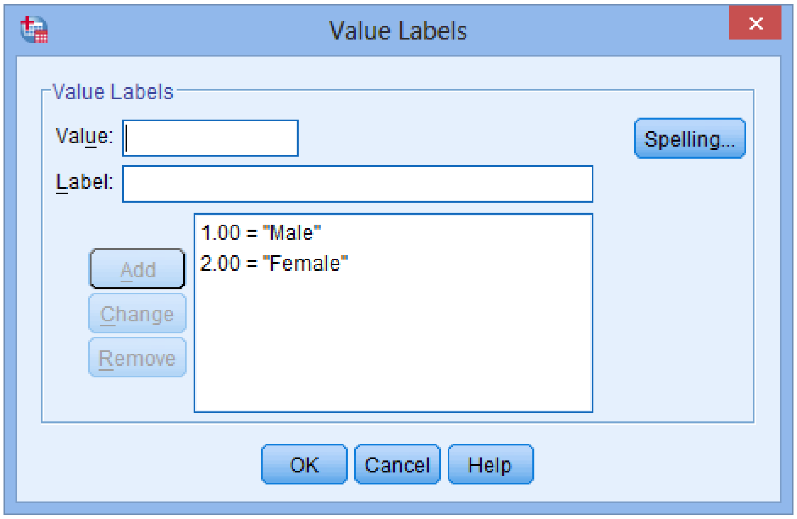


| | School_Class |
|---|---|
| 1 | 2 |
| 2 | 1 |
| 3 | 1 |
| 4 | 3 |
| 5 | 1 |
| 6 | 1 |
| 7 | 4 |
| 8 | 4 |
| 9 | 1 |
| 10 | 1 |
| 11 | 4 |
| 12 | 1 |
| 13 | 3 |

2. Output window : This window is used to show the results that have been output from your data analysis. Depending on the analysis that you are carrying out this may include the Chart Editor Window or Pivot Table Window.

3. Syntax window : This window shows the underlying commands that have executed your data analysis. If you are a confident coder this is where you can amend the code or write your own from scratch and then run your own custom analysis on your data set.

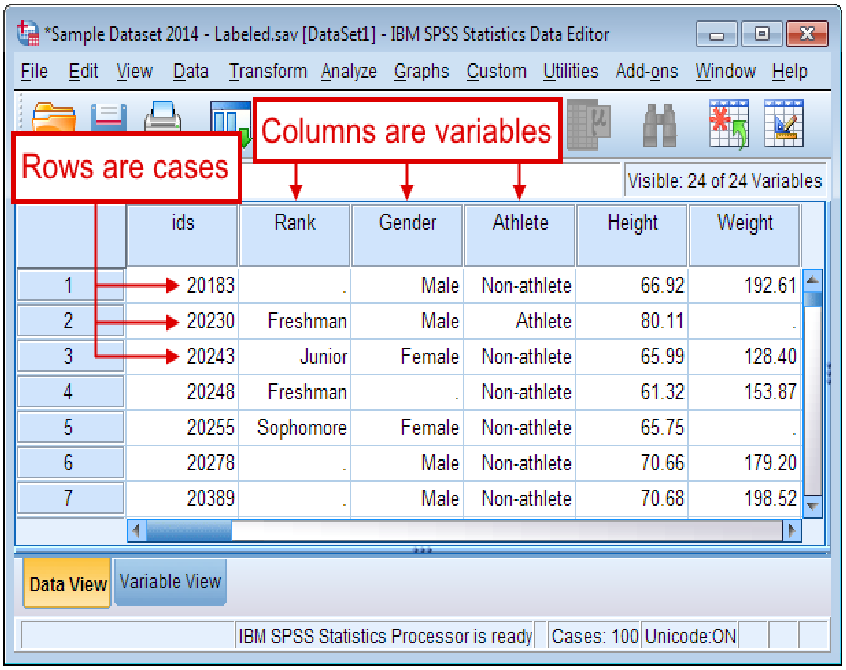
4. Script window :Scripts can be used to customize operations within a particular stream and they are saved with that stream. Scripts can be used to specify a particular execution order for the terminal nodes within a stream. It is mainly used for coding and programming.



1.00 = "Male"
2.00 = "Female"

Q.3. Uses of SPSS in business?

- SPSS is used as a data collection tool by researchers. The data entry screen in SPSS looks much like any other spreadsheet software. We can enter variables and quantitative data and save the file as a data file. Furthermore, we can organize our data in SPSS by assigning properties to different variables.
- Once data is collected and entered the data sheet in SPSS, we can create an output file from the data. For example, we can create frequency distribution of our data to determine whether our data is normally distributed. The frequency distribution is displayed in an output file.
- The most obvious use for SPSS is to use the software to run statistical tests. SPSS has all of the most widely used statistical tests built-in to the software. Therefore, we won't have to do any mathematical equations by hand.
- SPSS helps to create reports of questionnaire data in the form of graphical presentations which are ready for publications and reporting.
- SPSS research tool can compare and explore the differences between responses to two or more questions. It's very easy to find the difference between to batches of data.
- Using this tool we can make analysis according to the expected research goals and obtain the gained results.
- SPSS is the powerful tool for data analyzing and it's also used for SPSS data entry.  It's the first step in statistical process and it should be very important to input data correctly.



Columns are variables

Rows are cases

Visible: 24 of 24 Variables

| | ids | Rank | Gender | Athlete | Height | Weight |
|---|---|---|---|---|---|---|
| 1 | 20183 | . | Male | Non-athlete | 66.92 | 192.61 |
| 2 | 20230 | Freshman | Male | Athlete | 80.11 | . |
| 3 | 20243 | Junior | Female | Non-athlete | 65.99 | 128.40 |
| 4 | 20248 | Freshman | . | Non-athlete | 61.32 | 153.87 |
| 5 | 20255 | Sophomore | Female | Non-athlete | 65.75 | . |
| 6 | 20278 | . | Male | Non-athlete | 70.66 | 179.20 |
| 7 | 20389 | . | Male | Non-athlete | 70.68 | 198.52 |

IBM SPSS Statistics Processor is ready | Cases: 100 | Unicode:ON

The students are expected to perform the following functions in SPSS:

1. Sorting file
2. Split file
3. Compute
4. Recode
5. Select cases

## SORTING FILE

Sorting data allows us to re-organize the data in ascending or descending order with respect to a specific variable. Same procedures in SPSS require that your data be sorted in a certain way before the procedure will execute. **Using SPSS and Pasw/sorting variables.**

One of the function you will often want to perform in the data view of the data editor is sorting by a variables values to bring those of similar value together. This can be a very useful when exploring the raw data in your datasheet. There are two ways to do this. One is quite simple, the other allows sorting on more than one variable.

### STEPS

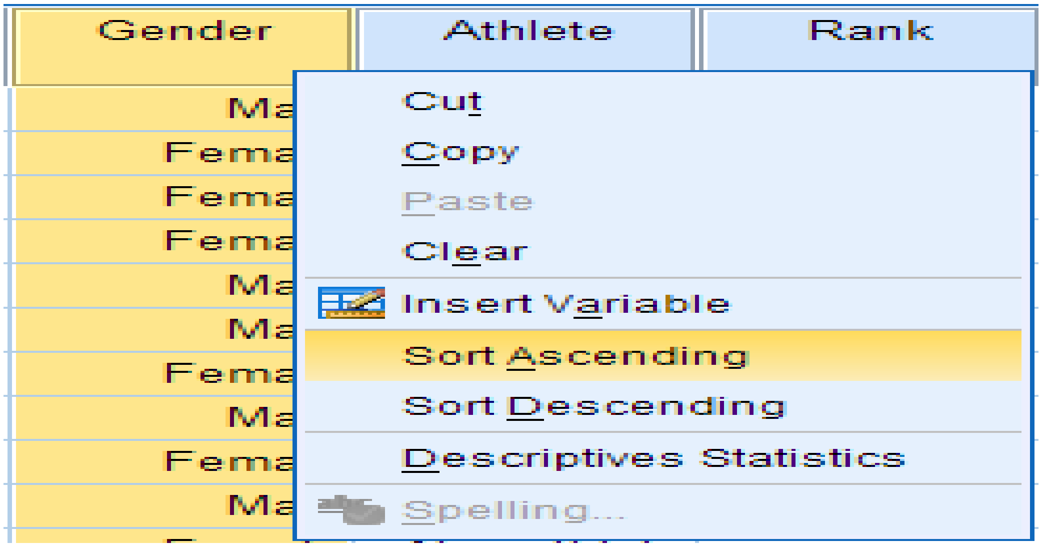
The simple way to sort variable values is to make sure you are looking at the data view tab. Then scroll to the variable by which you want to sort. Right click on its column heading and a context menu will appear. At the bottom of the context menu are two sorting options: " sort ascending" and "sort descending". Choosing the first will move the smallest values to the top of the data set while a descending sort will bring up the largest value. Keep in mind that SPSS automatically (unlike excel) moves all rows in unison. So you don't need to worry about cases becoming misaligned when sorting.

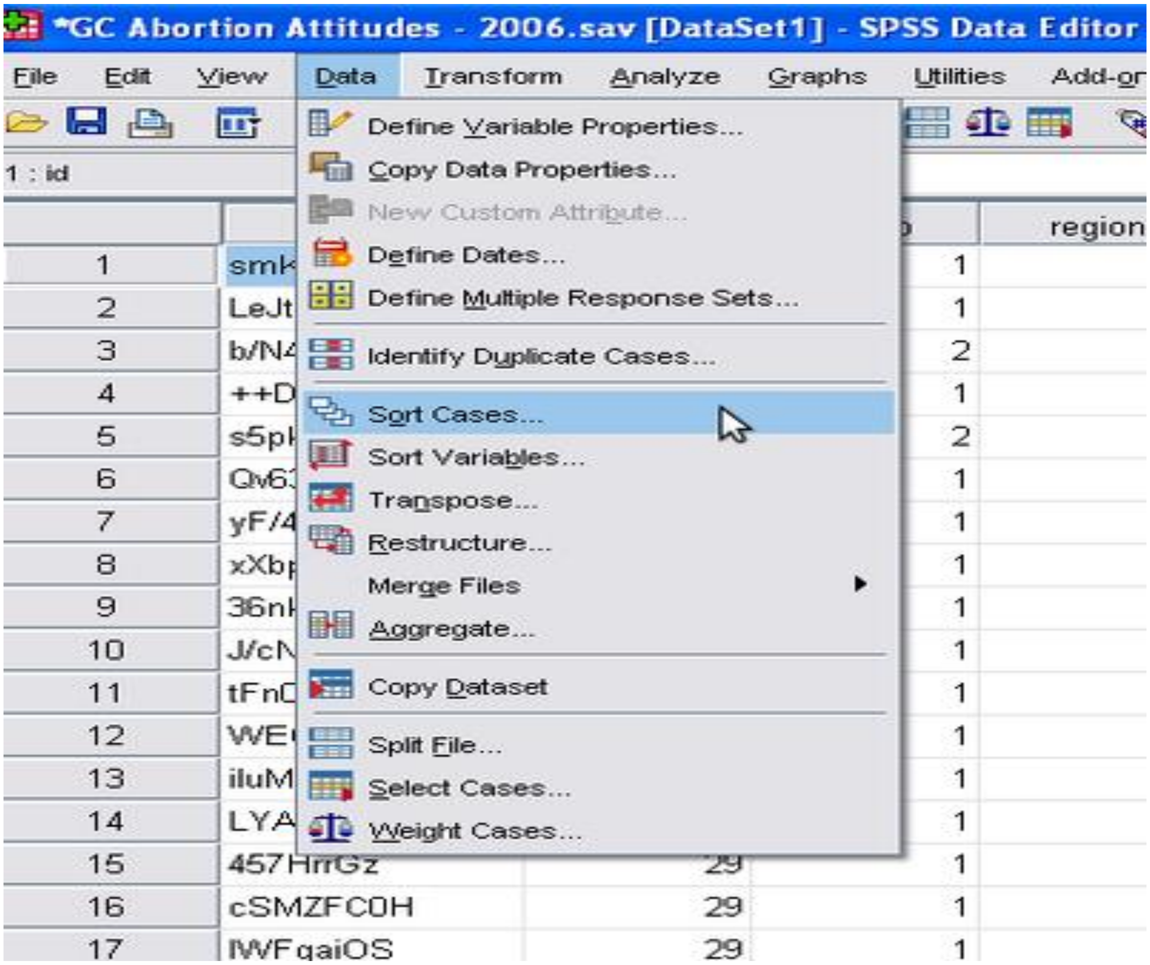
Sometimes its useful to sort on multiple variables, which means that SPSS while sort the data set by the values of the first variable than breaks ties in that sort by sorting on the values of the second variable and so on. The sort on multiple variable at once choose " data" >"sort cases" :

You will get dialog window:

The "sort order" box allows the choice between ascending or descending order once again. On the left is a list of all the variables in the data set. Choose them the most important sort first by clicking the variable involved and then the blue arrow and repeat.

Save your data set after sorting to retain it as a default sorting.


Gender
Athlete
Rank
Ma
Fema
Fema
Fema
Ma
Ma
Fema
Ma
Fema
Ma
Cut
Copy
Paste
Clear
Insert Variable
Sort Ascending
Sort Descending
Descriptives Statistics
Spelling...


*GC Abortion Attitudes - 2006.sav [DataSet1] - SPSS Data Editor
File
Edit
View
Data
Transform
Analyze
Graphs
Utilities
Add-on
1 : id
region
Define Variable Properties...
Copy Data Properties...
New Custom Attribute...
Define Dates...
Define Multiple Response Sets...
Identify Duplicate Cases...
Sort Cases...
Sort Variables...
Transpose...
Restructure...
Merge Files
Aggregate...
Copy Dataset
Split File...
Select Cases...
Weight Cases...
1 smk 1
2 LeJt 1
3 b/N4 2
4 ++D 1
5 s5pl 2
6 Qv6 1
7 yF/4 1
8 xXb 1
9 36nl 1
10 J/cN 1
11 tFnC 1
12 WE 1
13 iluM 1
14 LYA 1
15 457HnGz 29 1
16 cSMZFC0H 29 1
17 lWFgaiOS 29 1

**SPLIT FILE**

Split file is a SPSS facility to perform any analysis sequence for each group defined by a categorical split variables. When you turn split file on. It will be active until you turn it off explicity  or you replace the current data set with a different one.

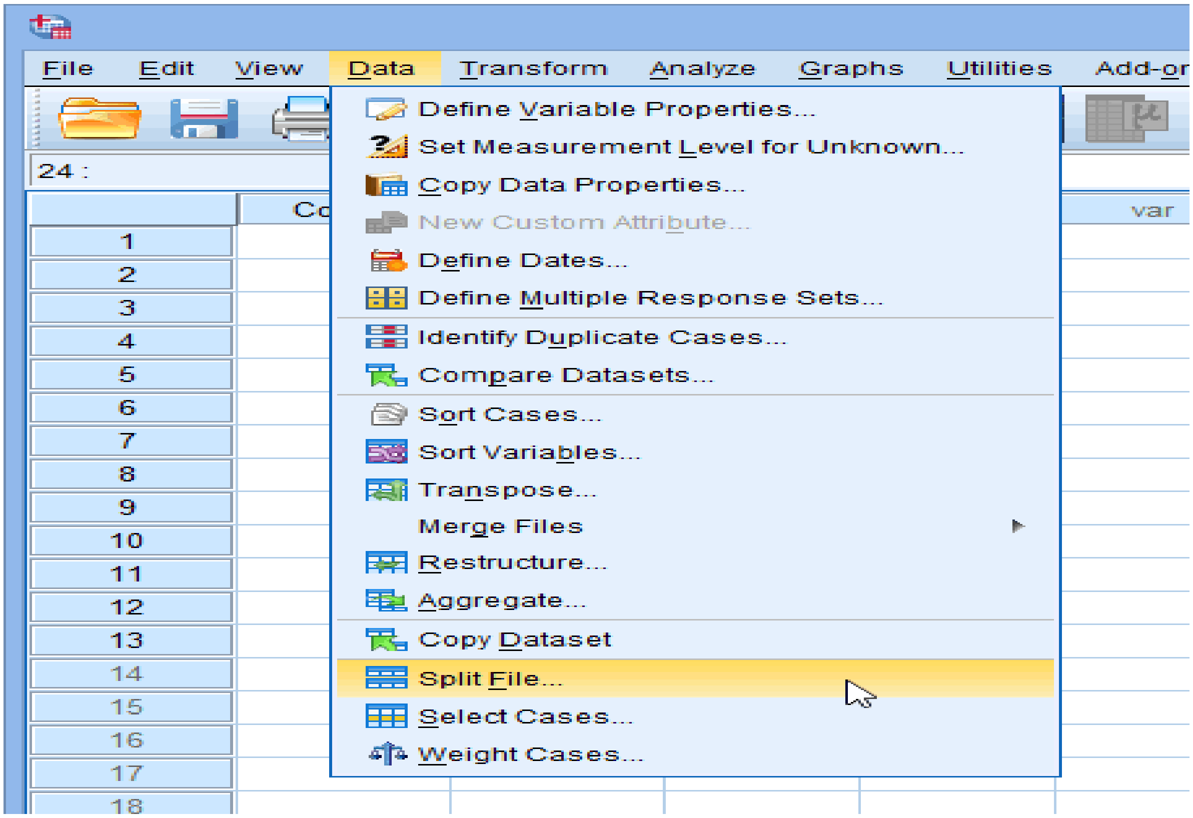
**STEPS**

The data>split file dialog lets you control this mechanism. To active

- Choose either
- Compare groups: produces single tables including all groups
- Organize output by groups: produces separate tables for each group
- Select a variable region in our example for the group based on field
- Make sure to check sort the file by grouping variables is selected, if you are not certain that the file is sorted on the grouping variable.

After clicking OK, split is activated ( you can see “split file on” in the status line of the SPSS window) and any procedure you invoke now will be performed. Separately on each group defined by the grouping variables as long as you do not turn off split file.

If you need to know what is current split variable you will have to open the data> split file dialog.



**COMPUTE**

Sometimes you may need to compute a new variable based on existing information (from other variables) in your data.

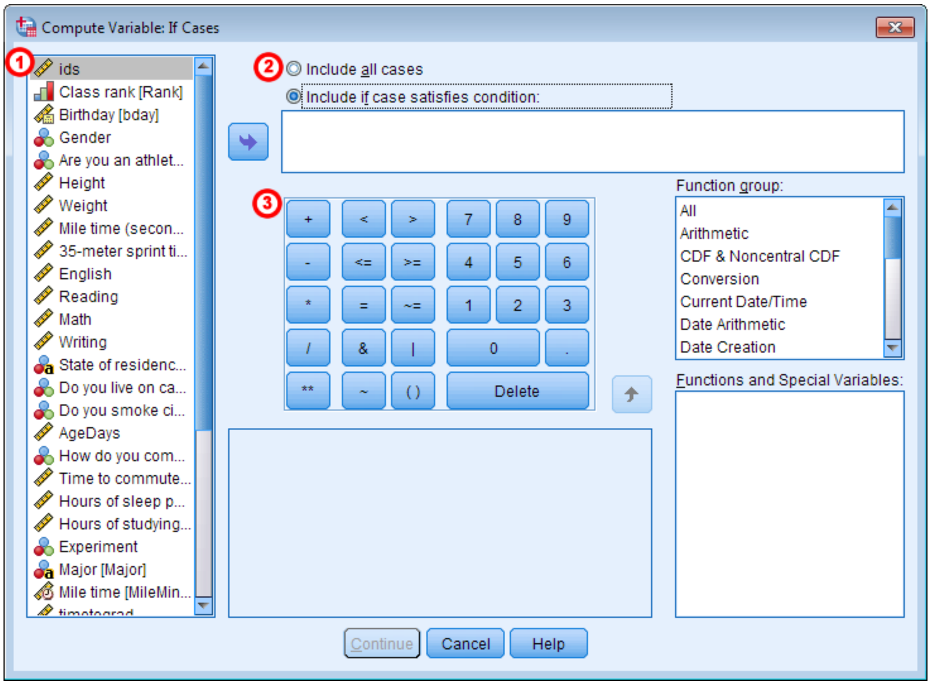
For example: you may want to:

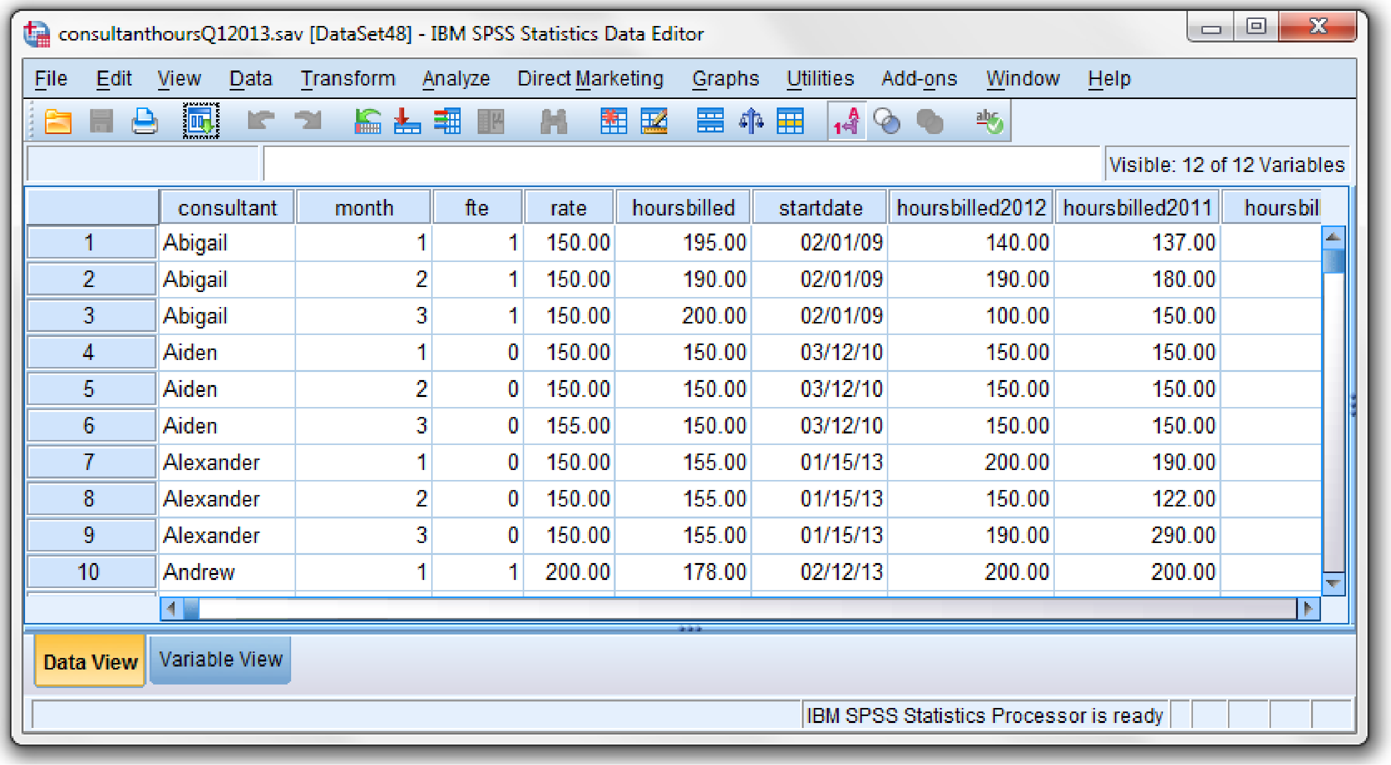
- Convert the units of a variable from feet to meters.
- Use a subjects height and weight to compute their BMI.
- Compute a sub scale score from items on a survey
- Apply a competition conditionally so that a new variable is only computed for cases where certain conditions are meet

In this tutorial we will discus how to compute variables in SPSS using numeric expressions, built –in functions and conditional logic.

