SPSS – PART 1

SHORTCOURSE HANDOUT

TEXAS TECH UNIVERSITY
Information Technology Division
Technology Support

Texas Tech University **|** Carlee M. DeYoung, M.A.

Tables of Contents

IBM SPSS Statistics 28 – Part 1
ShortCourse Handout

Credit: This document is adapted from SPSS ([Getting Help](), [Tutorial](), [Statistics Coach](), and [Case Studies]()) for SPSS 28 and the SPSS 25 Part-1 ShortCourse Handout created by [Heide Mansouri]().

Introduction

IBM SPSS Statistics 28 is a suite of statistical software available for **Windows**, **Macintosh**, and the **UNIX** platforms. IBM SPSS Statistics 28 stores and outputs data files with **.spv** file extensions.

IBM SPSS Statistics 28 provides statistical analysis and data management system in an accessible **point-and-click interface**.

This course is designed for people who are just starting to use SPSS Statistics and will only cover a few of the many basic statistical procedures available with **SPSS Statistics 28**. This course assumes the user is familiar with **elementary statistics**.

Course Objectives

- Create a data set using **Data Editor**
- Import data from Microsoft **Excel** and **Word**
- Prepare data for analysis
- Run an analysis (e.g. **frequency** or **correlation**)
- Create **new variables**
- Work with SPSS output

Preparing for this lesson

- Go to the following link:
  https://www.depts.ttu.edu/itts/training/shortcourses/handouts.php
- Under "Mathematics and Statistics," click the practice file link under SPSS 1 to download the excel file "grades.xlsx".
- Save this file somewhere where you can quickly access it later (desktop).

Accessing IBM SPSS

A desktop download of IBM SPSS Statistics software is offered free to TTU students and at a cost to TTU faculty/ staff on the [TTU software license site]().

Starting the Program

- Once SPSS is open the following window will appear Click "close"

- From the top menu choose: File → Open→ Data… or select the file icon on the top toolbar.

- Select C:\ Program Files\IBM\SPSS\Statistics\28\Sample\English\demo.sav
- Click "Open"
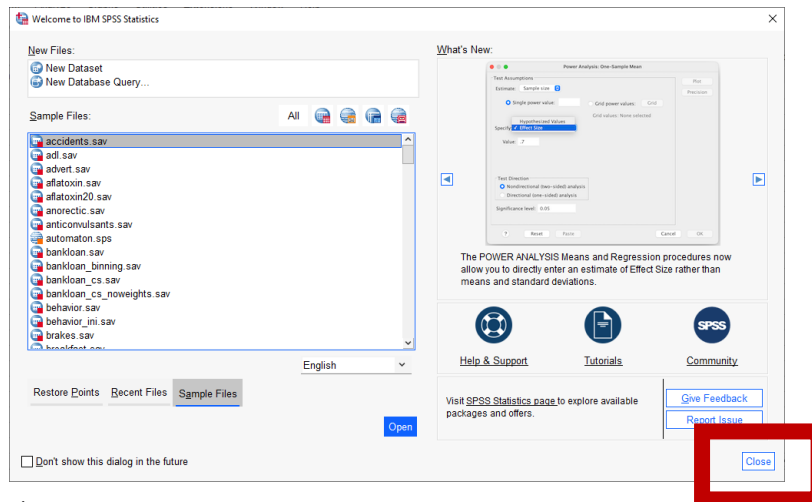
Demo.sav Data File