To compute a new variable. Click transform > compute variable

The compute variable window will open where you specify how to calculate your new variable




consultanthoursQ12013.sav [DataSet48] - IBM SPSS Statistics Data Editor
File
Edit
View
Data
Transform
Analyze
Direct Marketing
Graphs
Utilities
Add-ons
Window
Help
Visible: 12 of 12 Variables
Data View
Variable View
IBM SPSS Statistics Processor is ready

| | consultant | month | fte | rate | hoursbilled | startdate | hoursbilled2012 | hoursbilled2011 | hoursbil |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Abigail | 1 | 1 | 150.00 | 195.00 | 02/01/09 | 140.00 | 137.00 | |
| 2 | Abigail | 2 | 1 | 150.00 | 190.00 | 02/01/09 | 190.00 | 180.00 | |
| 3 | Abigail | 3 | 1 | 150.00 | 200.00 | 02/01/09 | 100.00 | 150.00 | |
| 4 | Aiden | 1 | 0 | 150.00 | 150.00 | 03/12/10 | 150.00 | 150.00 | |
| 5 | Aiden | 2 | 0 | 150.00 | 150.00 | 03/12/10 | 150.00 | 150.00 | |
| 6 | Aiden | 3 | 0 | 155.00 | 150.00 | 03/12/10 | 150.00 | 150.00 | |
| 7 | Alexander | 1 | 0 | 150.00 | 155.00 | 01/15/13 | 200.00 | 190.00 | |
| 8 | Alexander | 2 | 0 | 150.00 | 155.00 | 01/15/13 | 150.00 | 122.00 | |
| 9 | Alexander | 3 | 0 | 150.00 | 155.00 | 01/15/13 | 190.00 | 290.00 | |
| 10 | Andrew | 1 | 1 | 200.00 | 178.00 | 02/12/13 | 200.00 | 200.00 | |

## RECODE

Sometimes you will want to transform a variable by grouping its categories or values together. For example: you may want to change a continuous variable into a categorical variable or you may want to merge the categories of a normal variable. In SPSS this type of transform is called recoding.

In SPSS there are three basic options for recoding variables.

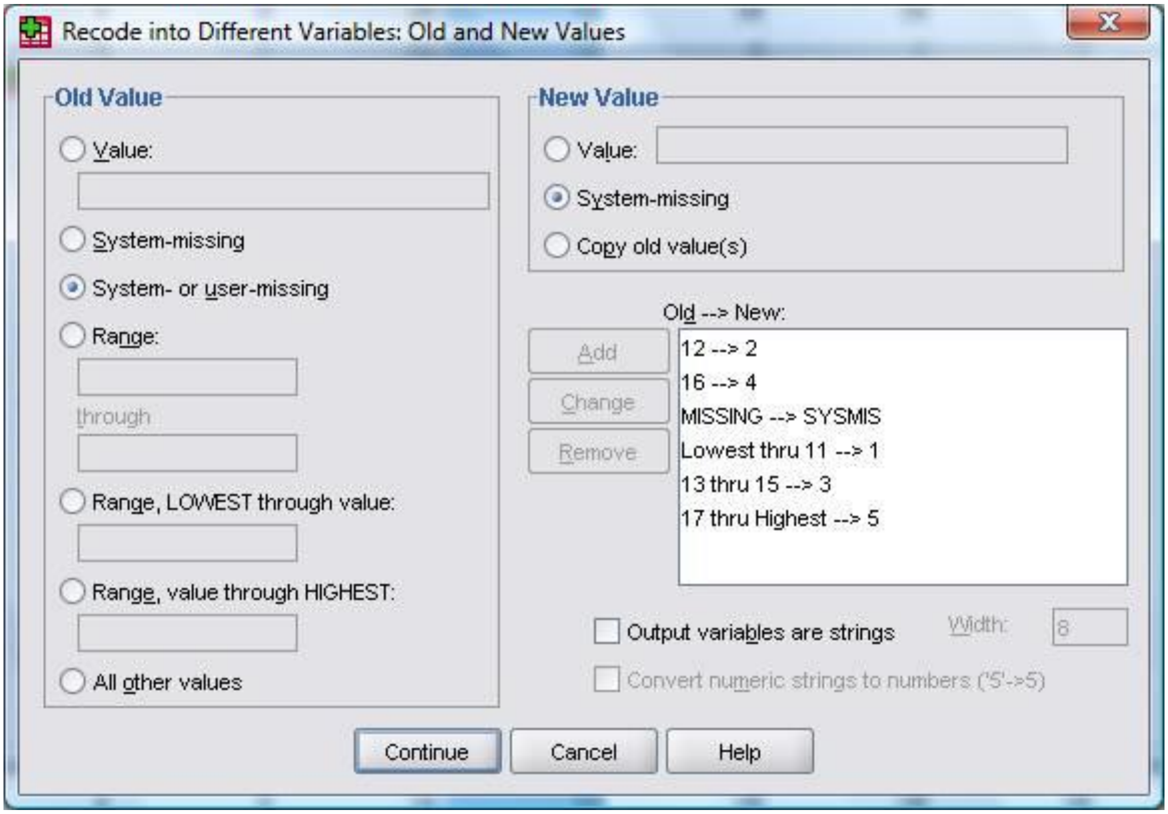
- Recode in different variables
- Recode into same variables
- Do if syntax

Recoding into a different variable transforms an original variable into a new variable. That is the changes do not overwrite the original variable, they are instead applied to a copy of the original variable under a new name.

## STEPS

To recode into different variables click transform >recode into different variables

The recode into different variables window will appear. The left column lists all of the variables in your data set. Select the variables you wish to recode by clicking it. Click the arrow in the center to move the selected variable to the center text box.



Recode into Different Variables: Old and New Values

Old Value

○ Value:

○ System-missing

● System- or user-missing

○ Range:

through

○ Range, LOWEST through value:

○ Range, value through HIGHEST:

○ All other values

New Value

○ Value:

● System-missing

○ Copy old value(s)

Old --> New:

12 --> 2
16 --> 4
MISSING --> SYSMIS
Lowest thru 11 --> 1
13 thru 15 --> 3
17 thru Highest --> 5

□ Output variables are strings Width: 8

□ Convert numeric strings to numbers ('5'->5)

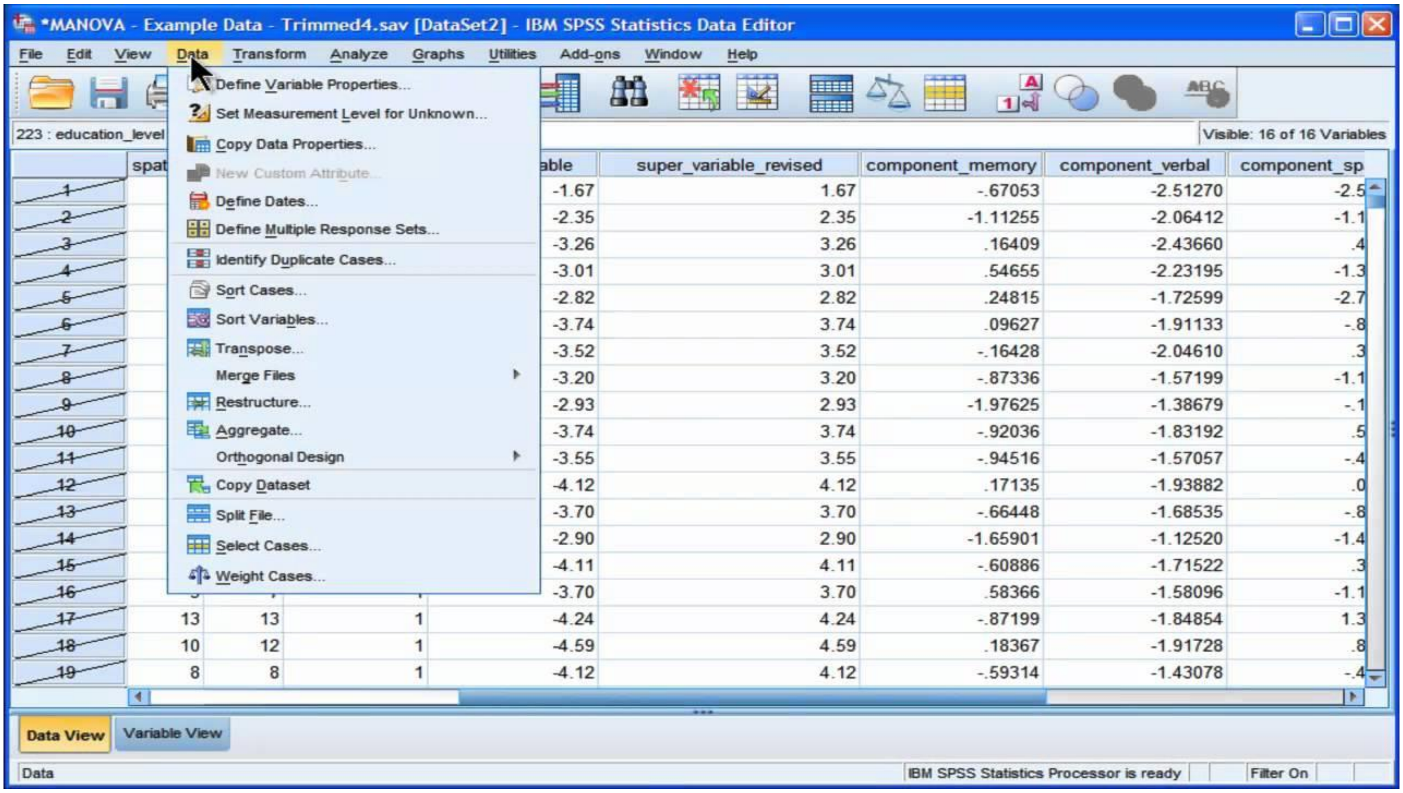
## SELECT CASES

Data files are not always organized to meet specific users need. For example: users may wish to select specific subjects or split the data file into separate groups for analysis. If you have two or more subjects groups in your data and you want to analyze each subject independently you want to analyze each subject you can use the select case option.

## STEPS

- Open the part1.sav data file provided in your computer. Make sure you are on data view
- Click the data menu, and then click select cases. The select cases dialog box opens. Select the if condition is satisfied option

- Click the if button. The select cases: if dialog box opens. Select the gender variables in the left box, and then click the transfer arrow button to move it to the right box. Click the = button and then click the 1 button. Because the symbol 1 represents formals according to our value levels we are telling SPSS to select only female participants.
- Click the continue button to return to the select cases dialog box. Click the OK button to return to the data view. All males will be excluded from the statistical analysis.
- Run an analysis. Note the crossed out participants in the data file. Those represents all the male participants.
- To undo the select cases open the select cases box and click the reset button.



| | spat | | | able | super_variable_revised | component_memory | component_verbal | component_sp |
|---|---|---|---|---|---|---|---|---|
| ~~1~~ | | | | -1.67 | 1.67 | -.67053 | -2.51270 | -2.5 |
| ~~2~~ | | | | -2.35 | 2.35 | -1.11255 | -2.06412 | -1.1 |
| ~~3~~ | | | | -3.26 | 3.26 | .16409 | -2.43660 | .4 |
| ~~4~~ | | | | -3.01 | 3.01 | .54655 | -2.23195 | -1.3 |
| ~~5~~ | | | | -2.82 | 2.82 | .24815 | -1.72599 | -2.7 |
| ~~6~~ | | | | -3.74 | 3.74 | .09627 | -1.91133 | -.8 |
| ~~7~~ | | | | -3.52 | 3.52 | -.16428 | -2.04610 | .3 |
| ~~8~~ | | | | -3.20 | 3.20 | -.87336 | -1.57199 | -1.1 |
| ~~9~~ | | | | -2.93 | 2.93 | -1.97625 | -1.38679 | -.1 |
| ~~10~~ | | | | -3.74 | 3.74 | -.92036 | -1.83192 | .5 |
| ~~11~~ | | | | -3.55 | 3.55 | -.94516 | -1.57057 | -.4 |
| ~~12~~ | | | | -4.12 | 4.12 | .17135 | -1.93882 | .0 |
| ~~13~~ | | | | -3.70 | 3.70 | -.66448 | -1.68535 | -.8 |
| ~~14~~ | | | | -2.90 | 2.90 | -1.65901 | -1.12520 | -1.4 |
| ~~15~~ | | | | -4.11 | 4.11 | -.60886 | -1.71522 | .3 |
| ~~16~~ | | | | -3.70 | 3.70 | .58366 | -1.58096 | -1.1 |
| ~~17~~ | 13 | 13 | 1 | -4.24 | 4.24 | -.87199 | -1.84854 | 1.3 |
| ~~18~~ | 10 | 12 | 1 | -4.59 | 4.59 | .18367 | -1.91728 | .8 |
| ~~19~~ | 8 | 8 | 1 | -4.12 | 4.12 | -.59314 | -1.43078 | -.4 |

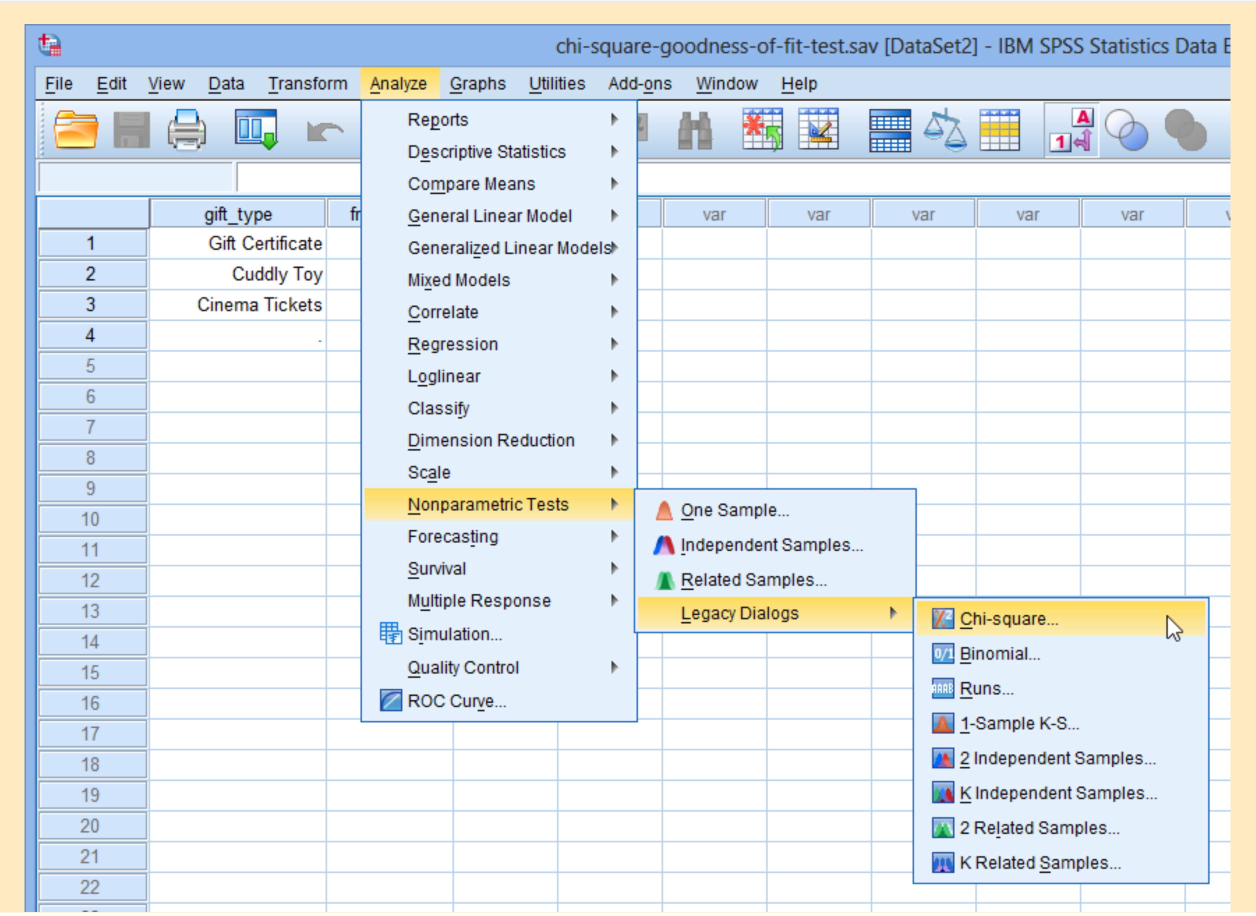
# Experiment 2: A Chi-square goodness-of-fit test in SPSS

The four steps below show you how to analyse your data using a chi-square goodness-of-fit test in SPSS Statistics when you have hypothesised that you have equal expected proportions (N.B., if you are unclear about the differences between equal and unequal expected proportions, see the Introduction). Also, it is important to note that this procedure will only give you the correct results if you have set up your data correctly in SPSS Statistics (N.B., if you have entered the summated frequencies for each group of your categorical variable, this procedure will only work if you have already "weighted" your cases, as we explained in the Data Setup section earlier, but if you have entered all of your data into SPSS Statistics in raw form, this procedure will not give the correct results). In our enhanced chi-square goodness-of-fit test guide, we show all the SPSS Statistics procedures for when you have equal and unequal expected proportions, as well as when you have to weight your cases or have not summated your data. If you only need to follow this "quick start" guide for equal expected proportions (without the weighting of cases), the four steps you need are shown below. At the end of these four steps, we show you how to interpret the results from this test.

---

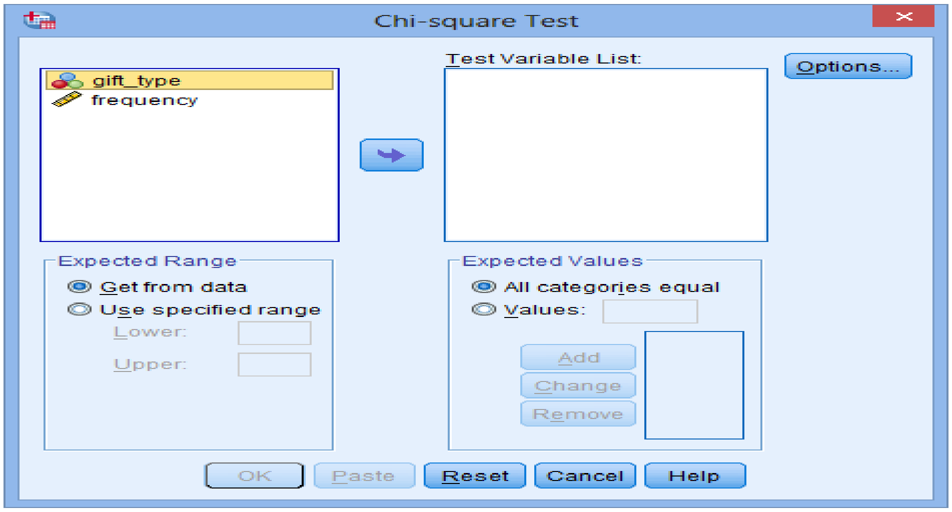
- Click **Analyze** > **Nonparametric Tests** > **Legacy Dialogs** > **Chi-square...** on the top menu as shown below:

**Note:** If you are on older versions of SPSS Statistics, you will not have to go through the **Legacy Dialogs** me



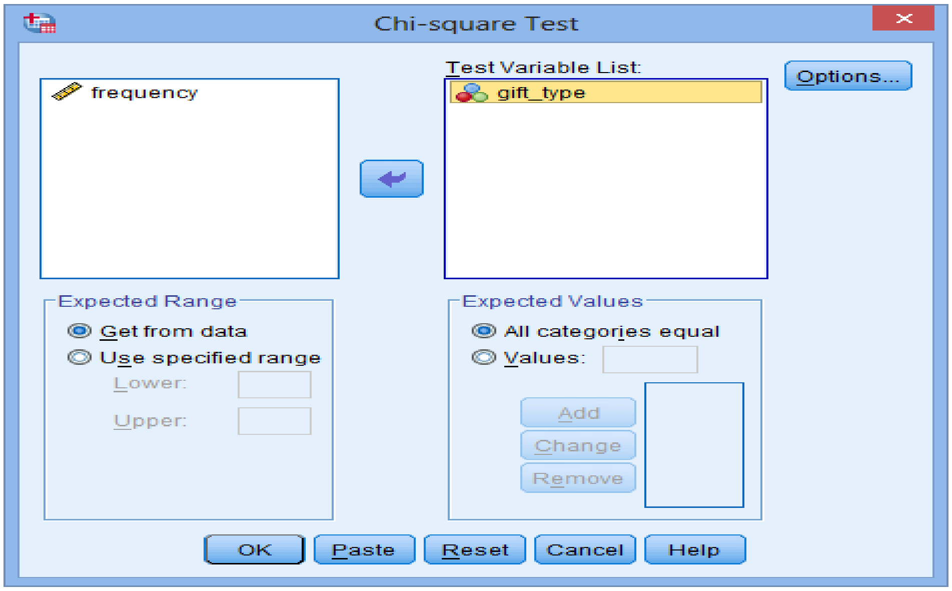
Published with written permission from SPSS Statistics, IBM Corporation.

- You will be presented with the **Chi-square Test** dialogue box, as shown below:



Published with written permission from SPSS Statistics, IBM Corporation.

- Transfer the gift_type variable into the Test Variable List: box by using the ➡ button, as shown below:



Chi-square Test

frequency

Test Variable List:
gift_type

Expected Range
● Get from data
○ Use specified range
Lower:
Upper:

Expected Values
● All categories equal
○ Values:

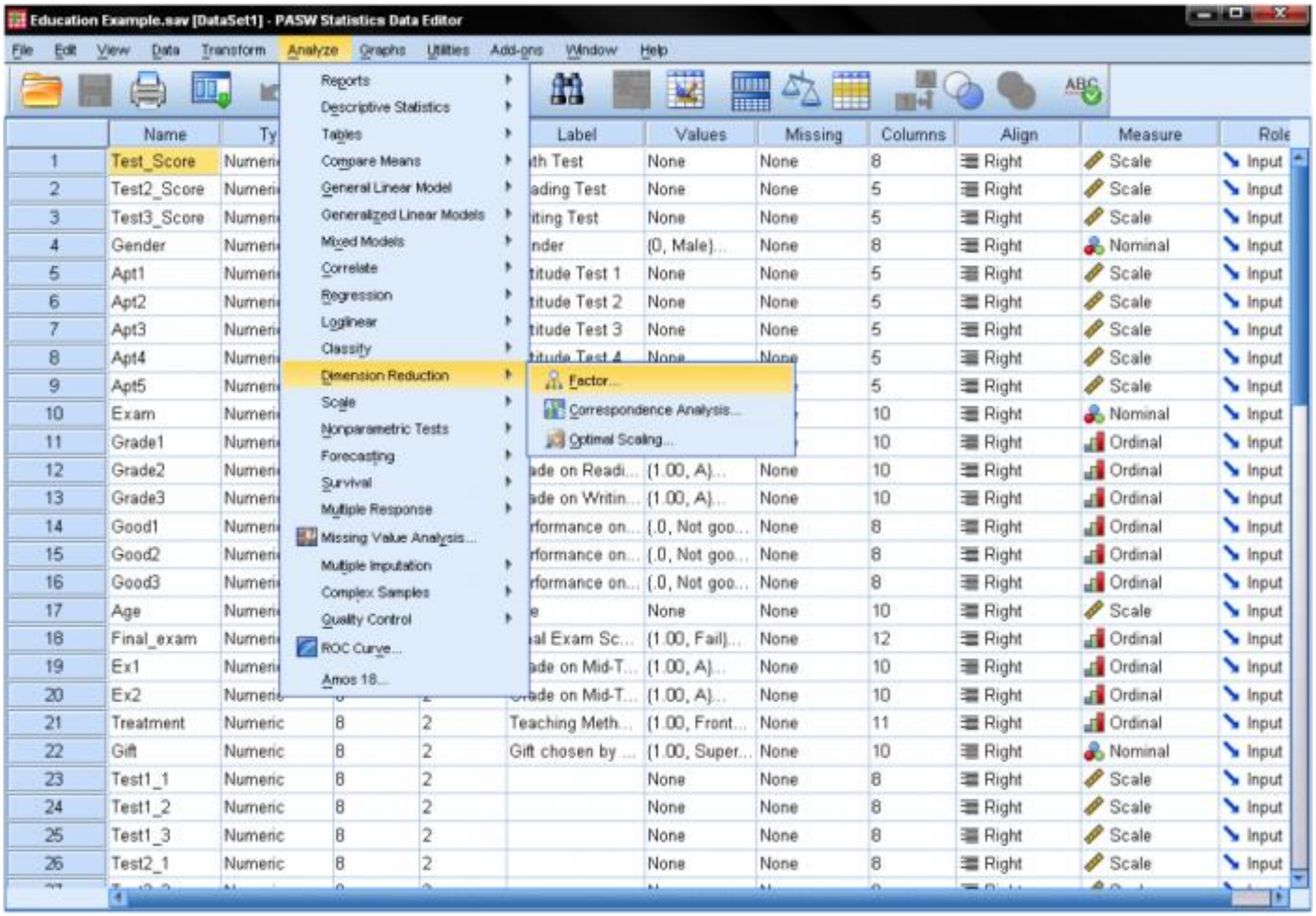
Published with written permission from SPSS Statistics, IBM Corporation.

Keep the All categories equal option selected in the –Expected Values– area as we are assuming equal proportions for each category.

- Click the OK button to generate the output.

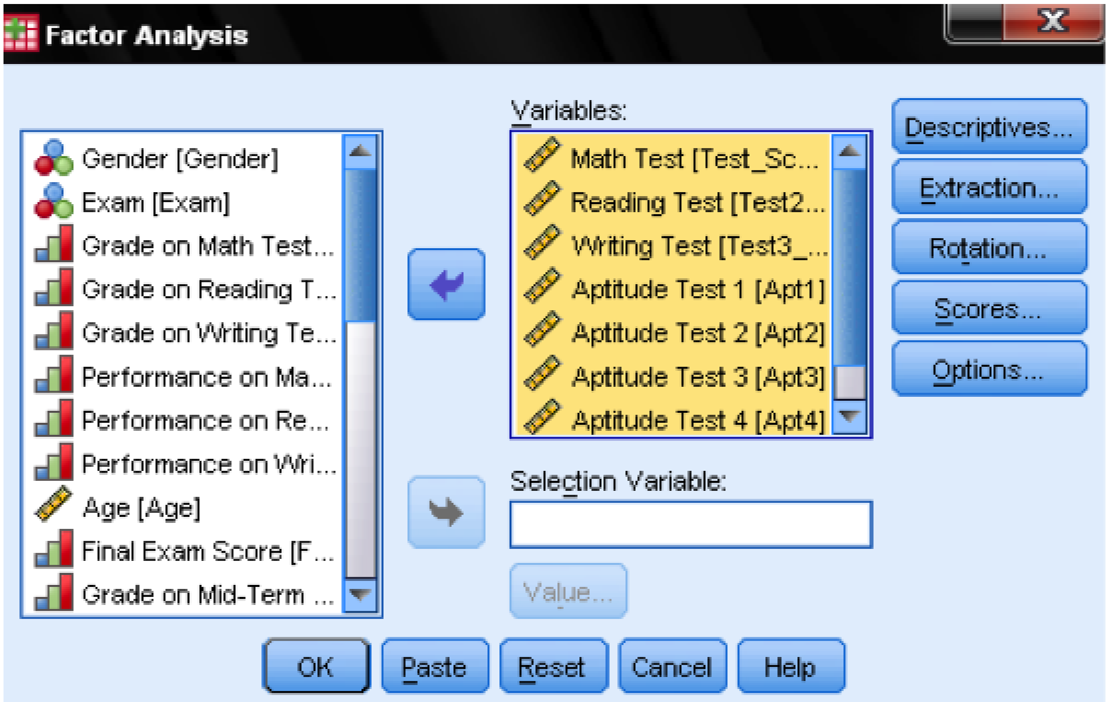
# Experiment 3: Exploratory Factor Analysis

The Exploratory factor analysis can be found in *Analyze/Dimension Reduction/Factor…*



| | Name | Ty | | | Label | Values | Missing | Columns | Align | Measure | Role |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Test_Score | Numeri | | | th Test | None | None | 8 | Right | Scale | Input |
| 2 | Test2_Score | Numeri | | | ading Test | None | None | 5 | Right | Scale | Input |
| 3 | Test3_Score | Numeri | | | iting Test | None | None | 5 | Right | Scale | Input |
| 4 | Gender | Numeri | | | nder | {0, Male}... | None | 8 | Right | Nominal | Input |
| 5 | Apt1 | Numeri | | | titude Test 1 | None | None | 5 | Right | Scale | Input |
| 6 | Apt2 | Numeri | | | titude Test 2 | None | None | 5 | Right | Scale | Input |
| 7 | Apt3 | Numeri | | | titude Test 3 | None | None | 5 | Right | Scale | Input |
| 8 | Apt4 | Numeri | | | titude Test 4 | None | None | 5 | Right | Scale | Input |
| 9 | Apt5 | Numeri | | | | | | 5 | Right | Scale | Input |
| 10 | Exam | Numeri | | | | | | 10 | Right | Nominal | Input |
| 11 | Grade1 | Numeri | | | | | | 10 | Right | Ordinal | Input |
| 12 | Grade2 | Numeri | | | ade on Readi... | {1.00, A}... | None | 10 | Right | Ordinal | Input |
| 13 | Grade3 | Numeri | | | ade on Writin... | {1.00, A}... | None | 10 | Right | Ordinal | Input |
| 14 | Good1 | Numeri | | | rformance on... | {.0, Not goo... | None | 8 | Right | Ordinal | Input |
| 15 | Good2 | Numeri | | | rformance on... | {.0, Not goo... | None | 8 | Right | Ordinal | Input |
| 16 | Good3 | Numeri | | | rformance on... | {.0, Not goo... | None | 8 | Right | Ordinal | Input |
| 17 | Age | Numeri | | | e | None | None | 10 | Right | Scale | Input |
| 18 | Final_exam | Numeri | | | al Exam Sc... | {1.00, Fail}... | None | 12 | Right | Ordinal | Input |
| 19 | Ex1 | Numeri | | | ade on Mid-T... | {1.00, A}... | None | 10 | Right | Ordinal | Input |
| 20 | Ex2 | Numeric | | | ade on Mid-T... | {1.00, A}... | None | 10 | Right | Ordinal | Input |
| 21 | Treatment | Numeric | 8 | 2 | Teaching Meth... | {1.00, Front... | None | 11 | Right | Ordinal | Input |
| 22 | Gift | Numeric | 8 | 2 | Gift chosen by ... | {1.00, Super... | None | 10 | Right | Nominal | Input |
| 23 | Test1_1 | Numeric | 8 | 2 | | None | None | 8 | Right | Scale | Input |
| 24 | Test1_2 | Numeric | 8 | 2 | | None | None | 8 | Right | Scale | Input |
| 25 | Test1_3 | Numeric | 8 | 2 | | None | None | 8 | Right | Scale | Input |
| 26 | Test2_1 | Numeric | 8 | 2 | | None | None | 8 | Right | Scale | Input |

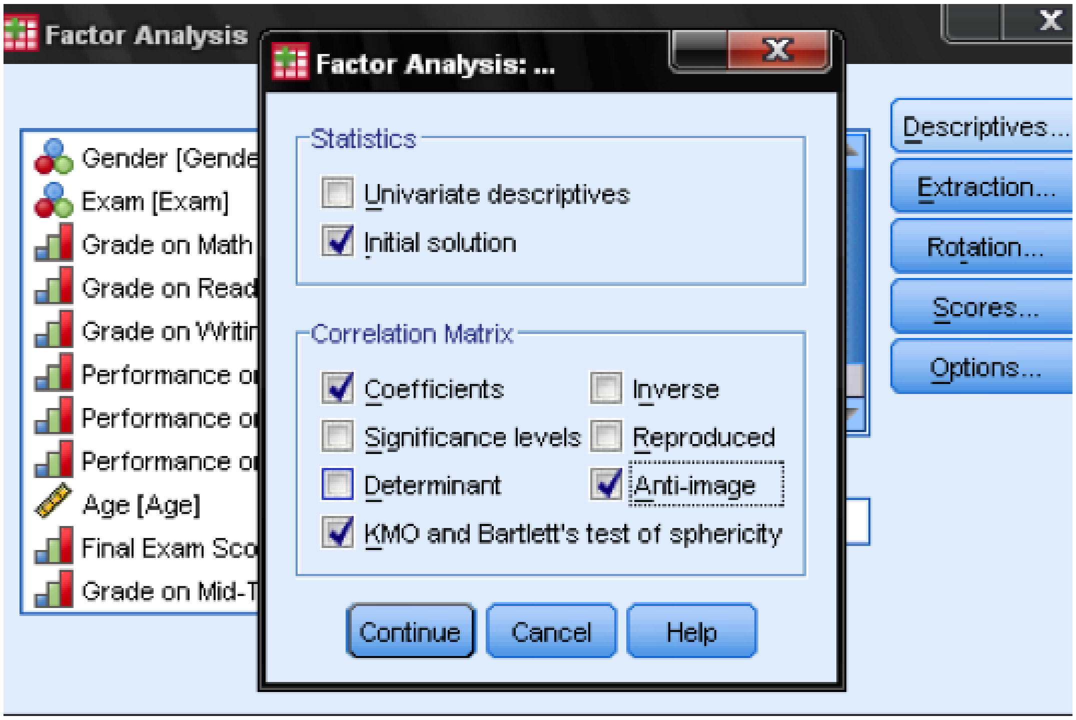
In the dialog box of the *factor analysis* we start by adding our variables (the standardized tests math, reading, and writing, as well as the aptitude tests 1-5) to the list of variables.



Variables:

- Math Test [Test_Sc...
- Reading Test [Test2...
- Writing Test [Test3_...
- Aptitude Test 1 [Apt1]
- Aptitude Test 2 [Apt2]
- Aptitude Test 3 [Apt3]
- Aptitude Test 4 [Apt4]

In the dialog *Descriptives…* we need to add a few statistics to verify the assumptions made by the factor analysis. To verify the assumptions, we need the KMO test of sphericity and the Anti-Image Correlation matrix.



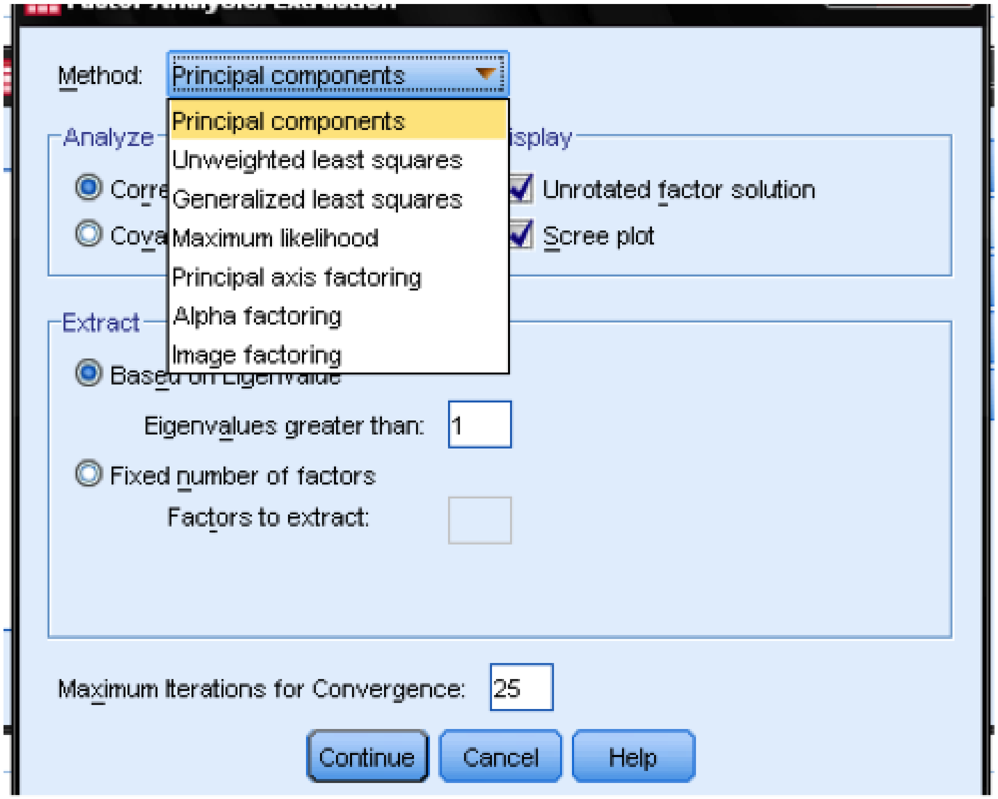
Statistics

□ Univariate descriptives
☑ Initial solution

Correlation Matrix

☑ Coefficients □ Inverse
□ Significance levels □ Reproduced
□ Determinant ☑ Anti-image
☑ KMO and Bartlett's test of sphericity

The dialog box *Extraction…* allows us to specify the extraction method and the cut-off value for the extraction. Generally, SPSS can extract as many factors as we have variables. In an exploratory analysis, the *eigenvalue* is calculated for each factor extracted and can be used to determine the number of factors to extract.  A cutoff value of 1 is generally used to determine factors based on eigenvalues.



Method: Principal components

Analyze

● Corre

○ Cova

Display

☑ Unrotated factor solution

☑ Scree plot

Extract

● Based on Eigenvalue

Eigenvalues greater than: 1

○ Fixed number of factors

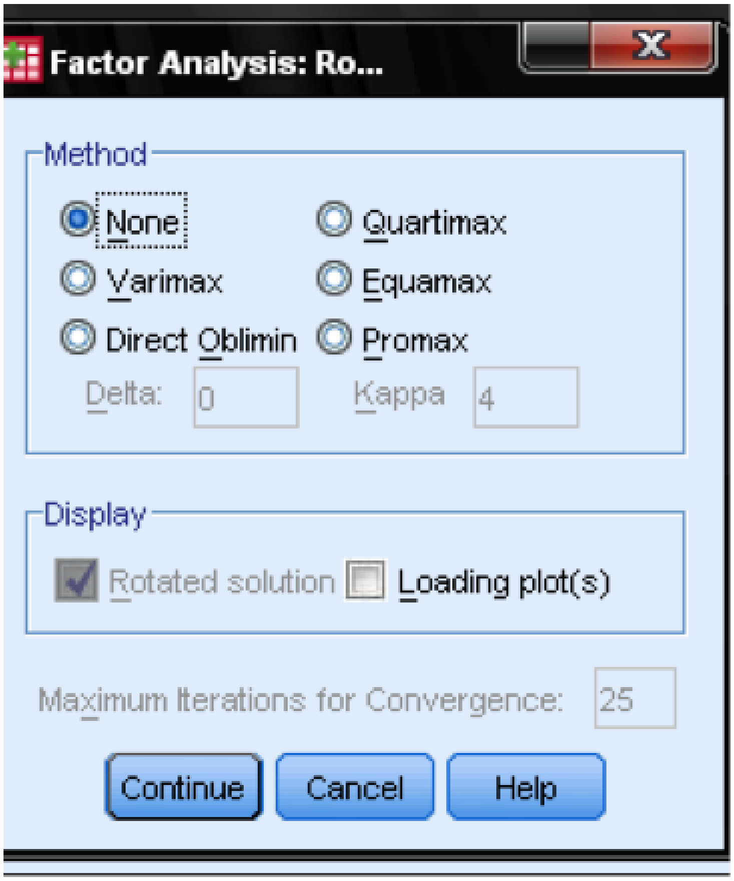
Factors to extract:

Maximum Iterations for Convergence: 25

Next, an appropriate extraction method need to be selected.  *Principal components* is the default extraction method in SPSS.  It extracts uncorrelated linear combinations of the variables and gives the first factor maximum amount of explained variance.  All following factors explain smaller and smaller portions of the variance and are all uncorrelated with each other.  This method is appropriate when the goal is to reduce the data, but is not appropriate when the goal is to identify latent constructs.

The second most common extraction method is *principal axis factoring*.  This method is appropriate when attempting to identify latent constructs, rather than simply reducing the data. In our research question, we are interested in the dimensions behind the variables, and therefore we are going to use principal axis factoring.

The next step is to select a rotation method.  After extracting the factors, SPSS can rotate the factors to better fit the data.  The most commonly used method is *varimax*.  *Varimax* is an orthogonal rotation method that tends produce factor loading that are either very high or very low, making it easier to match each item with a single factor.  If non-orthogonal factors are desired (i.e., factors that can be correlated), a *direct oblimin* rotation is appropriate.  Here, we choose varimax.



Method

● None ○ Quartimax
○ Varimax ○ Equamax
○ Direct Oblimin ○ Promax

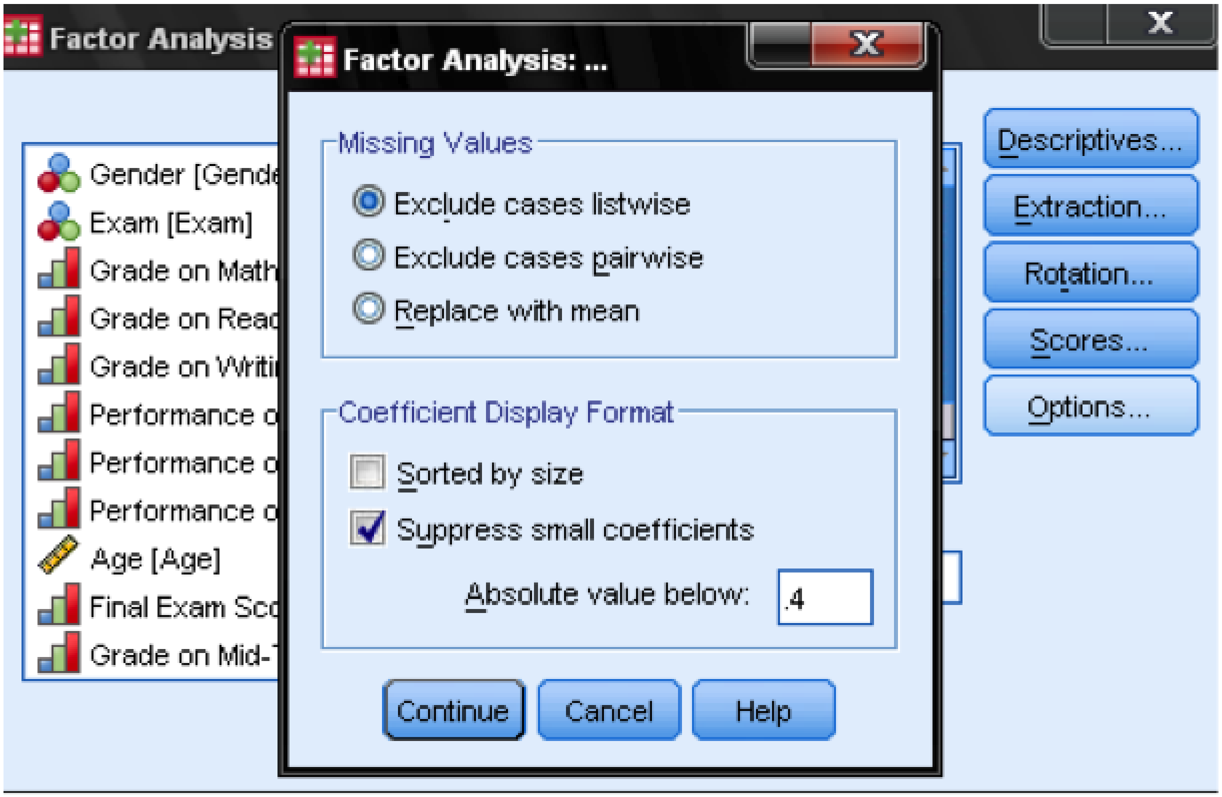
Delta: 0 Kappa 4

Display

☑ Rotated solution □ Loading plot(s)

Maximum Iterations for Convergence: 25

In the dialog box *Options* we can manage how missing values are treated – it might be appropriate to replace them with the mean, which does not change the correlation matrix but ensures that we do not over penalize missing values. Also, we can specify in the output if we do not want to display all factor loadings. The factor loading tables are much easier to read when we suppress small factor loadings. Default value is 0.1, but in this case, we will increase this value to 0.4. The last step would be to save the results in the *Scores…* dialog. This automatically creates standardized scores representing each extracted factor.