- This data file is a fictitious survey of several thousand people, containing basic demographic and consumer information.
- In the Data Editor with the Data View tab selected, put the cursor on a variable name located in the column headings. A descriptive variable label will appear in a light-yellow textbox.
- Many times, descriptive datapoints will be numerically coded for analysis purposes. These numerically coded values are displayed by default.
- To display the descriptive, non-numerical labels for datapoints:
  - From "View" in the top menu select "Value Labels" OR use the "Value Labels" button located on the top tool bar
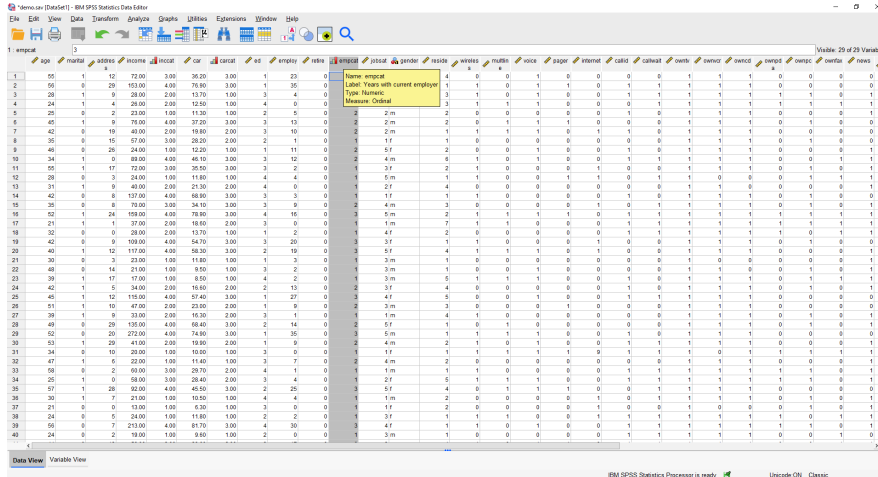
Windows in SPSS Statistics

- **Data Editor**
  - The Data Editor displays the contents of the data file. The information in the Data Editor consists of variables and cases.
    - In *Data View*, columns represent variables, and rows represent cases (observations).
    - In *Variable View*, each row is a variable, and each column is an attribute that is associated with that variable.
- **Viewer** (Output Window)
  - The results from running a statistical procedure are displayed in the Viewer. The output produced can be statistical tables, charts, graphs, or text, depending on the choices you make when you run the procedure.
- **Pivot Table Editor**
  - Most output will be presented in Pivot Tables. These tables can be modified using the Pivot Table Editor. **Note:** to create a Pivot Table, double-click on a table in your output to open the **Pivoting Trays** dialogue box. Also use **Pivot** or **Format** menu.
- **Chart Editor**
  - The Chart Editor provides various methods for manipulating charts.
- **Text Output Editor**
  - Use this editor to edit your output before saving or printing (adding header or editing text). Text output that is not part of the output pivot tables can be modified with eh Text Output Editor.
- **Syntax Editor**

- o You can paste your dialog box choices into a syntax window (using the paste button located on most dialog boxes), where your selections appear in the form of command syntax.

Menus in SPSS Statistics for Windows

- Each window in SPSS statistics has its own menu bar.
- The **Data Editor** window, for example, has the following menu associated with its tool bar:
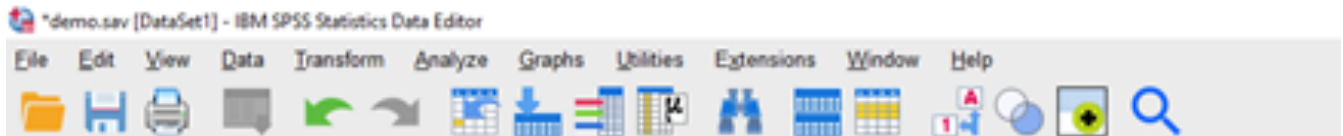
Common Menus between SPSS Statistics Windows