Missing Values

● Exclude cases listwise

○ Exclude cases pairwise

○ Replace with mean

Coefficient Display Format

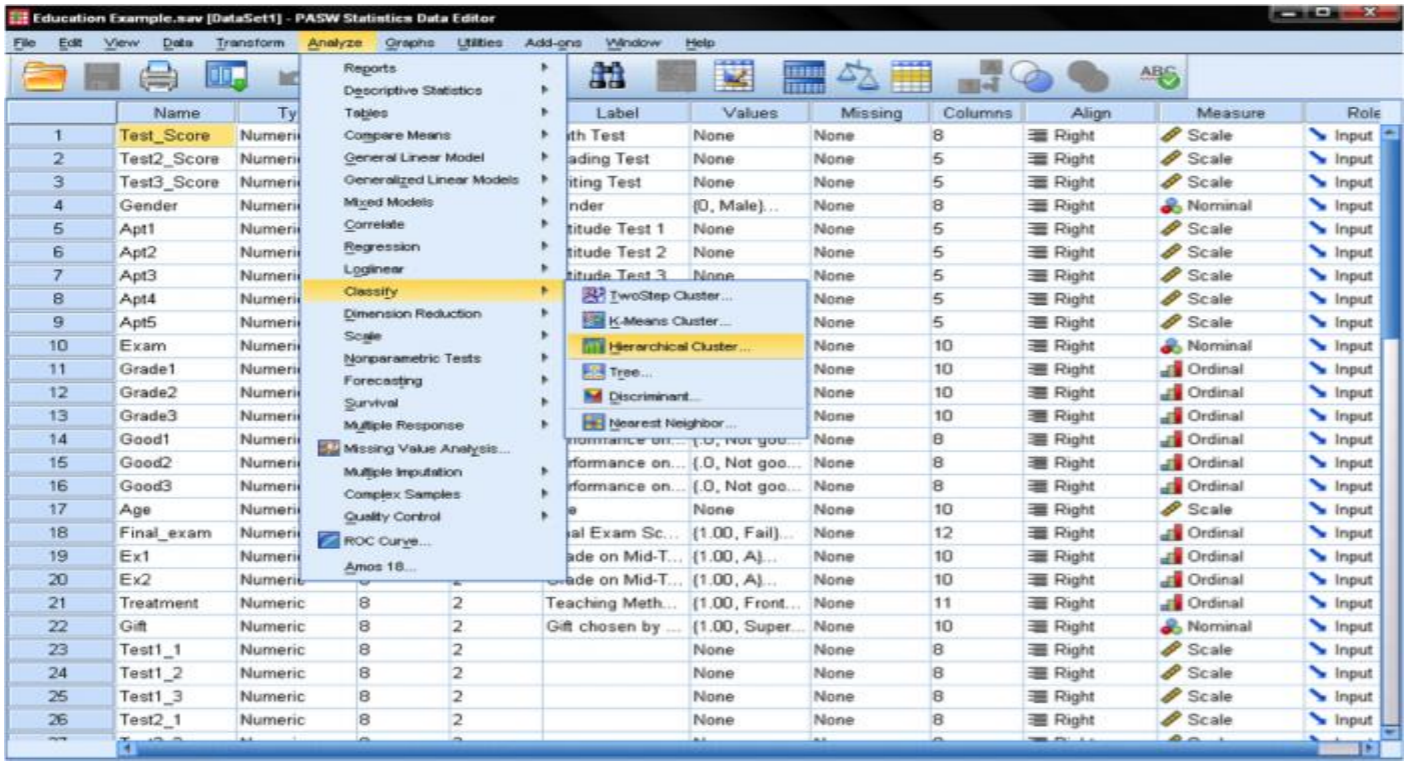
□ Sorted by size

☑ Suppress small coefficients

Absolute value below: .4

# Experiment 4: The Cluster Analysis in SPSS

Our research question for this example cluster analysis is as follows:
What homogenous clusters of students emerge based on standardized test scores in mathematics, reading, and writing?
In SPSS Cluster Analyses can be found in *Analyze/Classify…*. SPSS offers three methods for the cluster analysis: *K-Means Cluster*, *Hierarchical Cluster*, and *Two-Step Cluster*.
*K-means cluster* is a method to quickly cluster large data sets. The researcher define the number of clusters in advance. This is useful to test different models with a different assumed number of clusters.
*Hierarchical cluster* is the most common method. It generates a series of models with cluster solutions from *1* (all cases in one cluster) to *n* (each case is an individual cluster). Hierarchical cluster also works with variables as opposed to cases; it can cluster variables together in a manner somewhat similar to factor analysis. In addition, hierarchical cluster analysis can handle nominal, ordinal, and scale data; however it is not recommended to mix different levels of measurement.
*Two-step cluster* analysis identifies groupings by running pre-clustering first and then by running hierarchical methods. Because it uses a quick cluster algorithm upfront, it can handle large data sets that would take a long time to compute with hierarchical cluster methods. In this respect, it is a combination of the previous two approaches. Two-step clustering can handle scale and ordinal data in the same model, and it automatically selects the number of clusters.

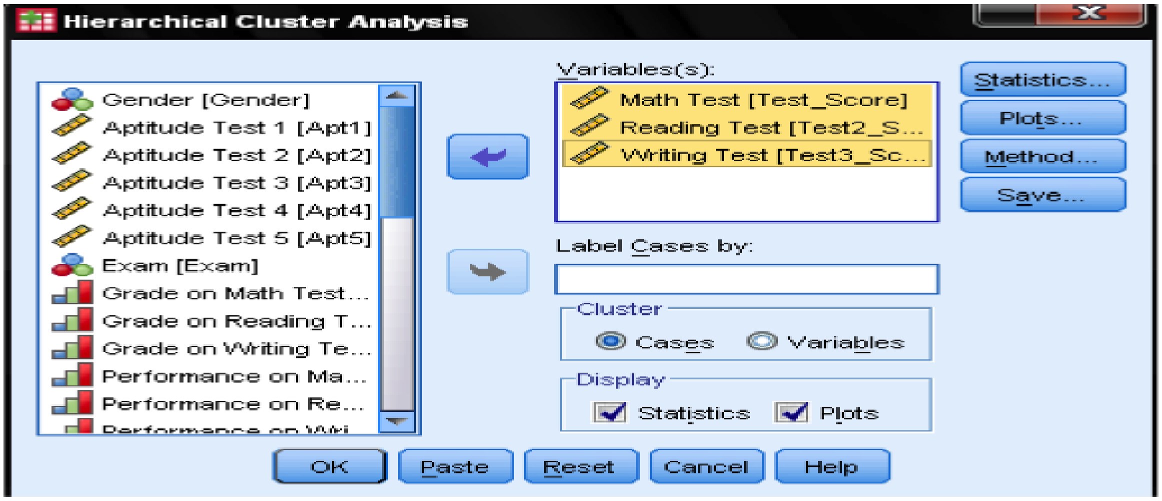


Education Example.sav [DataSet1] - PASW Statistics Data Editor

| | Name | Ty | | | Label | Values | Missing | Columns | Align | Measure | Role |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Test_Score | Numeri | | | th Test | None | None | 8 | Right | Scale | Input |
| 2 | Test2_Score | Numeri | | | ading Test | None | None | 5 | Right | Scale | Input |
| 3 | Test3_Score | Numeri | | | iting Test | None | None | 5 | Right | Scale | Input |
| 4 | Gender | Numeri | | | nder | {0, Male}... | None | 8 | Right | Nominal | Input |
| 5 | Apt1 | Numeri | | | titude Test 1 | None | None | 5 | Right | Scale | Input |
| 6 | Apt2 | Numeri | | | titude Test 2 | None | None | 5 | Right | Scale | Input |
| 7 | Apt3 | Numeri | | | titude Test 3 | None | None | 5 | Right | Scale | Input |
| 8 | Apt4 | Numeri | | | | | None | 5 | Right | Scale | Input |
| 9 | Apt5 | Numeri | | | | | None | 5 | Right | Scale | Input |
| 10 | Exam | Numeri | | | | | None | 10 | Right | Nominal | Input |
| 11 | Grade1 | Numeri | | | | | None | 10 | Right | Ordinal | Input |
| 12 | Grade2 | Numeri | | | | | None | 10 | Right | Ordinal | Input |
| 13 | Grade3 | Numeri | | | | | None | 10 | Right | Ordinal | Input |
| 14 | Good1 | Numeri | | | [illegible] | [illegible] | None | 8 | Right | Ordinal | Input |
| 15 | Good2 | Numeri | | | rformance on... | {.0, Not goo... | None | 8 | Right | Ordinal | Input |
| 16 | Good3 | Numeri | | | rformance on... | {.0, Not goo... | None | 8 | Right | Ordinal | Input |
| 17 | Age | Numeri | | | e | None | None | 10 | Right | Scale | Input |
| 18 | Final_exam | Numeri | | | al Exam Sc... | {1.00, Fail}... | None | 12 | Right | Ordinal | Input |
| 19 | Ex1 | Numeri | | | ade on Mid-T... | {1.00, A}... | None | 10 | Right | Ordinal | Input |
| 20 | Ex2 | Numeric | | | ade on Mid-T... | {1.00, A}... | None | 10 | Right | Ordinal | Input |
| 21 | Treatment | Numeric | 8 | 2 | Teaching Meth... | {1.00, Front... | None | 11 | Right | Ordinal | Input |
| 22 | Gift | Numeric | 8 | 2 | Gift chosen by ... | {1.00, Super... | None | 10 | Right | Nominal | Input |
| 23 | Test1_1 | Numeric | 8 | 2 | | None | None | 8 | Right | Scale | Input |
| 24 | Test1_2 | Numeric | 8 | 2 | | None | None | 8 | Right | Scale | Input |
| 25 | Test1_3 | Numeric | 8 | 2 | | None | None | 8 | Right | Scale | Input |
| 26 | Test2_1 | Numeric | 8 | 2 | | None | None | 8 | Right | Scale | Input |

The hierarchical cluster analysis follows three basic steps: 1) calculate the distances, 2) link the clusters, and 3) choose a solution by selecting the right number of clusters.

First, we have to select the variables upon which we base our clusters. In the dialog window we add the math, reading, and writing tests to the list of variables. Since we want to cluster cases we leave the rest of the tick marks on the default.



Hierarchical Cluster Analysis

Gender [Gender]
Aptitude Test 1 [Apt1]
Aptitude Test 2 [Apt2]
Aptitude Test 3 [Apt3]
Aptitude Test 4 [Apt4]
Aptitude Test 5 [Apt5]
Exam [Exam]
Grade on Math Test...
Grade on Reading T...
Grade on Writing Te...
Performance on Ma...
Performance on Re...
Performance on Wri...

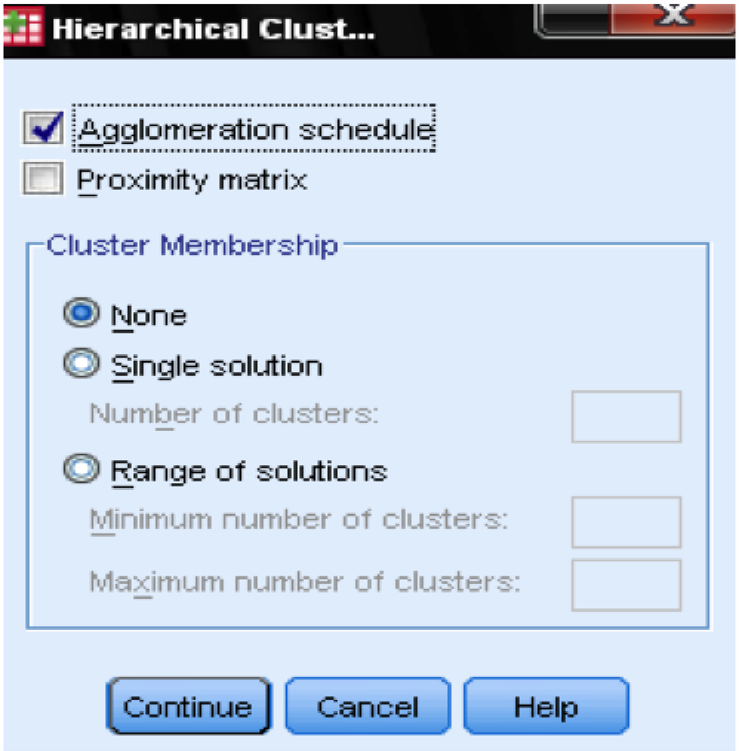
Variables(s):
Math Test [Test_Score]
Reading Test [Test2_S...
Writing Test [Test3_Sc...

Label Cases by:

Cluster
● Cases ○ Variables

Display
☑ Statistics ☑ Plots

In the dialog box *Statistics…* we can specify whether we want to output the proximity matrix (these are the distances calculated in the first step of the analysis) and the predicted cluster membership of the cases in our observations. Again, we leave all settings on default.

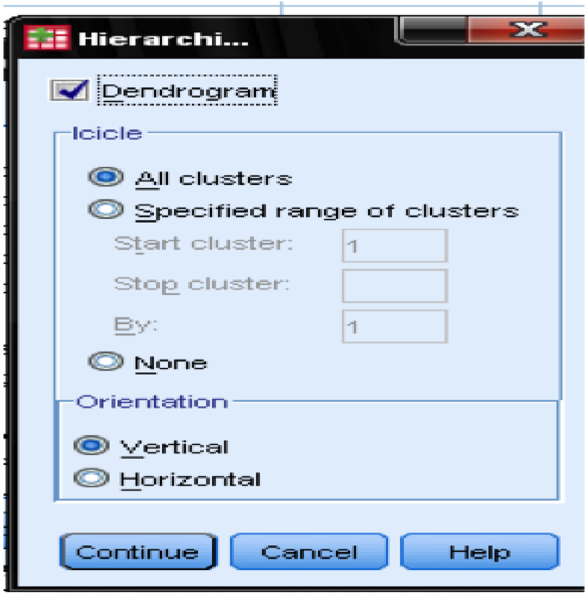


Hierarchical Clust...

☑ Agglomeration schedule
□ Proximity matrix

Cluster Membership
● None
○ Single solution
Number of clusters:
○ Range of solutions
Minimum number of clusters:
Maximum number of clusters:

In the dialog box *Plots…* we should add the *Dendrogram*. The *Dendrogram* will graphically show how the clusters are merged and allows us to identify what the appropriate number of clusters is.



Hierarchi...

☑ Dendrogram

Icicle

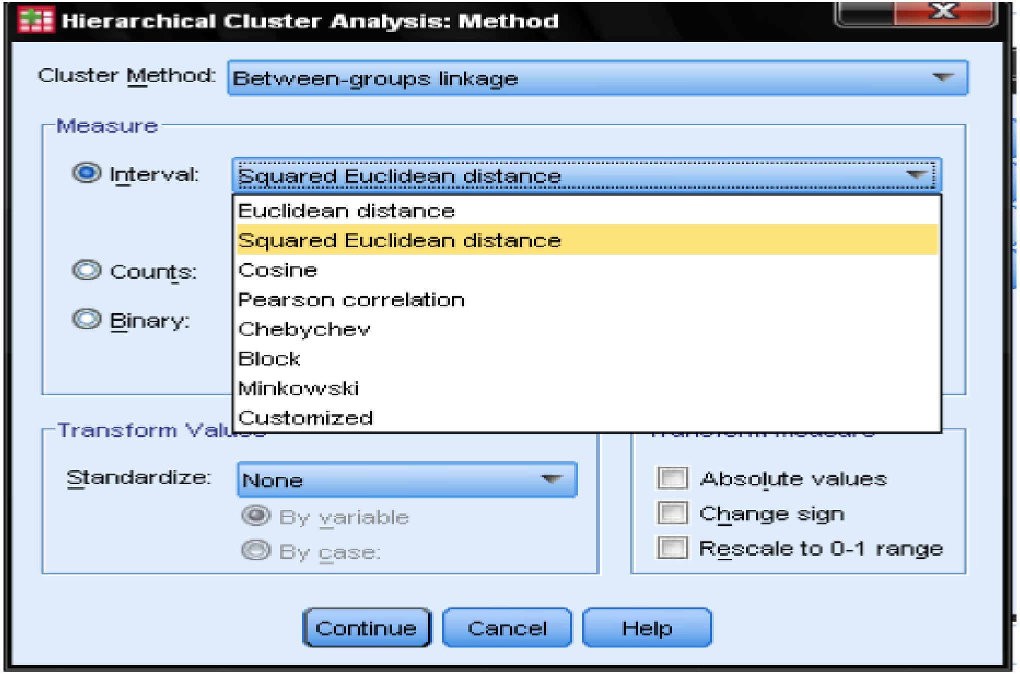
● All clusters
○ Specified range of clusters
Start cluster: 1
Stop cluster:
By: 1
○ None

Orientation

● Vertical
○ Horizontal

The dialog box *Method…* allows us to specify the distance measure and the clustering method. First, we need to define the correct distance measure. SPSS offers three large blocks of distance measures for interval (scale), counts (ordinal), and binary (nominal) data.

For interval data, the most common is *Square Euclidian Distance*. It is based on the Euclidian Distance between two observations, which is the square root of the sum of squared distances. Since the Euclidian Distance is squared, it increases the importance of large distances, while weakening the importance of small distances.



Hierarchical Cluster Analysis: Method

Cluster Method: Between-groups linkage

Measure

● Interval: Squared Euclidean distance
○ Counts:
○ Binary:

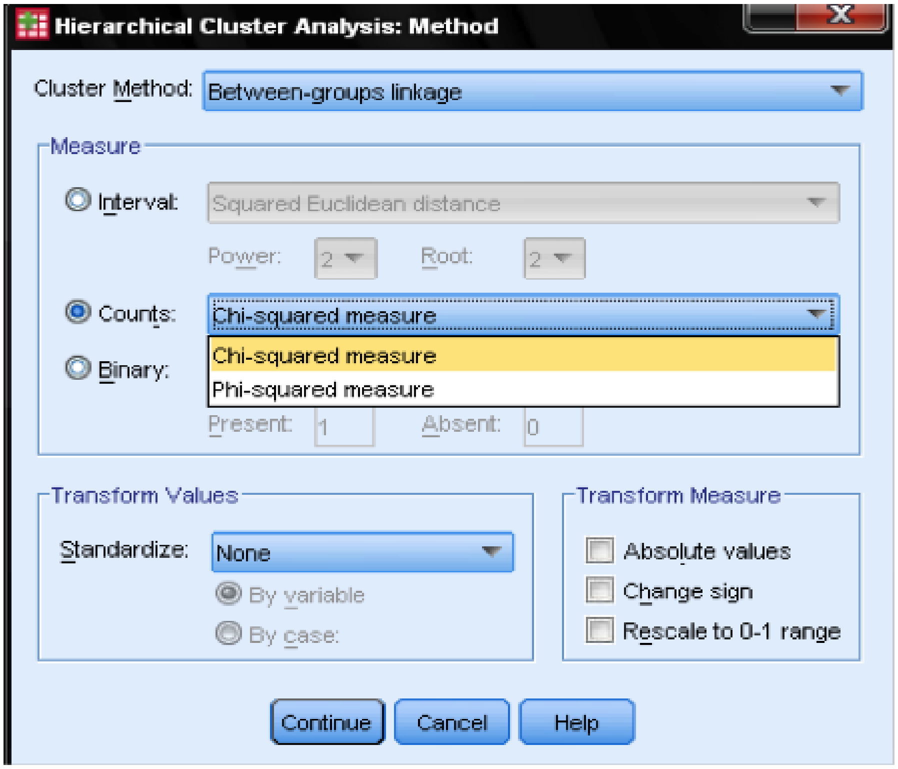
Transform Values

Standardize: None
○ By variable
○ By case:

Transform Measure

□ Absolute values
□ Change sign
□ Rescale to 0-1 range

If we have ordinal data (counts) we can select between Chi-Square or a standardized Chi-Square called *Phi-Square*. For binary data, the Squared Euclidean Distance is commonly used.



Hierarchical Cluster Analysis: Method

Cluster Method: Between-groups linkage

Measure

○ Interval: Squared Euclidean distance

Power: 2 Root: 2

● Counts: Chi-squared measure

○ Binary:

Present: 1 Absent: 0

Transform Values

Standardize: None

○ By variable

○ By case:

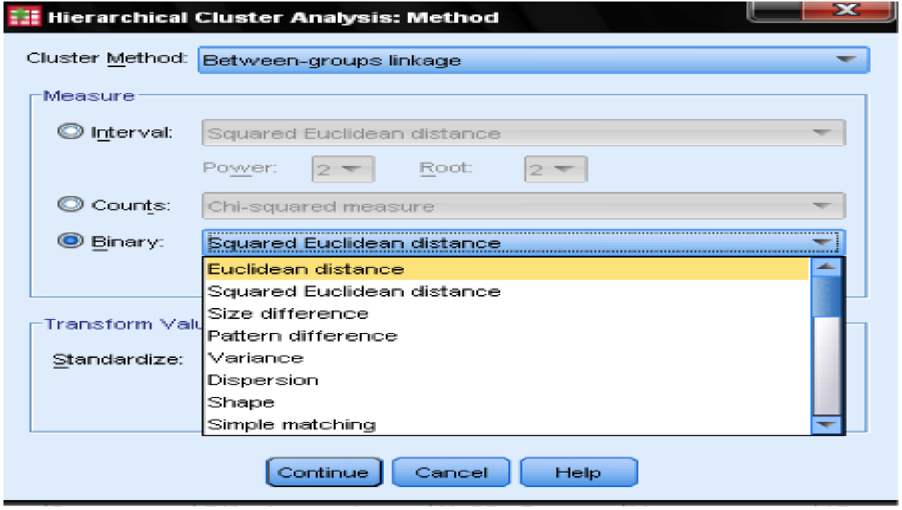
Transform Measure

□ Absolute values

□ Change sign

□ Rescale to 0-1 range

In our example, we choose *Interval* and *Square Euclidean Distance*.



Hierarchical Cluster Analysis: Method

Cluster Method: Between-groups linkage

Measure

○ Interval: Squared Euclidean distance

Power: 2 Root: 2

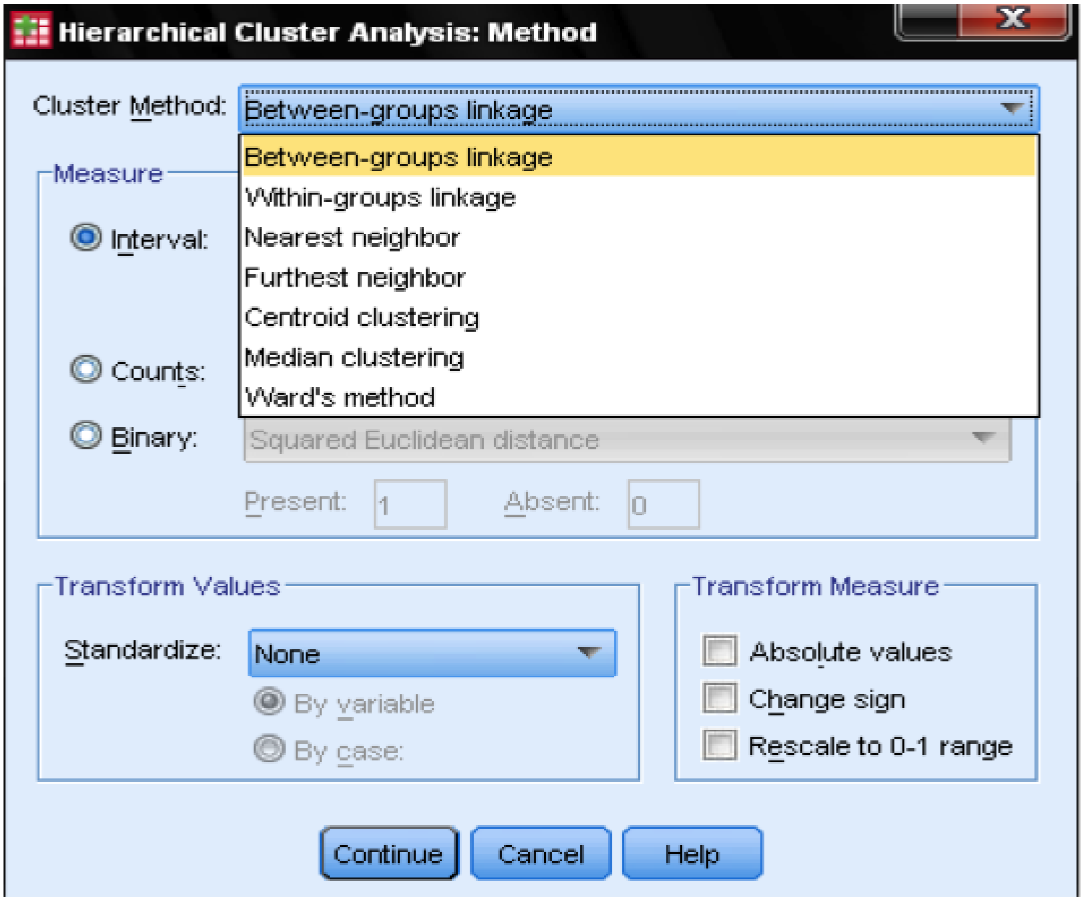
○ Counts: Chi-squared measure

● Binary: Squared Euclidean distance

Transform Valu

Standardize:

Next, we have to choose the *Cluster Method*. Typically, choices are between-groups linkage (distance between clusters is the average distance of all data points within these clusters), nearest neighbor (single linkage: distance between clusters is the smallest distance between two data points), furthest neighbor (complete linkage: distance is the largest distance between two data points), and Ward's method (distance is the distance of all clusters to the grand average of the sample). Single linkage works best with long chains of clusters, while complete linkage works best with dense blobs of clusters. Between-groups linkage works with both cluster types. It is recommended is to use single linkage first. Although single linkage tends to create chains of clusters, it helps in identifying outliers.  After excluding these outliers, we can move onto Ward's method.  Ward's method uses the *F* value (like in ANOVA) to maximize the significance of differences between clusters.



Cluster Method: Between-groups linkage

Measure

● Interval:

○ Counts:

○ Binary: Squared Euclidean distance

Present: 1 Absent: 0

Transform Values

Standardize: None

● By variable

○ By case:

Transform Measure

□ Absolute values

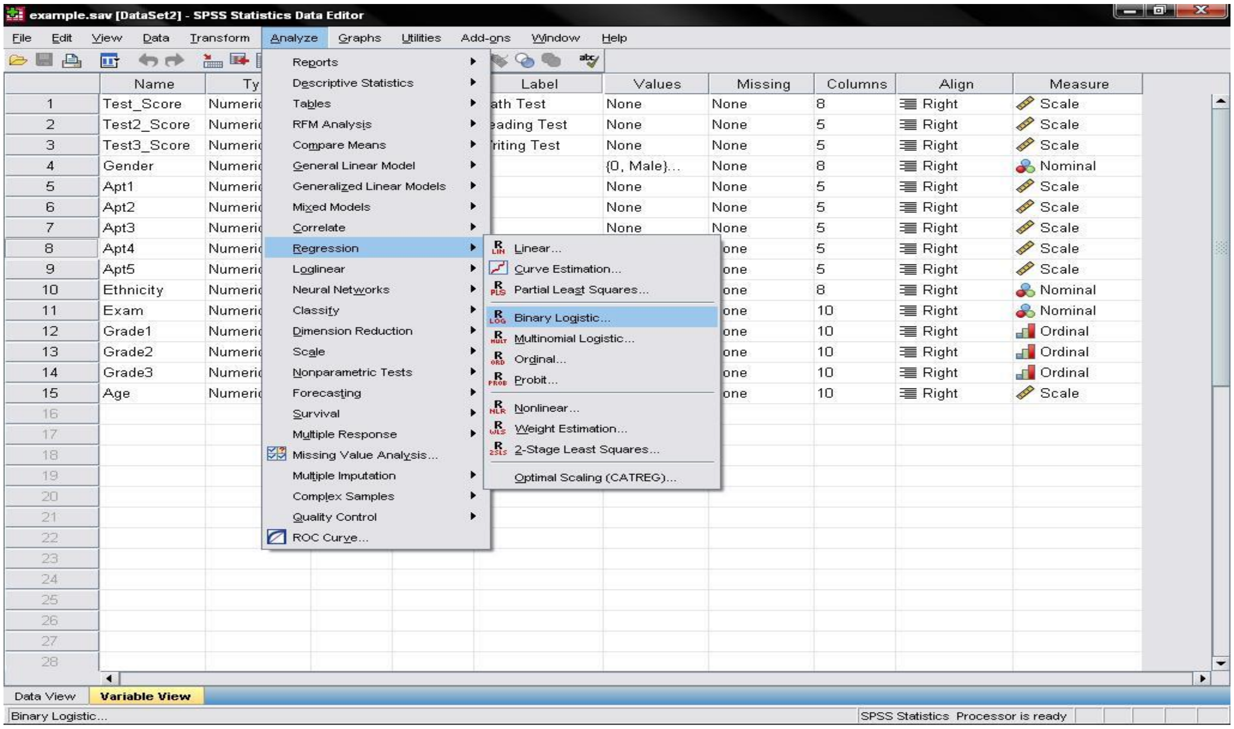
□ Change sign

□ Rescale to 0-1 range

A last consideration is *standardization*.  If the variables have different scales and means we might want to standardize either to *Z scores* or by centering the scale.  We can also transform the values to absolute values if we have a data set where this might be appropriate.

# Experiment 5: The Logistic Regression Analysis in SPSS

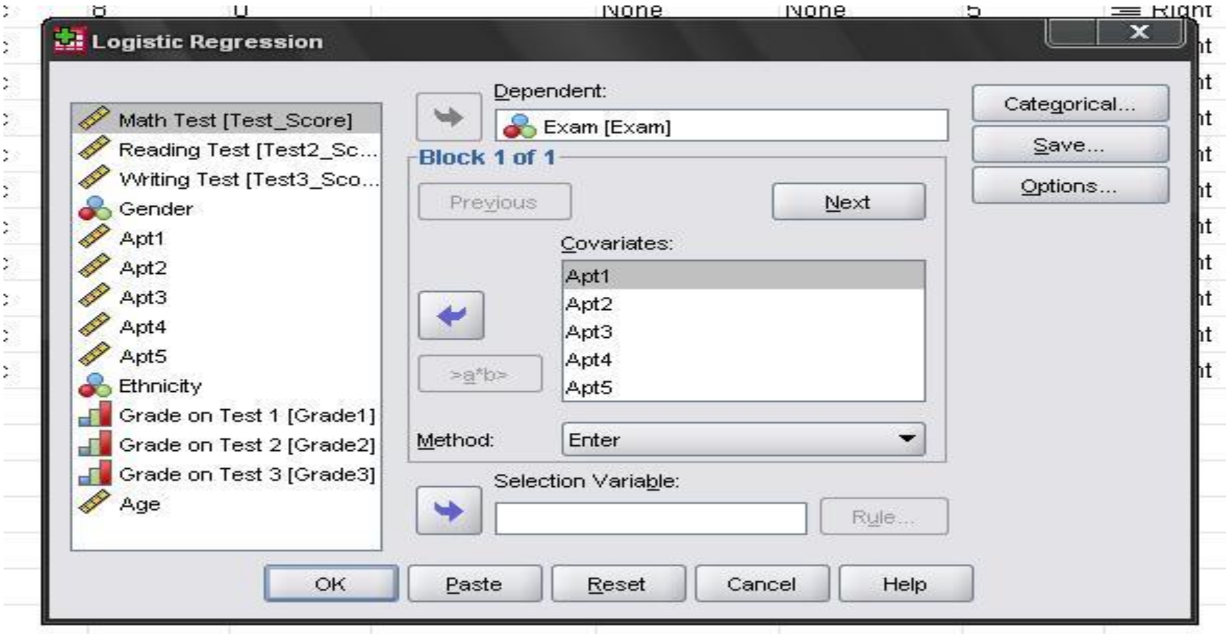
First we need to check that all cells in our model are populated. Although the logistic regression is robust against multivariate normality and therefore better suited for smaller samples than a probit model, we still need to check, because we don't have any categorical variables in our design we will skip this step.
Logistic Regression is found in SPSS under Analyze/Regression/Binary Logistic…



| | Name | Ty | Label | Values | Missing | Columns | Align | Measure |
|---|---|---|---|---|---|---|---|---|
| 1 | Test_Score | Numeric | ath Test | None | None | 8 | Right | Scale |
| 2 | Test2_Score | Numeric | eading Test | None | None | 5 | Right | Scale |
| 3 | Test3_Score | Numeric | riting Test | None | None | 5 | Right | Scale |
| 4 | Gender | Numeric | | {0, Male}... | None | 8 | Right | Nominal |
| 5 | Apt1 | Numeric | | None | None | 5 | Right | Scale |
| 6 | Apt2 | Numeric | | None | None | 5 | Right | Scale |
| 7 | Apt3 | Numeric | | None | None | 5 | Right | Scale |
| 8 | Apt4 | Numeric | | | one | 5 | Right | Scale |
| 9 | Apt5 | Numeric | | | one | 5 | Right | Scale |
| 10 | Ethnicity | Numeric | | | one | 8 | Right | Nominal |
| 11 | Exam | Numeric | | | one | 10 | Right | Nominal |
| 12 | Grade1 | Numeric | | | one | 10 | Right | Ordinal |
| 13 | Grade2 | Numeric | | | one | 10 | Right | Ordinal |
| 14 | Grade3 | Numeric | | | one | 10 | Right | Ordinal |
| 15 | Age | Numeric | | | one | 10 | Right | Scale |

Binary Logistic...
SPSS Statistics Processor is ready

This opens the dialogue box to specify the model



Dependent: Exam [Exam]

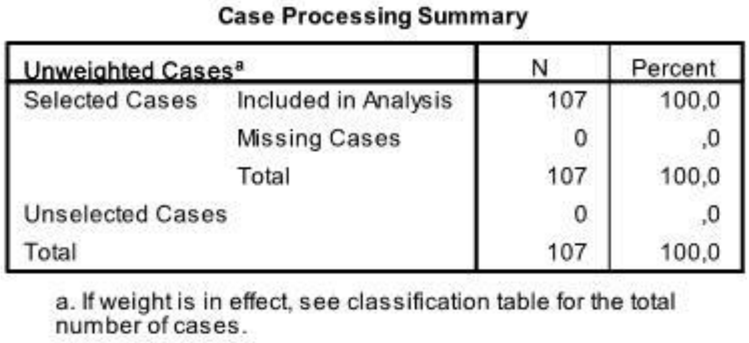
Covariates: Apt1, Apt2, Apt3, Apt4, Apt5

Method: Enter

Here we need to enter the nominal variable Exam (pass = 1, fail = 0) into the dependent variable box and we enter all aptitude tests as the first block of covariates in the model.

The menu categorical… allows to specify contrasts for categorical variables (which we do not have in our logistic regression model), and options offers several additional statistics, which don't need.

The first table just shows the sample size.

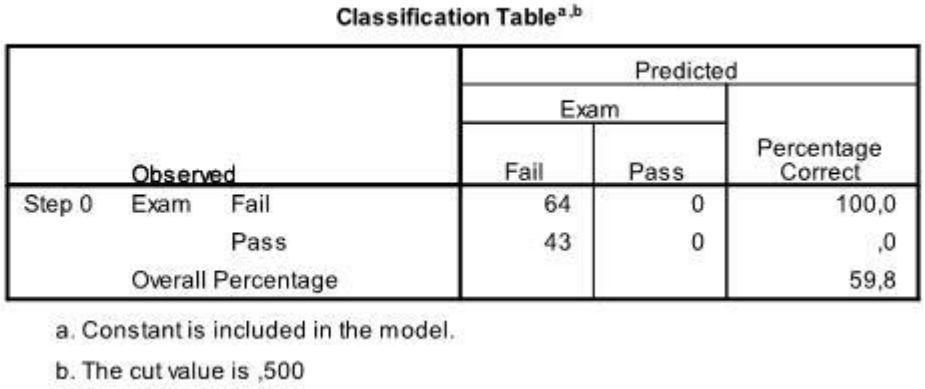


**Case Processing Summary**

| Unweighted Cases[a] | | N | Percent |
|---|---|---|---|
| Selected Cases | Included in Analysis | 107 | 100,0 |
| | Missing Cases | 0 | ,0 |
| | Total | 107 | 100,0 |
| Unselected Cases | | 0 | ,0 |
| Total | | 107 | 100,0 |

a. If weight is in effect, see classification table for the total number of cases.

The next 3 tables are the results fort he intercept model. That is the Maximum Likelihood model if only the intercept is included without any of the dependent variables in the analysis. This is basically only interesting to calculate the Pseudo R² that describe the goodness of fit for the logistic model.

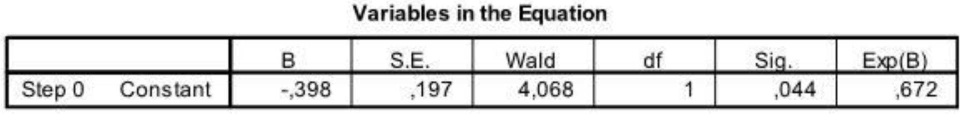


**Classification Table[a,b]**

| | | | Predicted | | |
|---|---|---|---|---|---|
| | | | Exam | | Percentage Correct |
| Observed | | | Fail | Pass | |
| Step 0 | Exam | Fail | 64 | 0 | 100,0 |
| | | Pass | 43 | 0 | ,0 |
| | Overall Percentage | | | | 59,8 |

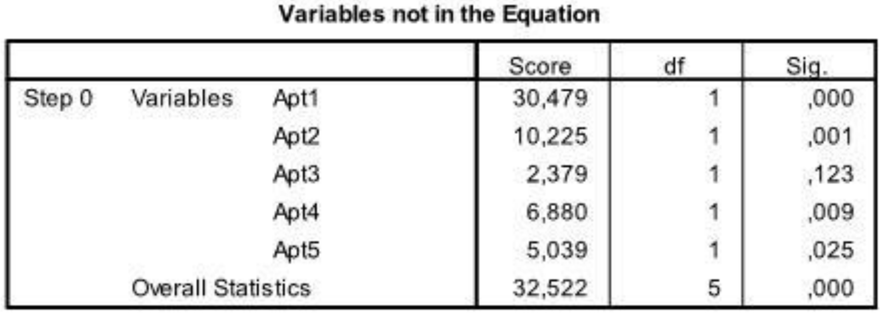
a. Constant is included in the model.

b. The cut value is ,500



**Variables in the Equation**

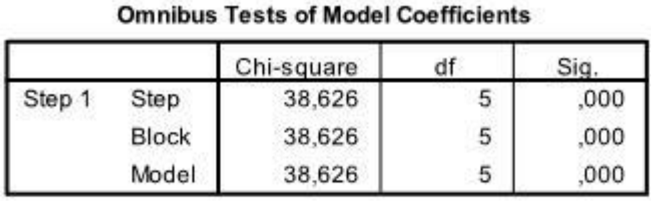
| | | B | S.E. | Wald | df | Sig. | Exp(B) |
|---|---|---|---|---|---|---|---|
| Step 0 | Constant | -,398 | ,197 | 4,068 | 1 | ,044 | ,672 |



**Variables not in the Equation**

| | | | Score | df | Sig. |
|---|---|---|---|---|---|
| Step 0 | Variables | Apt1 | 30,479 | 1 | ,000 |
| | | Apt2 | 10,225 | 1 | ,001 |
| | | Apt3 | 2,379 | 1 | ,123 |
| | | Apt4 | 6,880 | 1 | ,009 |
| | | Apt5 | 5,039 | 1 | ,025 |
| | Overall Statistics | | 32,522 | 5 | ,000 |

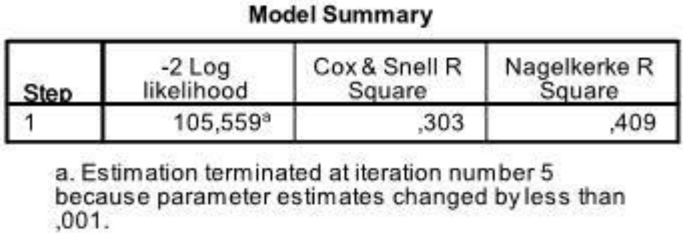
The relevant tables can be found in the section 'Block 1' in the SPSS output of our logistic regression analysis. The first table includes the Chi-Square goodness of fit test. It has the null hypothesis that intercept and all coefficients are zero. We can reject this null hypothesis.



**Omnibus Tests of Model Coefficients**

| | | Chi-square | df | Sig. |
|---|---|---|---|---|
| Step 1 | Step | 38,626 | 5 | ,000 |
| | Block | 38,626 | 5 | ,000 |
| | Model | 38,626 | 5 | ,000 |

The next table includes the Pseudo R², the -2 log likelihood is the minimization criteria used by SPSS. We see that Nagelkerke's R² is 0.409 which indicates that the model is good but not great. Cox & Snell's R² is the nth root (in our case the 107th of the -2log likelihood improvement. Thus we can interpret this as 30% probability of the event passing the exam is explained by the logistic model.

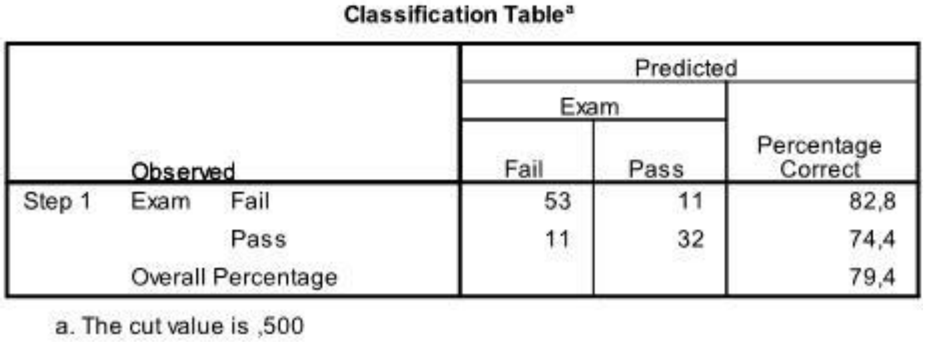


**Model Summary**

| Step | -2 Log likelihood | Cox & Snell R Square | Nagelkerke R Square |
|---|---|---|---|
| 1 | 105,559[a] | ,303 | ,409 |

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than ,001.

The next table contains the classification results, with almost 80% correct classification the model is not too bad – generally a discriminant analysis is better in classifying data correctly.

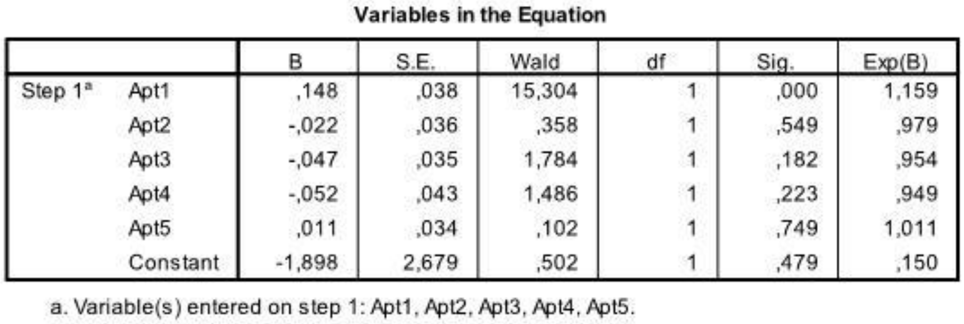


**Classification Table[a]**

| Observed | | | Predicted | | |
|---|---|---|---|---|---|
| | | | Exam | | Percentage Correct |
| | | | Fail | Pass | |
| Step 1 | Exam | Fail | 53 | 11 | 82,8 |
| | | Pass | 11 | 32 | 74,4 |
| | Overall Percentage | | | | 79,4 |

a. The cut value is ,500

The last table is the most important one for our logistic regression analysis. It shows the regression function -1.898 + .148*x1 – .022*x2 – .047*x3 – .052*x4 + .011*x5. The table also includes the test of significance for each of the coefficients in the logistic regression model. For small samples the t-values are not valid and the Wald statistic should be used instead. Wald is basically t² which is Chi-Square distributed with df=1. However, SPSS gives the significance levels of each coefficient. As we can see, only Apt1 is significant all other variables are not.



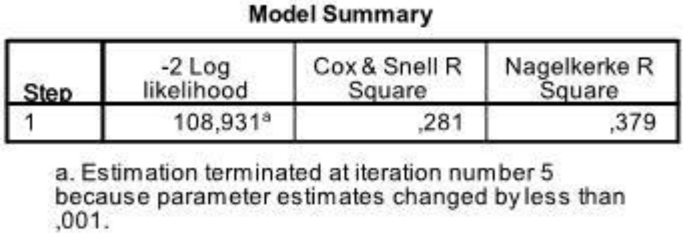
**Variables in the Equation**

| | | B | S.E. | Wald | df | Sig. | Exp(B) |
|---|---|---|---|---|---|---|---|
| Step 1[a] | Apt1 | ,148 | ,038 | 15,304 | 1 | ,000 | 1,159 |
| | Apt2 | -,022 | ,036 | ,358 | 1 | ,549 | ,979 |
| | Apt3 | -,047 | ,035 | 1,784 | 1 | ,182 | ,954 |
| | Apt4 | -,052 | ,043 | 1,486 | 1 | ,223 | ,949 |
| | Apt5 | ,011 | ,034 | ,102 | 1 | ,749 | 1,011 |
| | Constant | -1,898 | 2,679 | ,502 | 1 | ,479 | ,150 |

a. Variable(s) entered on step 1: Apt1, Apt2, Apt3, Apt4, Apt5.

If we change the method from Enter to Forward:Wald the quality of the logistic regression improves. Now only the significant coefficients are included in the logistic regression equation. In our case this is Apt1 and the intercept.

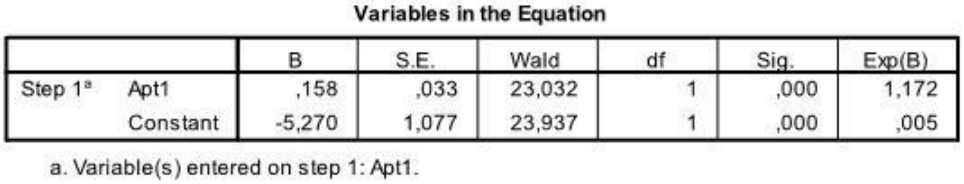
We see that $p = \frac{1}{1+e^{-(-5.270+.158 Apt1)}}$, and we know that a 1 point higher score in the Apt1 test multiplies the odds of passing the exam by 1.17 (exp(.158)). We can also calculate the critical value which is Apt1 > -intercept/coefficient > -5.270/.158 > 33.35. That is if a pupil scored higher than 33.35 on the Aptitude Test 1 the logistic regression predicts that this pupil will pass the final exam.