- **File** – Use this menu to create a new SPSS Statistics system file, open an existing file, database, Excel file, or read Text Data.
- **Edit** – Use this menu to cut, copy and paste data values from the Data Editor; modify or copy text from the Viewer, Syntax Editor, etc.
- **View** – Use this menu to turn toolbars and the status bar and gridlines on and off.
- **Data** – Use this menu to make global changes to SPSS Statistics data files, such as transposing variables, cases, or creating subsets of cases for analysis and merging files. **Note:** These changes are only temporary and do not affect the permanent file unless you save the file with the changes.
- **Transform** – Use this menu to make changes to selected variables based on the values of existing ones. **Note:** These changes are only temporary and do not affect the permanent file unless you save the file with the changes.
- **Analyze –** Use this menu for various statistical procedures such as cross tabulation, analysis of variance, correlation, linear regression, and factor analysis.
- **Graphs –** Use this menu to create different types of charts and graphs.
- **Utilities –** Use this menu to display information about the variables in the working data file and control the list of variables from all window types; change the designed Viewer and Syntax Editor.
- **Window –** Use this menu to switch between SPSS Statistics windows or to minimize all open SPSS Statistics windows.
- **Help –** This menu opens a window containing information on how to use the many features of SPSS Statistics.

One-click Descriptive Statistics and more Tools

- **One-click descriptive statistics** -- This tool allows you to run descriptive statistics, quickly, with one click.

- **Merge data files without pre-sorting** – Merge data files by values of key variables without pre-sorting the flies based on key values (Data -> Merge Files -> Add Cases… or Add Variables…)

- **Password protect and encrypt data and output files**.

- **PivotTable editing** – After creating pivot tables you can:
  o Toggle the display of names, values, and labels.
  o Sort table rows.
  o Insert rows and columns.
  o Change the output language.

- **Export Output** -- This unction saves Viewer output in **HTML**, **text**, **Word/ RTF**, **Excel**, **PowerPoint**, and **PDF** formats.
  o To export output:
    ▪ Make the **Output Viewer** the active window by clicking anywhere in the window.
    ▪ From the menu select: **File** -> **Export**.. -> follow the steps.

Using the Help System

From the **Help** menu, choose:

- **Topics –** use Contents and Index tabs to find topics.
  o **Tutorial –** provides access to the introductory tutorial.
  o **Case studies –** provides hands-on examples. The 100 sample data files used in the examples are provided.
  o **Statistics Coach –** helps you find the right statistical or charting procedure for what you want to do.

Using the Data Editor

Data Editor is a spreadsheet-like window which contains the data to be analyzed. The data editor has two views:

- In **Data View**, columns represent variables, and rows represent cases (datapoints, observations).

- In **Variable View**, each row is a variable, and each column is an attribute that is associated with that variable.
  o **Variables** are used to represent the different types of data that you have compiled. For example, the response to each question on a survey is a variable.
  o **Variables** represent different types of data that can change (or vary) between people (e.g. IQ, BMI) and things (e.g. profit, number of cancerous cells etc.)
  o **Variables** may include **numbers, strings, currency, and dates**.

- Data can be entered into the **Data Editor**, when you have small data files or when you want to make minor edits to larger data files.

Rules for Variable Names
- The column names in the data files are **variable names**.
- Each variable name must be **unique**.
- Variable names can be up to **64 characters long**, and the first character must be a **letter**, or the characters **@, #, or $.**
- Variable names **cannot** contain **spaces**.
- The **period**, **underscore**, or the characters **@, #, or $** can be used within a variable name.
- Variable names ending with a **period** should be **avoided**, since the period may be interpreted as a command terminator.
- Variable names ending in **underscores** should be **avoided**.
- Reserved keywords cannot be used as variable names. Reserved keywords are **ALL, AND, BY, EQ, GE, GT, LE, LT, NE, NOT, OR, TO,** and **WITH.**
- Variable names can be defined with any mixture of uppercase and lowercase characters, and case is preserved for display purposes.