**Model Summary**

| Step | -2 Log likelihood | Cox & Snell R Square | Nagelkerke R Square |
|---|---|---|---|
| 1 | 108,931[a] | ,281 | ,379 |

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than ,001.



**Variables in the Equation**

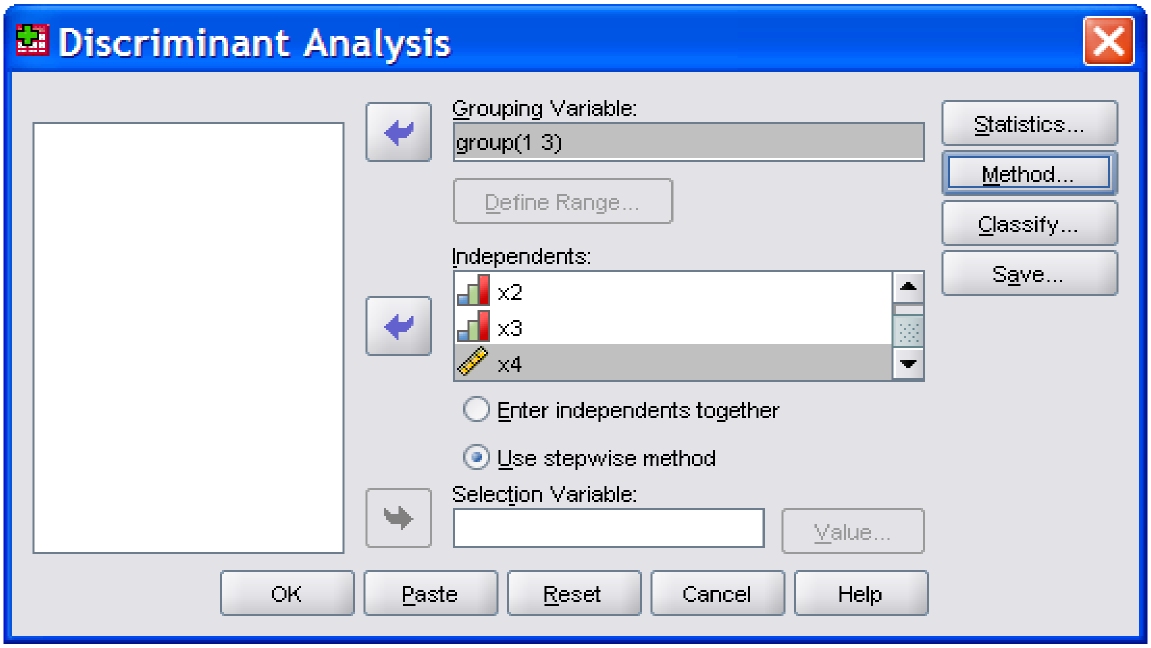
| | | B | S.E. | Wald | df | Sig. | Exp(B) |
|---|---|---|---|---|---|---|---|
| Step 1[a] | Apt1 | ,158 | ,033 | 23,032 | 1 | ,000 | 1,172 |
| | Constant | -5,270 | 1,077 | 23,937 | 1 | ,000 | ,005 |

a. Variable(s) entered on step 1: Apt1.

# Experiment 6: Discriminant Analysis

SPSS will do **stepwise DFA**. You simply specify which method you wish to employ for selecting predictors. The most economical method is the **Wilks lambda** method," which selects predictors that **minimize Wilks lambda**. As with stepwise multiple regression, you may set the **criteria for entry and removal** (*F* criteria or *p* criteria), or you may take the defaults.

Imagine that you are working as a statistician for the Internal Revenue Service. You are told that another IRS employee has developed four composite scores ($X_1$ - $X_4$), easily computable from the information that taxpayers provide on their income tax returns and from other databases to which the IRS has access. These composite scores were developed in the hope that they would be useful for discriminating tax cheaters from other persons. To see if these composite scores actually have any predictive validity, the IRS selects a random sample of taxpayers and audits their returns. Based on this audit, each taxpayer is placed into one of three groups: Group 1 is persons who overpaid their taxes by a considerable amount, Group 2 is persons who paid the correct amount, and Group 3 is persons who underpaid their taxes by a considerable amount. $X_1$ through $X_4$ are then computed for each of these taxpayers. You are given a data file with group membership, $X_1$, $X_2$, $X_3$, and $X_4$ for each taxpayer, with an equal number of subjects in each group. Your job is to use discriminant function analysis to develop a pair of discriminant functions (weighted sums of $X_1$ through $X_4$) to predict group membership. You use a fully stepwise selection procedure to develop a (maybe) reduced (less than four predictors) model. You employ the WILKS method of selecting variables to be entered or deleted, using the default *p* criterion for entering and removing variables.



Discriminant Analysis

Grouping Variable: group(1 3)

Independents: x2, x3, x4

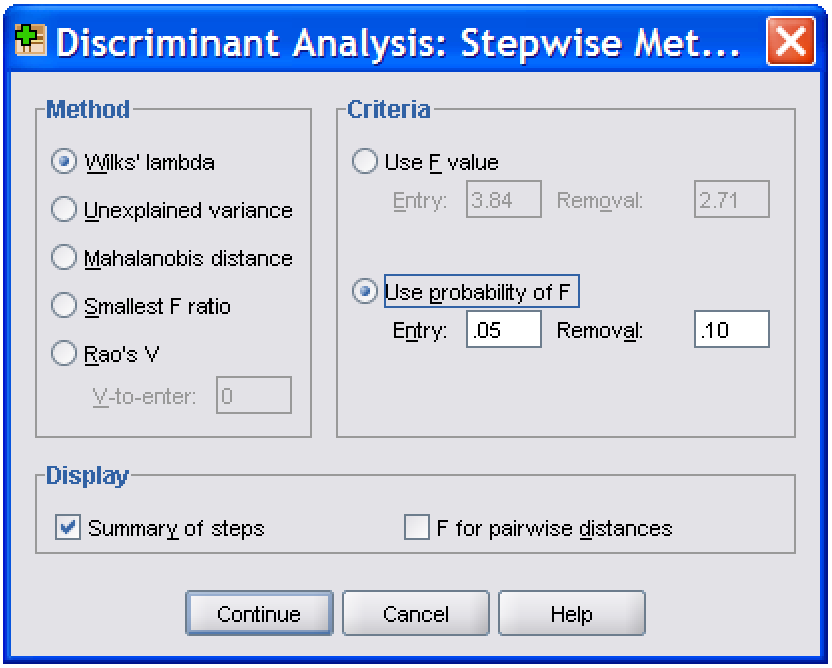
○ Enter independents together

● Use stepwise method

Selection Variable:

Your data file is DFA-STEP.sav, which is available on Karl's SPSS-Data page -- download it and then bring it into SPSS. To do the DFA, click Analyze, Classify, and then put Group into the Grouping Variable box, defining its range from 1 to 3. Put X1 through X4 in the "Independents" box, and select the stepwise method.

Click Method and select "Wilks' lambda" and "Use probability of *F*." Click Continue.



Discriminant Analysis: Stepwise Met...

Method

● Wilks' lambda

○ Unexplained variance

○ Mahalanobis distance

○ Smallest F ratio

○ Rao's V

V-to-enter: 0

Criteria

○ Use F value

Entry: 3.84 Removal: 2.71

● Use probability of F

Entry: .05 Removal: .10

Display

☑ Summary of steps

□ F for pairwise distances

Under Statistics, ask for the group means. Under Classify, ask for a territorial map. Continue, OK.

Look at the output, "Variables Not in the Analysis." At Step 0 the tax groups (overpaid, paid correct, underpaid) differ most on $X_3$ ($\Lambda$ drops to .636 if $X_3$ is entered) and "Sig. of *F* to enter" is less than .05, so that predictor is entered first. After entering $X_3$, all remaining predictors are eligible for entry, but $X_1$ most reduces lambda, so it enters. The Wilks lambda is reduced from .635 to .171. On the next step, only $X_2$ is eligible to enter, and it does, lowering Wilks lambda to .058. At this point no variable already in meets the criterion for removal and no variable out meets the criterion for entry, so the analysis stops.

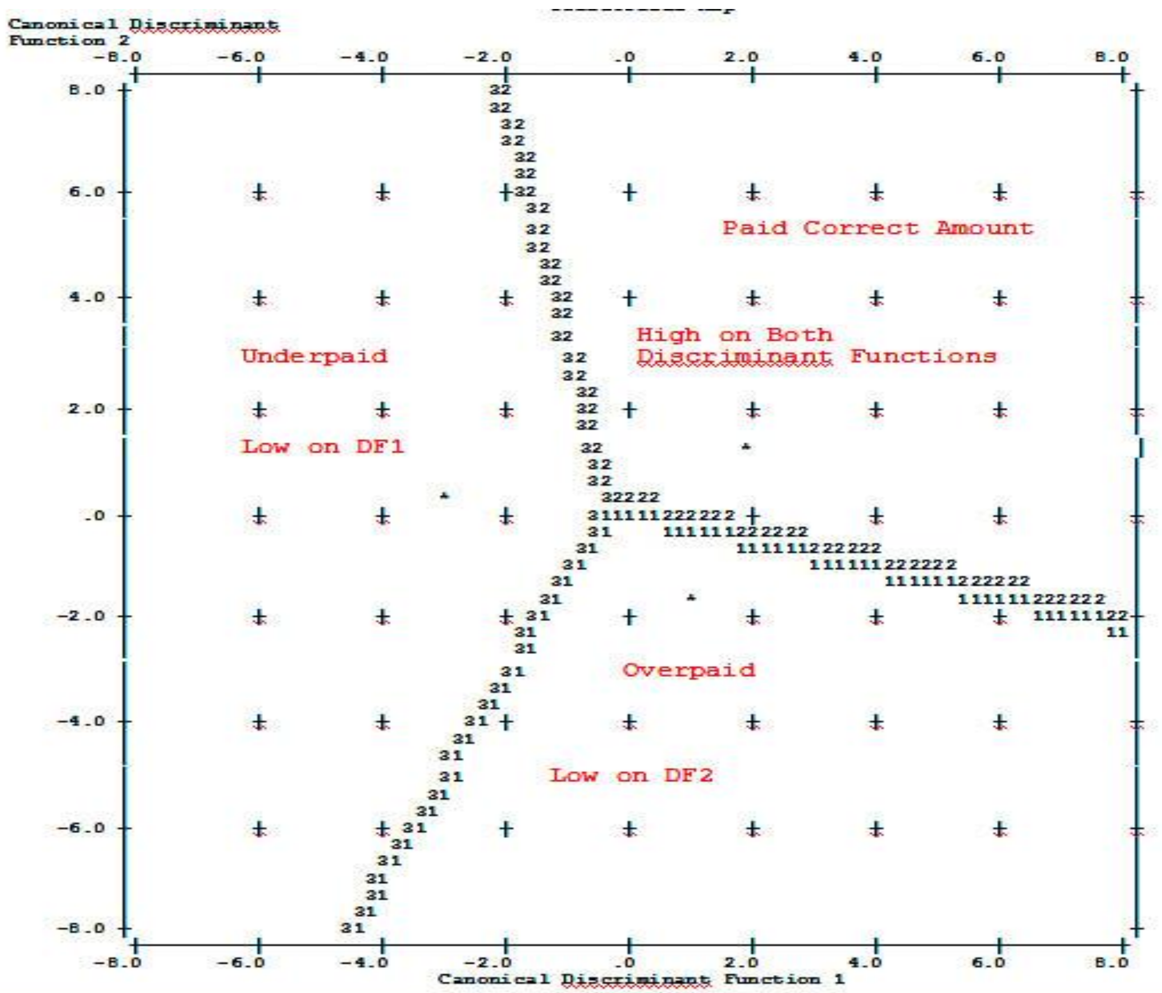
Look back at the Step 0 statistics. Only $X_2$ and $X_3$ were eligible for entry. Note, however, that after $X_3$ was entered, the *p* to enter dropped for all remaining predictors. Why? $X_3$ must suppress irrelevant variance in the other predictors (and vice versa). After $X_1$ is added to $X_3$, *p* to enter for $X_4$ rises, indicating redundancy of $X_4$ with $X_1$.

### Interpretation of the Output from the Example Program

If you look at the **standardized coefficients and loadings** you will see that high scores on $DF_1$ result from high $X_3$ and low $X_1$. If you look back at the **group means** you will see that those who underpaid are characterized by having low $X_3$ and high $X_1$, and thus low $DF_1$. This suggests that $DF_1$ is good for discriminating the cheaters (those who underpaid) from the others. The **centroids** confirm this.

If you look at the standardized coefficients and loadings for $DF_2$ you will see that high $DF_2$ scores come from having high $X_2$ and low $X_1$. From the group means you see that those who overpaid will have low $DF_2$ (since they have a low $X_2$ and a high $X_1$). $DF_2$ seems to be good for separating those who overpaid from the others, as confirmed by the centroids for $DF_2$.

In the **territorial map** the underpayers are on the left, having a low $DF_1$ (high $X_1$ and low $X_3$). The overpayers are on the lower right, having a high $DF_1$ and a low $DF_2$ (low $X_2$, high $X_3$, high $X_1$). Those who paid the correct amount are in the upper right, having a high $DF_1$ and a high $DF_2$ (low $X_1$, high $X_2$, high $X_3$).



# Experiment 7: Confirmatory Factor Analysis

**Confirmatory factor analysis (CFA)** is a multivariate statistical procedure that is used to test how well the measured variables represent the number of constructs. Confirmatory factor analysis (CFA) and exploratory factor analysis (EFA) are similar techniques, but in exploratory factor analysis (EFA), data is simply explored and provides information about the numbers of factors required to represent the data. In exploratory factor analysis, all measured variables are related to every latent variable. But in confirmatory factor analysis (CFA), researchers can specify the number of factors required in the data and which measured variable is related to which latent variable. Confirmatory factor analysis (CFA) is a tool that is used to confirm or reject the measurement theory.

**General Purpose – Procedure**

1. **Defining individual construct:** First, we have to define the individual constructs. The first step involves the procedure that defines constructs theoretically. This involves a pretest to

evaluate the construct items, and a confirmatory test of the measurement model that is conducted using confirmatory factor analysis (CFA), etc.

2. **Developing the overall measurement model theory:** In confirmatory factor analysis (CFA), we should consider the concept of unidimensionality between construct error variance and within construct error variance. At least four constructs and three items per constructs should be present in the research.
3. **Designing a study to produce the empirical results:** The measurement model must be specified. Most commonly, the value of one loading estimate should be one per construct. Two methods are available for identification; the first is rank condition, and the second is order condition.
4. **Assessing the measurement model validity:** Assessing the measurement model validity occurs when the theoretical measurement model is compared with the reality model to see how well the data fits. To check the measurement model validity, the number of the indicator helps us. For example, the factor loading latent variable should be greater than 0.7. Chi-square test and other goodness of fit statistics like RMR, GFI, NFI, RMSEA, SIC, BIC, etc., are some key indicators that help in measuring the model validity.

**Assumptions**

The assumptions of a CFA include multivariate normality, a sufficient sample size (*n* >200), the correct a priori model specification, and data must come from a random sample.

# Experiment 8: Conjoint Analysis

A graphical user interface is not yet available for the Conjoint procedure. To obtain a conjoint analysis, you must enter command syntax for a CONJOINT command into a syntax window and then run it. v For an example of command syntax for a CONJOINT command in the context of a complete conjoint analysis--including generating and displaying an orthogonal design--see . v For complete command syntax information about the CONJOINT command, see the Command Syntax Reference. To Run a Command from a Syntax Window From the menus choose: File > New > Syntax... This opens a syntax window. 1. Enter the command syntax for the CONJOINT command. 2. Highlight the command in the syntax window, and click the Run button (the right-pointing triangle) on the Syntax Editor toolbar. See the Core System User’s Guide for more information about running commands in syntax windows. Requirements The Conjoint procedure requires two files—a data file and a plan file—and the specification of how data were recorded (for example, each data point is a preference score from 1 to 100). The plan file consists of the set of product profiles to be rated by the subjects and should be generated using the Generate Orthogonal Design procedure. The data file contains the preference scores or rankings of those profiles collected from the subjects. The plan and data files are specified with the PLAN and DATA subcommands, respectively. The method of data recording is specified with the SEQUENCE, RANK, or SCORE subcommands. The following command syntax shows a minimal specification: CONJOINT PLAN=’CPLAN.SAV’ /DATA=’RUGRANKS.SAV’ /SEQUENCE=PREF1 TO PREF22. Specifying the Plan File and the Data File The CONJOINT command provides a number of options for specifying the plan file and the data file. v You can explicitly specify the filenames for the two files. For example: CONJOINT PLAN=’CPLAN.SAV’ /DATA=’RUGRANKS.SAV’ v If only a plan file or data file is specified, the CONJOINT command reads the specified file and uses the active dataset as the other. For example, if you specify a data file but omit a plan file (you cannot omit both), the active dataset is used as the plan, as shown in the following example: CONJOINT DATA=’RUGRANKS.SAV’ v You can use the asterisk (*) in place of a filename to indicate the active dataset, as shown in the following example: CONJOINT PLAN=’CPLAN.SAV’ /DATA=* The active dataset is used as the preference data. Note that you cannot use the asterisk (*) for both the plan file and the data file. © Copyright IBM Corporation 1989, 2013 9 Specifying How Data Were Recorded You must specify the way in which preference data were recorded. Data can be recorded in one of three ways: sequentially, as rankings, or as preference scores. These three methods are indicated by the SEQUENCE, RANK, and SCORE subcommands. You must specify one, and only one, of these subcommands as part of a CONJOINT command. SEQUENCE Subcommand The SEQUENCE subcommand indicates that data were recorded sequentially so that each data point in the data file is a profile number, starting with the most preferred profile and ending with the least preferred profile. This is how data are recorded if the subject is asked to order the profiles from the most to the least preferred. The researcher records which profile number was first, which profile number was second, and so on. CONJOINT PLAN=* /DATA=’RUGRANKS.SAV’ /SEQUENCE=PREF1 TO PREF22. v The variable PREF1 contains the profile number for the most preferred profile out of 22 profiles in the orthogonal plan. The variable PREF22 contains the profile number for the least preferred profile in the plan. RANK Subcommand The RANK subcommand indicates that each data point is a ranking, starting with the ranking of profile 1, then the ranking of profile 2, and so on. This is how the data are recorded if the subject is asked to assign a rank to each profile, ranging from 1 to n, where n is the number of profiles. A lower rank implies greater preference. CONJOINT PLAN=* /DATA=’RUGRANKS.SAV’ /RANK=RANK1 TO RANK22. v The variable RANK1 contains the ranking of profile 1, out of a total of 22 profiles in the orthogonal plan. The variable RANK22 contains the ranking of profile 22. SCORE Subcommand The SCORE subcommand indicates that each data point is a preference score assigned to the profiles, starting with the score of profile 1, then the score of profile 2, and so on. This type of data might be generated, for example, by asking subjects to assign a number from 1 to 100 to show how much they liked the profile. A higher score implies greater preference.

# Experiment 9: Time Series Analysis

The Apply Time Series Models procedure loads existing time series models from an external file and applies them to the active dataset. You can use this procedure to obtain forecasts for series for which new or revised data are available, without rebuilding your models. Models are generated using the Time Series Modeler procedure. Example. You are an inventory manager with a major retailer, and responsible for each of 5,000 products. You've used the Expert Modeler to create models that forecast sales for each product three months into the future. Your data warehouse is refreshed each month with actual sales data which you'd like to use to produce monthly updated forecasts. The Apply Time Series Models procedure allows you to accomplish this using the original models, and simply reestimating model parameters to account for the new data. Statistics. Goodness-of-fit measures: stationary R-square, R-square (R 2 ), root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), maximum absolute error (MaxAE), maximum absolute percentage error (MaxAPE), normalized Bayesian information criterion (BIC). Residuals: autocorrelation function, partial autocorrelation function, Ljung-Box Q. Plots. Summary plots across all models: histograms of stationary R-square, R-square (R 2 ), root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), maximum absolute error (MaxAE), maximum absolute percentage error (MaxAPE), normalized Bayesian information criterion (BIC); box plots of residual autocorrelations and partial autocorrelations. Results for individual models: forecast values, fit values, observed values, upper and lower confidence limits, residual autocorrelations and partial autocorrelations. Apply Time Series Models Data Considerations Data. Variables (dependent and independent) to which models will be applied should be numeric. Assumptions. Models are applied to variables in the active dataset with the same names as the variables specified in the model. All such variables are treated as time series, meaning that each case represents a time point, with successive cases separated by a constant time interval. v Forecasts. For producing forecasts using models with independent (predictor) variables, the active dataset should contain values of these variables for all cases in the forecast period. If model parameters are reestimated, then independent variables should not contain any missing values in the estimation period. Defining Dates The Apply Time Series Models procedure requires that the periodicity, if any, of the active dataset matches the periodicity of the models to be applied. If you're simply forecasting using the same dataset (perhaps with new or revised data) as that used to the build the model, then this condition will be satisfied. If no periodicity exists for the active dataset, you will be given the opportunity to navigate to the Define Dates dialog box to create one. If, however, the models were created without specifying a periodicity, then the active dataset should also be without one.

**To Apply Models**

1. From the menus choose: **Analyze > Forecasting > Apply Models...**

2. Enter the file specification for a model file or click Browse and select a model file (model files are created with the Time Series Modeler procedure).

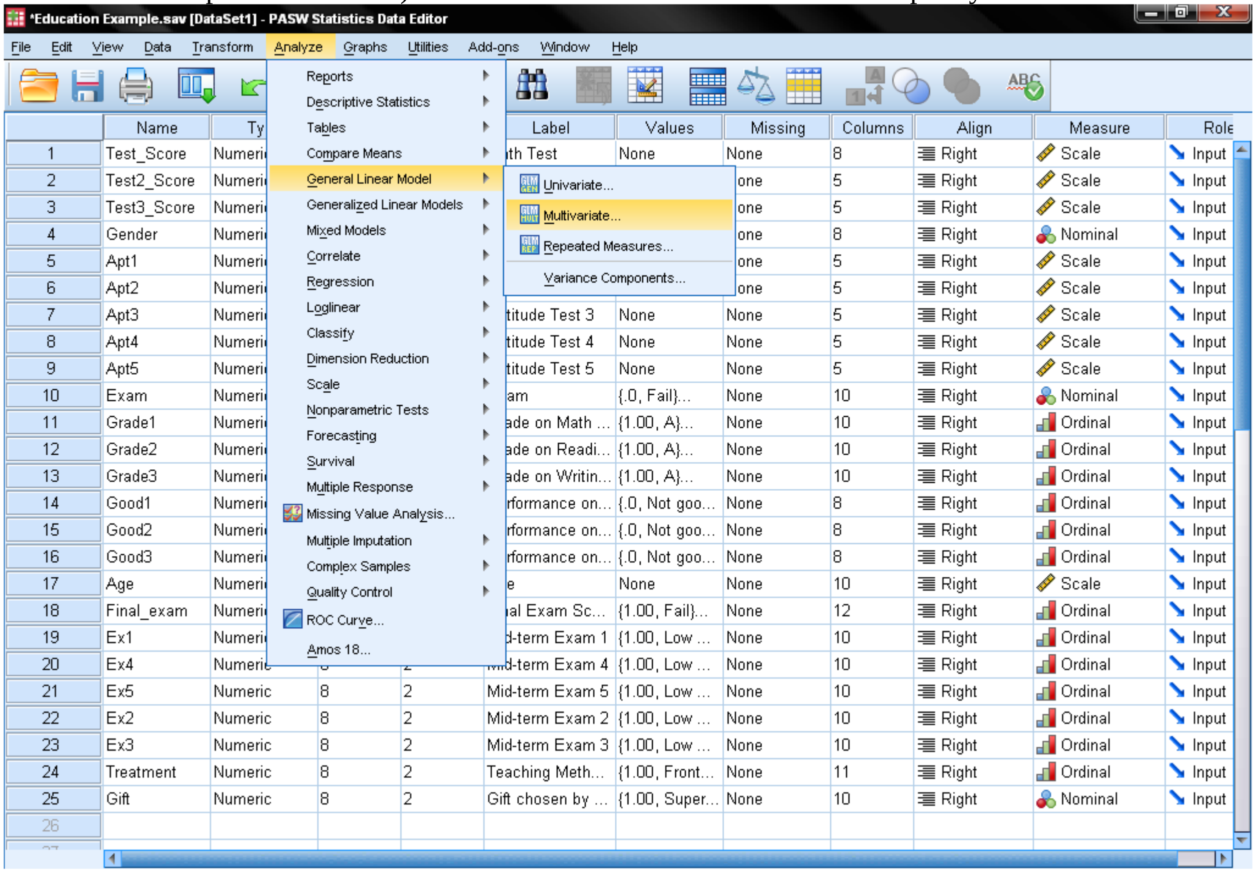
Optionally, you can: v Reestimate model parameters using the data in the active dataset. Forecasts are created using the reestimated parameters. v Save predictions, confidence intervals, and noise residuals. v Save reestimated models in XML format. Model Parameters and Goodness of Fit Measures Load from model file. Forecasts are produced using the model parameters from the model file without reestimating those parameters. Goodness of fit measures displayed in output and used to filter models (best- or worst-fitting) are taken from the model file and reflect the data used when each model was developed (or last updated). With this option, forecasts do not take into account historical data--for either dependent or independent variables--in the active dataset. You must choose Reestimate from data if you want historical data to impact the forecasts. In addition, forecasts do not take into account values of the dependent series in the forecast period--but they do take into account values of independent variables in the forecast period. If you have more current values of the dependent series and want them to be included in the forecasts, you need to reestimate, adjusting the estimation period to include these values. Reestimate from data. Model parameters are reestimated using the data in the active dataset. Reestimation of model parameters has no effect on model structure. For example, an ARIMA(1,0,1) model will remain so, but the autoregressive and moving-average parameters will be reestimated. Reestimation does not result in the detection of new outliers. Outliers, if any, are always taken from the model file. v Estimation Period. The estimation period defines the set of cases used to reestimate the model parameters. By default, the estimation period includes all cases in the active dataset. To set the estimation period, select Based on time or case range in the Select Cases dialog box. Depending on available data, the estimation period used by the procedure may vary by model and thus differ from the displayed value. For a given model, the true estimation period is the period left after eliminating any contiguous missing values, from the model's dependent variable, occurring at the beginning or end of the specified estimation period. Forecast Period The forecast period for each model always begins with the first case after the end of the estimation period and goes through either the last case in the active dataset or a user-specified date. If parameters are not reestimated (this is the default), then the estimation period for each model is the set of cases used when the model was developed (or last updated). v First case after end of estimation period through last case in active dataset. Select this option when the end of the estimation period is prior to the last case in the active dataset, and you want forecasts through the last case. v First case after end of estimation period through a specified date. Select this option to explicitly specify the end of the forecast period. Enter values for all of the cells in the Date grid. If no date specification has been defined for the active dataset, the Date grid shows the single column Observation. To specify the end of the forecast period, enter the row number (as displayed in the Data Editor) of the relevant case. The Cycle column (if present) in the Date grid refers to the value of the CYCLE_ variable in the active dataset. Output Available output includes results for individual models as well as results across all models. Results for individual models can be limited to a set of best- or poorest-fitting models based on user-specified criteria.

Statistics and Forecast Tables The Statistics tab provides options for displaying tables of model fit statistics, model parameters, autocorrelation functions, and forecasts. Unless model parameters are reestimated (Reestimate from data on the Models tab), displayed values of fit measures, Ljung-Box values, and model parameters are those from the model file and reflect the data used when each model was developed (or last updated). Outlier information is always taken from the model file. Display fit measures, Ljung-Box statistic, and number of outliers by model. Select (check) this option to display a table containing selected fit measures, Ljung-Box value, and the number of outliers for each model. Fit Measures. You can select one or more of the following for inclusion in the table containing fit measures for each model: v Stationary R-square v R-square v Root mean square error v Mean absolute percentage error v Mean absolute error v Maximum absolute percentage error v Maximum absolute error v Normalized BIC See the topic Chapter 6, "Goodness-of-Fit Measures," on page 25 for more information. Statistics for Comparing Models. This group of options controls the display of tables containing statistics across all models. Each option generates a separate table. You can select one or more of the following options: v Goodness of fit. Table of summary statistics and percentiles for stationary R-square, R-square, root mean square error, mean absolute percentage error, mean absolute error, maximum absolute percentage error, maximum absolute error, and normalized Bayesian Information Criterion. v Residual autocorrelation function (ACF). Table of summary statistics and percentiles for autocorrelations of the residuals across all estimated models. This table is only available if model parameters are reestimated (Reestimate from data on the Models tab). v Residual partial autocorrelation function (PACF). Table of summary statistics and percentiles for partial autocorrelations of the residuals across all estimated models. This table is only available if model parameters are reestimated (Reestimate from data on the Models tab). Statistics for Individual Models. This group of options controls display of tables containing detailed information for each model. Each option generates a separate table. You can select one or more of the following options: v Parameter estimates. Displays a table of parameter estimates for each model. Separate tables are displayed for exponential smoothing and ARIMA models. If outliers exist, parameter estimates for them are also displayed in a separate table. v Residual autocorrelation function (ACF). Displays a table of residual autocorrelations by lag for each estimated model. The table includes the confidence intervals for the autocorrelations. This table is only available if model parameters are reestimated (Reestimate from data on the Models tab). v Residual partial autocorrelation function (PACF). Displays a table of residual partial autocorrelations by lag for each estimated model. The table includes the confidence intervals for the partial autocorrelations. This table is only available if model parameters are reestimated

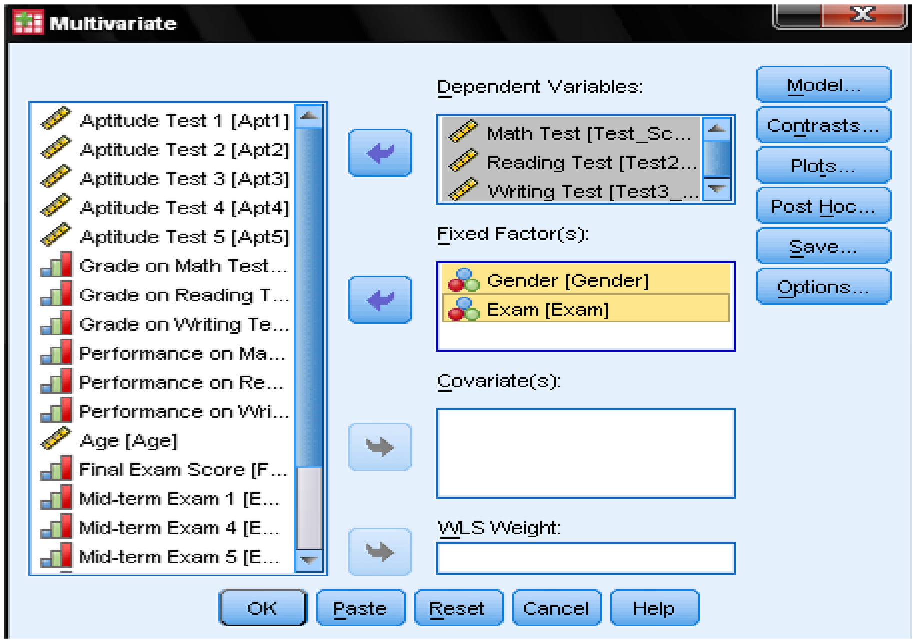
(Reestimate from data on the Models tab). Display forecasts. Displays a table of model forecasts and confidence intervals for each model.

# Experiment 10: MANOVA

The research question indicates that this analysis has multiple independent variables (exam and gender) and multiple dependent variables (math, reading, and writing test scores). We will skip the check for multivariate normality of the dependent variables; the sample we are going to look at has some violations of the assumption set forth by the MANOVA. The MANOVA can be found in SPSS in *Analyze/General Linear Model/Multivariate*□, which opens the dialog for multivariate GLM procedure (that is GLM with more than one dependent variable). The multivariate GLM model is used to specify the MANOVAs.



| | Name | Ty | | | Label | Values | Missing | Columns | Align | Measure | Role |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Test_Score | Numeri | | | th Test | None | None | 8 | Right | Scale | Input |
| 2 | Test2_Score | Numeri | | | | | one | 5 | Right | Scale | Input |
| 3 | Test3_Score | Numeri | | | | | one | 5 | Right | Scale | Input |
| 4 | Gender | Numeri | | | | | one | 8 | Right | Nominal | Input |
| 5 | Apt1 | Numeri | | | | | one | 5 | Right | Scale | Input |
| 6 | Apt2 | Numeri | | | | | one | 5 | Right | Scale | Input |
| 7 | Apt3 | Numeri | | | titude Test 3 | None | None | 5 | Right | Scale | Input |
| 8 | Apt4 | Numeri | | | titude Test 4 | None | None | 5 | Right | Scale | Input |
| 9 | Apt5 | Numeri | | | titude Test 5 | None | None | 5 | Right | Scale | Input |
| 10 | Exam | Numeri | | | am | {.0, Fail}... | None | 10 | Right | Nominal | Input |
| 11 | Grade1 | Numeri | | | ade on Math ... | {1.00, A}... | None | 10 | Right | Ordinal | Input |
| 12 | Grade2 | Numeri | | | ade on Readi... | {1.00, A}... | None | 10 | Right | Ordinal | Input |
| 13 | Grade3 | Numeri | | | ade on Writin... | {1.00, A}... | None | 10 | Right | Ordinal | Input |
| 14 | Good1 | Numeri | | | rformance on... | {.0, Not goo... | None | 8 | Right | Ordinal | Input |
| 15 | Good2 | Numeri | | | rformance on... | {.0, Not goo... | None | 8 | Right | Ordinal | Input |
| 16 | Good3 | Numeri | | | rformance on... | {.0, Not goo... | None | 8 | Right | Ordinal | Input |
| 17 | Age | Numeri | | | e | None | None | 10 | Right | Scale | Input |
| 18 | Final_exam | Numeri | | | al Exam Sc... | {1.00, Fail}... | None | 12 | Right | Ordinal | Input |
| 19 | Ex1 | Numeri | | | d-term Exam 1 | {1.00, Low ... | None | 10 | Right | Ordinal | Input |
| 20 | Ex4 | Numeric | 8 | 2 | id-term Exam 4 | {1.00, Low ... | None | 10 | Right | Ordinal | Input |
| 21 | Ex5 | Numeric | 8 | 2 | Mid-term Exam 5 | {1.00, Low ... | None | 10 | Right | Ordinal | Input |
| 22 | Ex2 | Numeric | 8 | 2 | Mid-term Exam 2 | {1.00, Low ... | None | 10 | Right | Ordinal | Input |
| 23 | Ex3 | Numeric | 8 | 2 | Mid-term Exam 3 | {1.00, Low ... | None | 10 | Right | Ordinal | Input |
| 24 | Treatment | Numeric | 8 | 2 | Teaching Meth... | {1.00, Front... | None | 11 | Right | Ordinal | Input |
| 25 | Gift | Numeric | 8 | 2 | Gift chosen by ... | {1.00, Super... | None | 10 | Right | Nominal | Input |
| 26 | | | | | | | | | | | |



Dependent Variables:

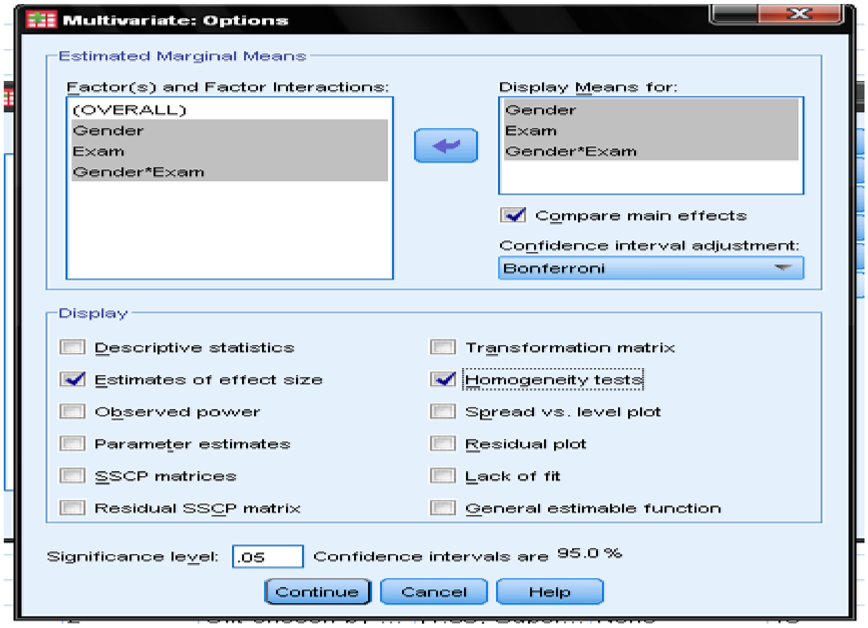
- Math Test [Test_Sc...
- Reading Test [Test2...
- Writing Test [Test3_...

Fixed Factor(s):

- Gender [Gender]
- Exam [Exam]

Covariate(s):

WLS Weight:



Estimated Marginal Means

Factor(s) and Factor Interactions:

- (OVERALL)
- Gender
- Exam
- Gender*Exam

Display Means for:

- Gender
- Exam
- Gender*Exam

☑ Compare main effects

Confidence interval adjustment: Bonferroni

Display

| | |
|---|---|
| □ Descriptive statistics | □ Transformation matrix |
| ☑ Estimates of effect size | ☑ Homogeneity tests |
| □ Observed power | □ Spread vs. level plot |
| □ Parameter estimates | □ Residual plot |
| □ SSCP matrices | □ Lack of fit |
| □ Residual SSCP matrix | □ General estimable function |

Significance level: .05 Confidence intervals are 95.0 %

# Additional Experiment:

## Experiment 11: Decision Tree Analysis

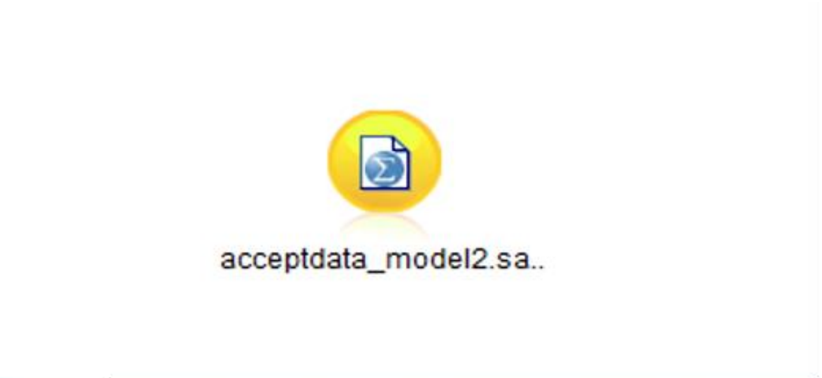
SPSS Modeler is statistical analysis software used for data analysis, data mining and forecasting. Statistical analysis allows us to use a sample of data to make predictions about a larger population. Creating predictive models utilizing the information currently at your fingertips to predict what decisions will impact your future success. Predictive analytics is hugely important as it allows you to see into the future and make quality decisions based on long term planning.

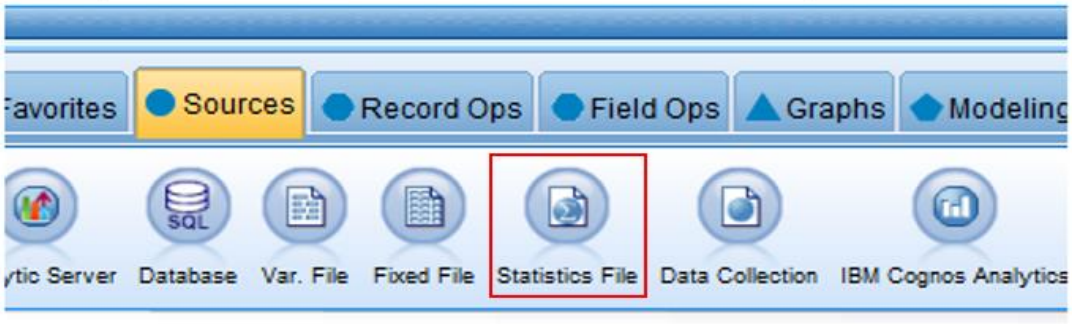
Decision tree analyses are popular models because they indicate which predictors are most strongly related to the target. The purpose of decision trees is to model a series of events and look at how it affects an outcome. This type of model calculates a set of conditional probabilities based on different scenarios.

This blog will detail how to create a simple predictive model using a CHAID analysis and how to interpret the decision tree results. In this example I will be predicting student enrollment, which has two categories Yes, meaning those students who did enroll in the university and No, those students who did not enroll.

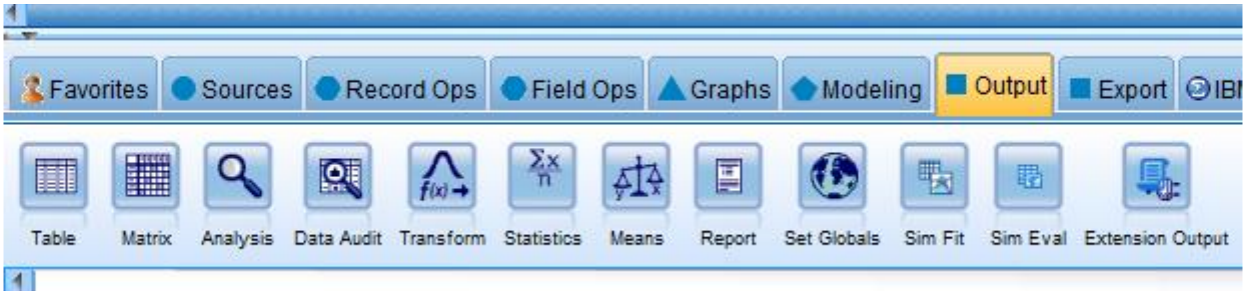
**Creating the Model:**

Starting from the sources tab I'm going to drag in a statistics file node and import the .sav file from my local machine.

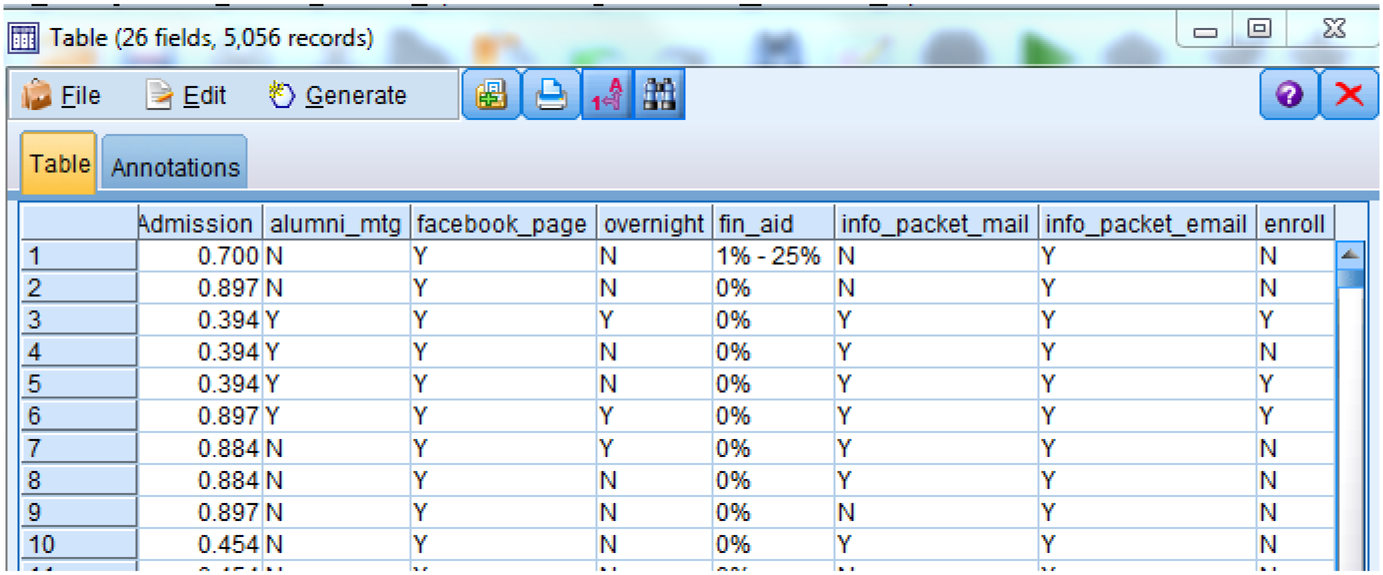




To view the data drag in a table node and attach it to the statistics node already on the canvas.