Variable Measurement Level
You can specify the level of measurement as **scale** (numeric data on an interval or ratio scale), **ordinal**, or **nominal**. Nominal and ordinal data can be either **string** (alpha numeric) or **numeric**.
- **Nominal** – A variable can be treated as nominal when its values represent categories with no intrinsic ranking (e.g., The department of a company in which an employee works).
    - Examples: region, zip code, and religious affiliation
- **Ordinal** – A variable can be treated as ordinal when its values represent categories with some intrinsic ranking (e.g., levels of service satisfaction from highly dissatisfied to highly satisfied.
    - Examples: socio economic status ("low income","middle income","high income"), education level ("high school","BS","MS","PhD"), satisfaction rating ("extremely dislike", "dislike", "neutral", "like", "extremely like").
- **Scale** -- A variable can be treated as scale (continuous) when its values represent ordered categories with a meaningful metric, so that distance comparisons between values are appropriate.
    - Examples: length, weight, time, age, income

Define Variable type
- Make the Data Editor the active window.
- If Data View is displayed, double-click the variable name at the top of the clo0umn in Data View or click the Variable View tab at the bottom.
- Click the ⬚ button in the **Type** cell for the variable that you want to define.

- Select the data type in the **Variable Type** dialogue box.
- Click **OK**.

Note: For ordinal string variables with values of low, medium, high, it is more reliable to use numeric codes to represent data.
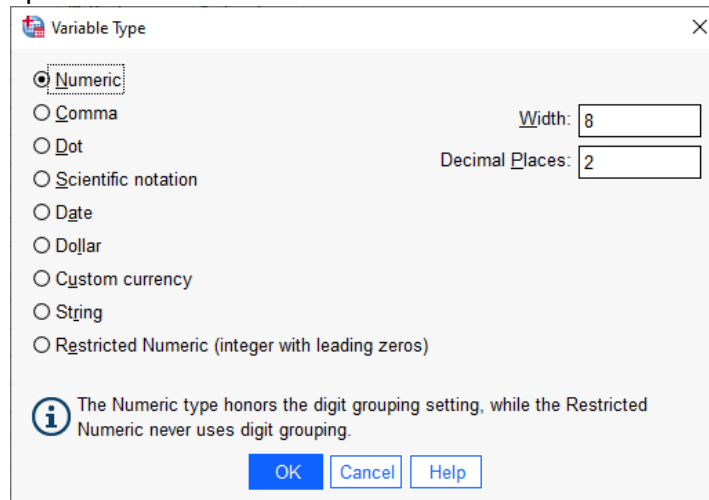
Exercise #1 Create a Dataset

Use the following 10 survey responses from a hypothetical study to create a data set. (adapted from https://ssric.org/trd/modules/spss accessed on 8/29/22).

Participant #: _3_
Gender: M (F)   Age: _31_   Year: Sn
Current GPA: _2.7_   High School GPA: _2.5_
On a scale of 1 to 10, how confident are you that
  you will learn statistics?
  1   2 (3) 4   5   6   7   8   9   10
  Not at All   Very
How many years of computer experience have you
  had? _15_

Participant #: _4_
Gender: (M) F   Age: _27_   Year: J
Current GPA: _3.9_   High School GPA: _3.6_
On a scale of 1 to 10, how confident are you that
  you will learn statistics?
  1   2   3   4   5   6   7   8   9 (10)
  Not at All   Very
How many years of computer experience have you
  had? _4_

Participant #: _5_
Gender: M (F)   Age: _27_   Year: So
Current GPA: _3.11_   High School GPA: _3.2_
On a scale of 1 to 10, how confident are you that
  you will learn statistics?
  1   2   3   4   5   6   7 (8) 9   10
  Not at All   Very
How many years of computer experience have you
  had? _8_

Participant #: _6_
Gender: (M) F   Age: _22_   Year: F
Current GPA: _3.0_   High School GPA: _3.4_
On a scale of 1 to 10, how confident are you that
  you will learn statistics?
  1   2   3   4   5   6 (7) 8   9   10
  Not at All   Very
How many years of computer experience have you
  had? _2_

Participant #: _7_
Gender: (M) F   Age: _21_   Year: Sn
Current GPA: _3.5_   High School GPA: _3.8_
On a scale of 1 to 10, how confident are you that
  you will learn statistics?
  1   2   3   4   5   6   7   8   9 (10)
  Not at All   Very
How many years of computer experience have you
  had? _9_

Participant #: _8_
Gender: M (F)   Age: _28_   Year: J
Current GPA: _2.22_   High School GPA: _2.7_
On a scale of 1 to 10, how confident are you that
  you will learn statistics?
  1   2 (3) 4   5   6   7   8   9   10
  Not at All   Very
How many years of computer experience have you
  had? _5_

Participant #: _9_
Gender: (M) F   Age: _21_   Year: Sn
Current GPA: _2.9_   High School GPA: _3.1_
On a scale of 1 to 10, how confident are you that
  you will learn statistics?
  1   2   3   4 (5) 6   7   8   9   10
  Not at All   Very
How many years of computer experience have you
  had? _1_

Participant #: 10_
Gender: (M) F   Age: _27_   Year: Sn
Current GPA: _3.3_   High School GPA: _3.1_
On a scale of 1 to 10, how confident are you that
  you will learn statistics?
  1   2   3   4   5   6   7 (8) 9   10
  Not at All   Very
How many years of computer experience have you
  had? _7_

To Set Up Columns:

- **File** -> **New** -> **Data**
- In **Variable View,** the first column is titled **Name**. Enter the name of your variables here (no spaces allowed).
- In the name column type: **Case**, **Gender**, **Age**, **GPA**, **HSGPA**, **Confidence** (variable name for survey item about confidence, **CompExp** (variable name for computer experience survey item), pressing enter after each entry. Note, across rows the cells fill with default settings.
- Too add more variables, click on the GPA variable to highlight the row, and then select **Edit-> Insert variable**
- Click on **VAR00001** and type **Year** to give the name variable a name.
- In **"Year"** row, move over to the **Values** column, click on the cell and then on the 3 dots button. A window will appear, press the green plus button on the right of this box **4 times**.

- o Type: **1** in the 1st **Value** box and **Freshman** in the corresponding **Label** box.
- o Type: **2** in the 2nd **Value** box and **Sophomore** in the corresponding **Label** box.
- o Type: **3** in the 3rd **Value** box and **Junior** in the corresponding **Label** box.
- o Type: **4** in the 4th **Value** box and **Senior** in the corresponding **Label** box, and then click **OK**.
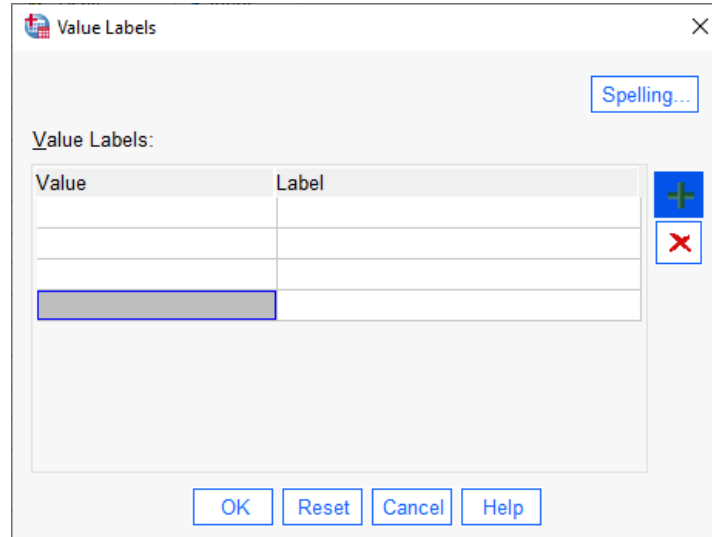
- In **"Gender"** row, move over to the **Values** column, click on the cell and then on the 3 dots button. A window will appear, press the green plus button on the right of this box **2 times**.
    - o Type: **1** in the 1st **Value** box and **Male** in the corresponding **Label** box.
    - o Type: **2** in the 2nd **Value** box and **Female** in the corresponding **Label** box, and then click **OK**.
- In "**Confidence**"row and under the **Label** column, type "on a scale of 1 – 10 with 10 being the most confident that you will learn statistics."
- In "**CompExp**" row and under the **Label** column, type "How many years of computer experience have you had?"
- In "**Case**" row and under the **Measure** column, click on the down arrow and select **Nominal**.
- In "**Gender**" row and under the **Measure** column, click on the down arrow and select **Nominal**.
- Using this same process change the remaining rows to **Scale**
- Change the value in the **Decimals** column to "0" for all variables but **GPA** and **HSGPA** variables.