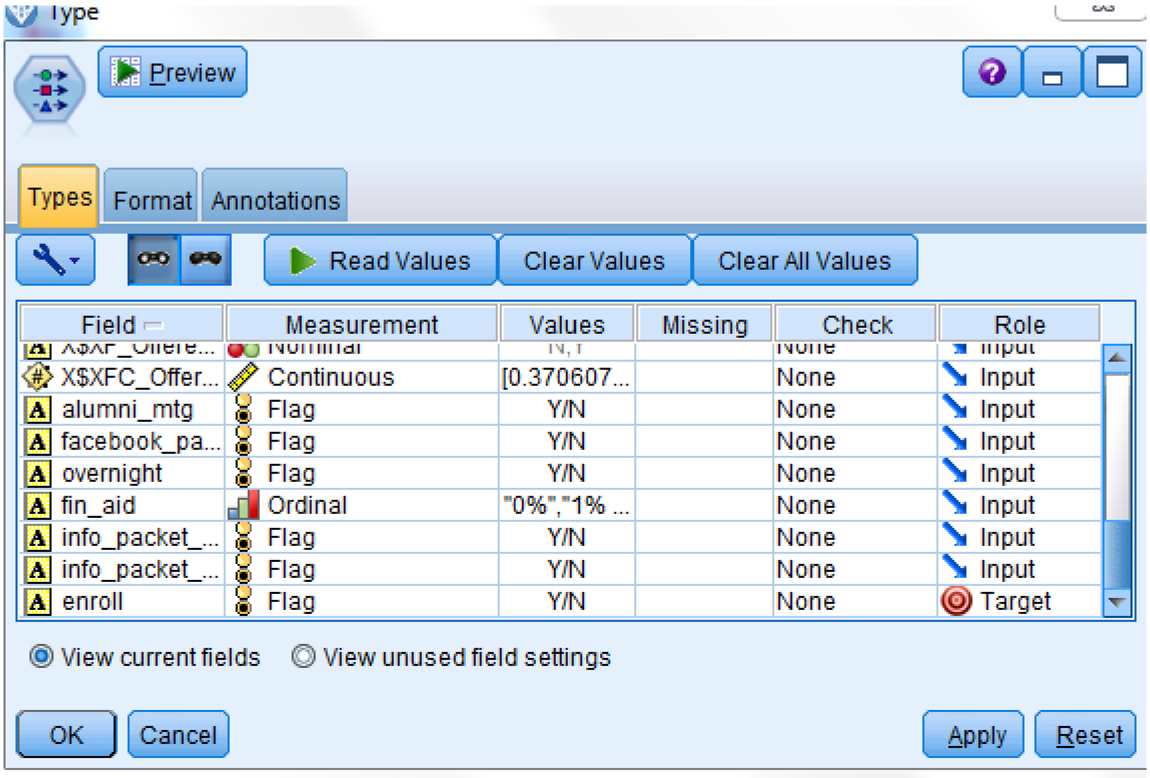


Click run and double click to view the table output.



| | Admission | alumni_mtg | facebook_page | overnight | fin_aid | info_packet_mail | info_packet_email | enroll |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.700 | N | Y | N | 1% - 25% | N | Y | N |
| 2 | 0.897 | N | Y | N | 0% | N | Y | N |
| 3 | 0.394 | Y | Y | Y | 0% | Y | Y | Y |
| 4 | 0.394 | Y | Y | N | 0% | Y | Y | N |
| 5 | 0.394 | Y | Y | N | 0% | Y | Y | Y |
| 6 | 0.897 | Y | Y | Y | 0% | Y | Y | Y |
| 7 | 0.884 | N | Y | Y | 0% | Y | Y | N |
| 8 | 0.884 | N | Y | N | 0% | Y | Y | N |
| 9 | 0.897 | N | Y | N | 0% | N | Y | N |
| 10 | 0.454 | N | Y | N | 0% | Y | Y | N |

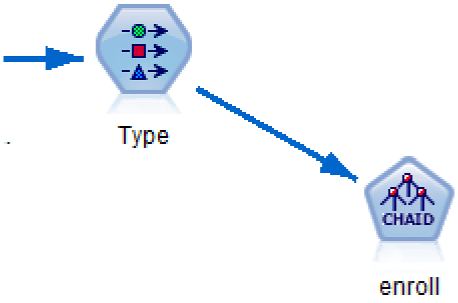
The next step in the process is to read in the data using a type node. The type node specifies metadata and data properties for each field: the measurement level, data values, the role and missing value definitions. From the screenshot you can see that the field enroll is our target.



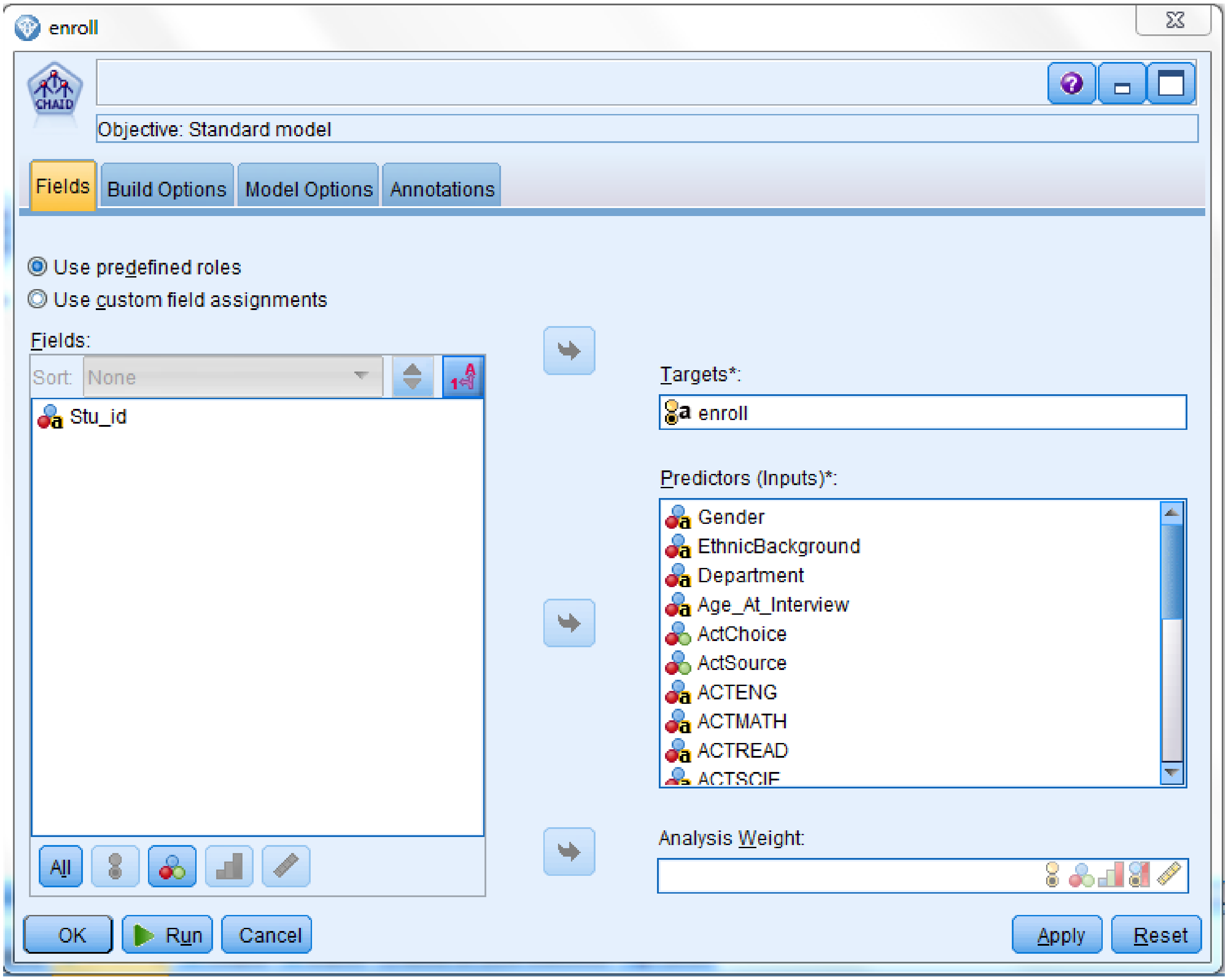
| Field | Measurement | Values | Missing | Check | Role |
|---|---|---|---|---|---|
| X$XF_Offere... | Nominal | N,Y | | None | Input |
| X$XFC_Offer... | Continuous | [0.370607... | | None | Input |
| alumni_mtg | Flag | Y/N | | None | Input |
| facebook_pa... | Flag | Y/N | | None | Input |
| overnight | Flag | Y/N | | None | Input |
| fin_aid | Ordinal | "0%","1% ... | | None | Input |
| info_packet_... | Flag | Y/N | | None | Input |
| info_packet_... | Flag | Y/N | | None | Input |
| enroll | Flag | Y/N | | None | Target |

● View current fields ○ View unused field settings

Next drag a CHAID node and attach it to the existing type node. CHAID stands for chi square automatic interaction detection and is one of the more popular decision tree models. And really what's going on behind the scenes is that the model is running the chi square test many times. This will make more sense in just a little bit, but essentially, the model is picking the predictors with the strongest relationship with the outcome field and that is determined by the field that has the highest chi square statistic.



Edit the CHAID node and make sure the target is set correctly. You can also remove any inputs or predictors that you don't want included in the model. When you're ready click the Run icon in the lower left hand corner to create the model.



enroll

Objective: Standard model

● Use predefined roles

○ Use custom field assignments

Fields:

Sort: None

Stu_id

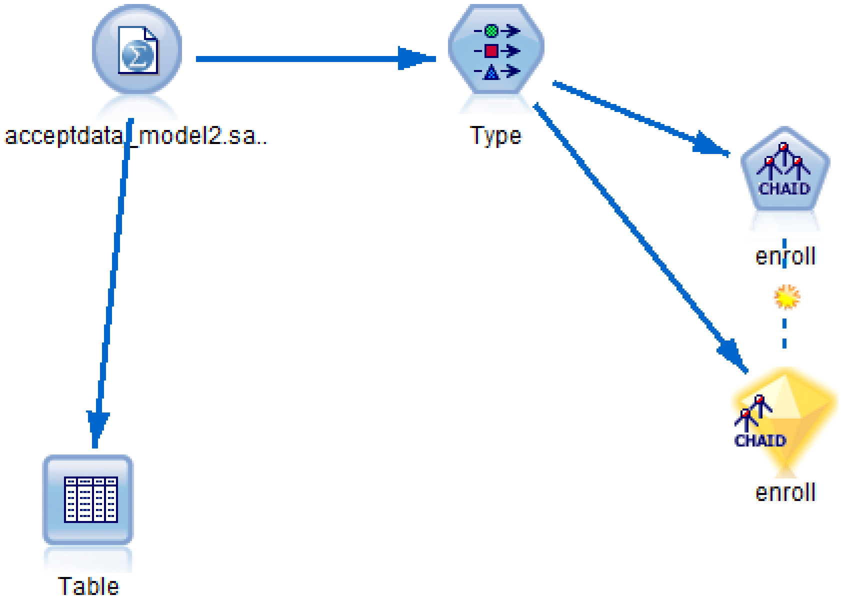
Targets*: enroll

Predictors (Inputs)*:

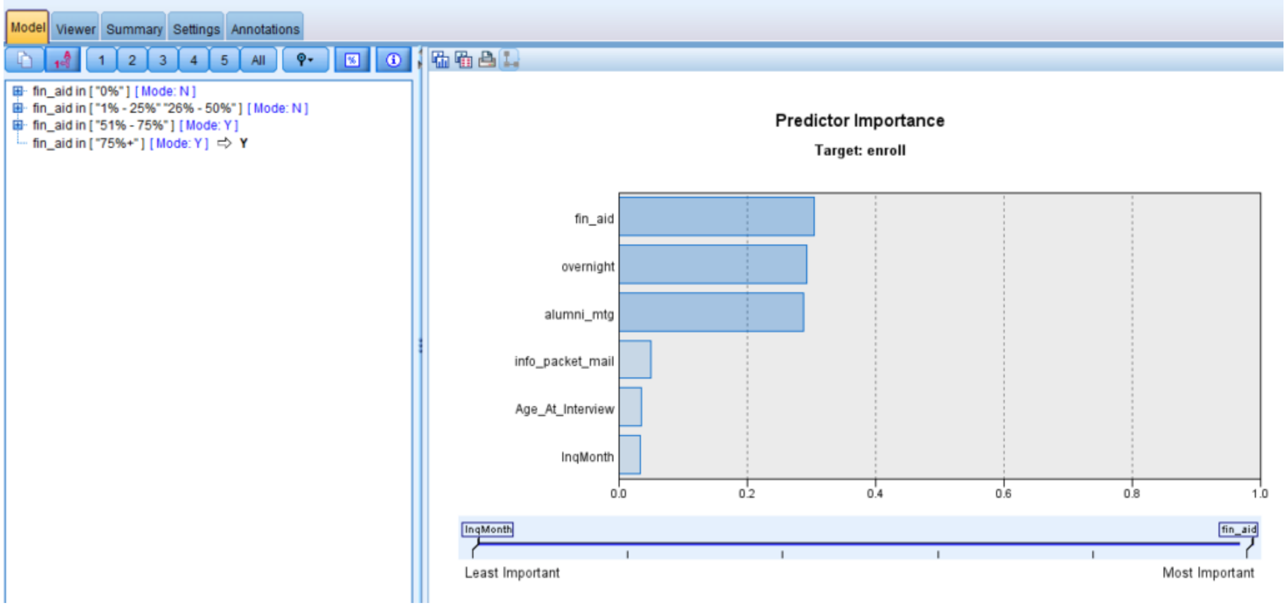
- Gender
- EthnicBackground
- Department
- Age_At_Interview
- ActChoice
- ActSource
- ACTENG
- ACTMATH
- ACTREAD
- ACTSCIE

Analysis Weight:

If all goes well you will get the golden nugget. Double click the nugget to see the results.



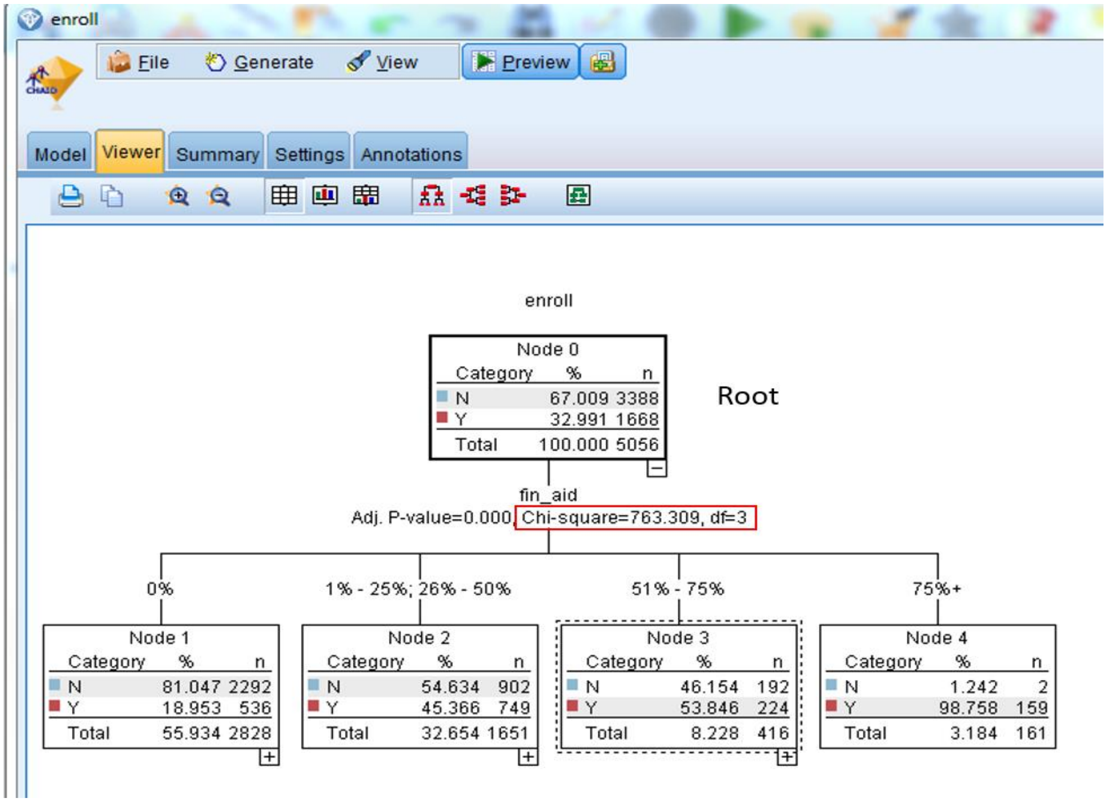
What we have here are the top predictors of enrollment. The top three predictors are financial aid, overnight visit and alumni meeting. In total there were six predictors that the model deemed important. To view the decision tree click on the Viewer tab.



fin_aid in [ "0%" ] [ Mode: N ]
fin_aid in [ "1% - 25%" "26% - 50%" ] [ Mode: N ]
fin_aid in [ "51% - 75%" ] [ Mode: Y ]
fin_aid in [ "75%+" ] [ Mode: Y ] ⇨ Y

**Interpreting the Results:**

The decision tree starts with the root node, which simply shows the distribution of the outcome field, which as we know is enrollment.  The data is then split based on statistical significance by the predictor with the strongest relationship with the target field, financial aid in this case. And you can see that there are five “buckets” that financial aid has been split into (0%, 1%-25%, 26%-5-%, 51%-75% and 75% +). Looking at those students who were offered a 51%-75% financial aid package, the model was able to predict that those students would enroll roughly 54% of the time. This prediction applied to 416 students and the model was accurate 224 times.



enroll

Node 0

| Category | % | n |
|---|---|---|
| 🟦 N | 67.009 | 3388 |
| 🟥 Y | 32.991 | 1668 |
| Total | 100.000 | 5056 |

Root

fin_aid
Adj. P-value=0.000, Chi-square=763.309, df=3

0%

Node 1

| Category | % | n |
|---|---|---|
| 🟦 N | 81.047 | 2292 |
| 🟥 Y | 18.953 | 536 |
| Total | 55.934 | 2828 |

1% - 25%; 26% - 50%

Node 2

| Category | % | n |
|---|---|---|
| 🟦 N | 54.634 | 902 |
| 🟥 Y | 45.366 | 749 |
| Total | 32.654 | 1651 |

51% - 75%

Node 3

| Category | % | n |
|---|---|---|
| 🟦 N | 46.154 | 192 |
| 🟥 Y | 53.846 | 224 |
| Total | 8.228 | 416 |

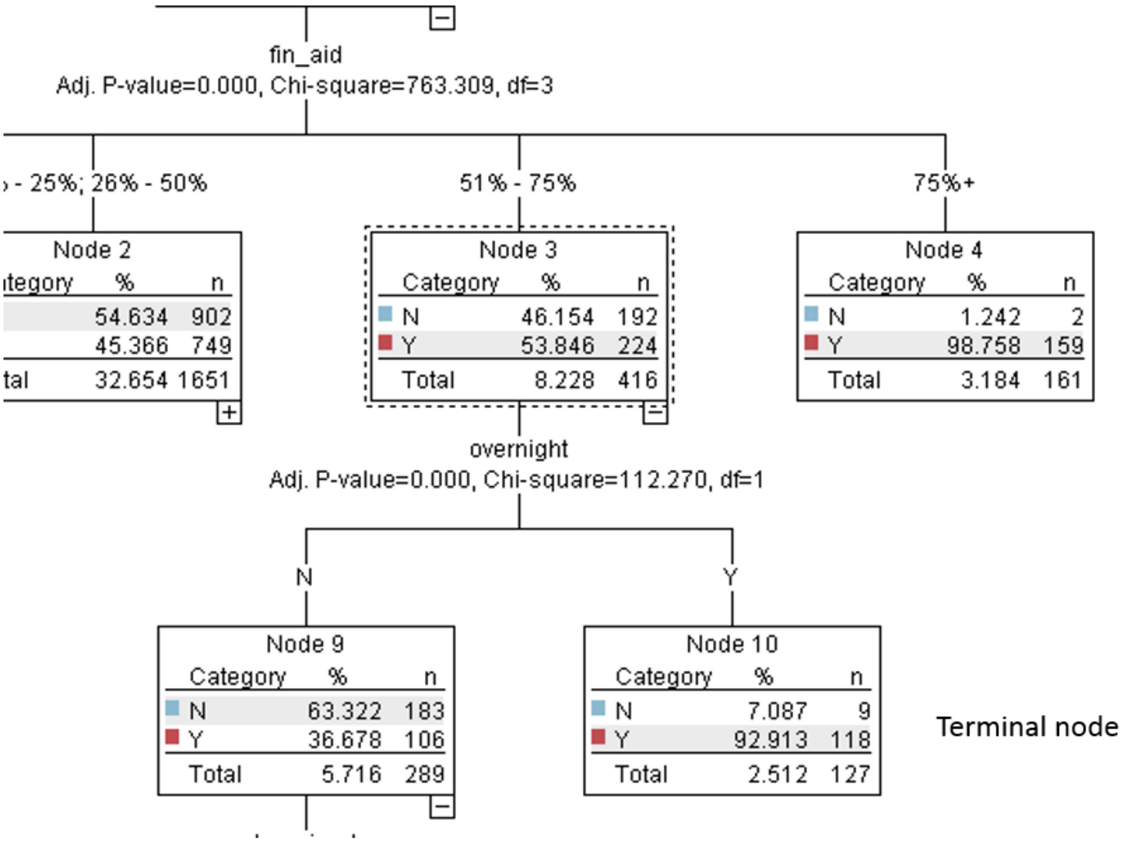
75%+

Node 4

| Category | % | n |
|---|---|---|
| 🟦 N | 1.242 | 2 |
| 🟥 Y | 98.758 | 159 |
| Total | 3.184 | 161 |

As we continue to work our way down the tree, we see that the next most important variable is an overnight visit. If a student was offered a financial aid package of 51%-75% and also took an overnight visit we were able to accurately predict that they would enroll around 93% of the time. Alternatively, if students did not take an overnight visit we predicted that they would not enroll 63% of the time. This rule applied to 289 students and we were accurate about 183 times.

And just like that we continue to work our way down the tree to the next most significant variable until we reach a terminal node, which signifies that the prediction has ended.


fin_aid
Adj. P-value=0.000, Chi-square=763.309, df=3
› - 25%; 26% - 50%
51% - 75%
75%+
overnight
Adj. P-value=0.000, Chi-square=112.270, df=1
N
Y
Terminal node

Node 2

| itegory | % | n |
|---|---|---|
|  | 54.634 | 902 |
|  | 45.366 | 749 |
| tal | 32.654 | 1651 |

Node 3

| Category | % | n |
|---|---|---|
| 🟦 N | 46.154 | 192 |
| 🟥 Y | 53.846 | 224 |
| Total | 8.228 | 416 |

Node 4

| Category | % | n |
|---|---|---|
| 🟦 N | 1.242 | 2 |
| 🟥 Y | 98.758 | 159 |
| Total | 3.184 | 161 |

Node 9

| Category | % | n |
|---|---|---|
| 🟦 N | 63.322 | 183 |
| 🟥 Y | 36.678 | 106 |
| Total | 5.716 | 289 |

Node 10

| Category | % | n |
|---|---|---|
| 🟦 N | 7.087 | 9 |
| 🟥 Y | 92.913 | 118 |
| Total | 2.512 | 127 |

This was a simple decision tree aimed as showing which variables help us to accurately predict student enrollment. Keep in mind that predictive analytics can be applied in a variety of industries including education, retail, healthcare and finance just to name a few.