Entering Data

- In **Data View**, from the **View** menu, select the **Value Labels** (checked), if it is not already selected.
- Enter the data from the survey results (see following table), using Tab or arrow keys.

Case	Gender	Age	Year	GPA	HSGPA	Confidence	CompExp
1	Male	23	Freshman	3.20	3.11	6	5
2	Male	25	Sophomore	3.00	2.90	8	2
3	Female	31	Senior	2.70	2.50	3	15
4	Male	27	Junior	3.90	3.60	10	4
5	Female	27	Sophomore	3.11	3.20	8	8
6	Male	22	Freshman	3.00	3.40	7	2
7	Male	22	Sophomore	3.50	3.80	10	9
8	Female	28	Junior	2.22	2.70	3	5
9	Male	22	Senior	2.90	3.10	5	1
10	Male	27	Senior	3.30	3.10	8	7

- The from the **File** menu, click **Save As** -> (choose a location).
- Type your last name in the **File Name** box.
- Select: **Save**.

Importing Data

- Data can be entered directly (like we just did above), or it can be imported from several different sources (e.g., **Microsoft Excel** or **Word**).
- Files created in **Microsoft Word** should be saved in **Plain Text** format before trying to import into SPSS Statistics.
- When importing **Excel** files, if your variable names contain spaces they will be **removed** during the importing process.

Exercise #2: Importing an Excel File

Suppose you have scores from three tests collected from a class of 10 students. For each student you have the following: an ID number, Gender, and scores for test one, two and three.

To import this Excel file:

- **File** -> **Import Data** -> **Excel…** Go to the location of the previously saved Grades.xlsx file and select it. Click **open.**
- Save the file as **Grades.SAV**, but do not close it.
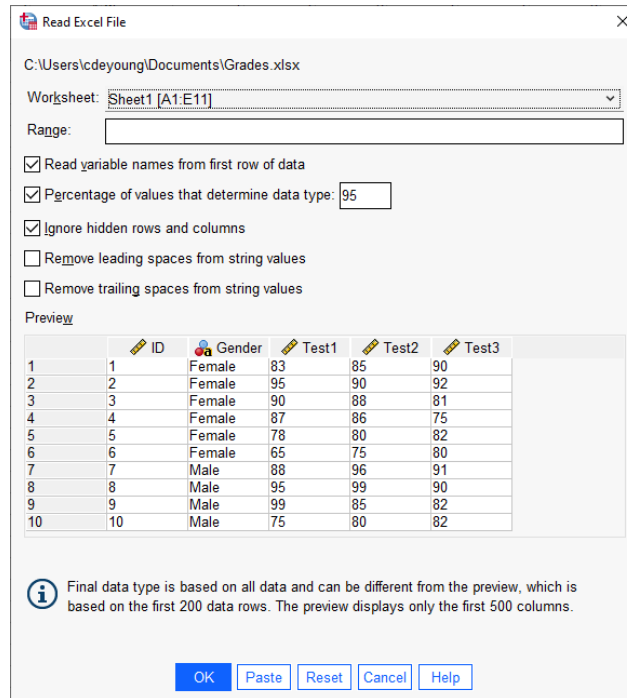
Exercise #3: Generating a New Variable "Average"

- From the **Transform** menu, click **Compute Variable…**
- In the box titled **Target Variable:** type what you would like your new variable to be called (lets title it **TestAverage**)
- In the box titled **Numerical Expression:** type how you would like the variable to be calculated. Since we would like to get test score averages, we will type **mean (test1, test2, test3)**, then click ok.
- A new Colum titled **TestAverage** will be displayed in the **Data Editor Window** (Data View tab) with the values of the mean score for each case.
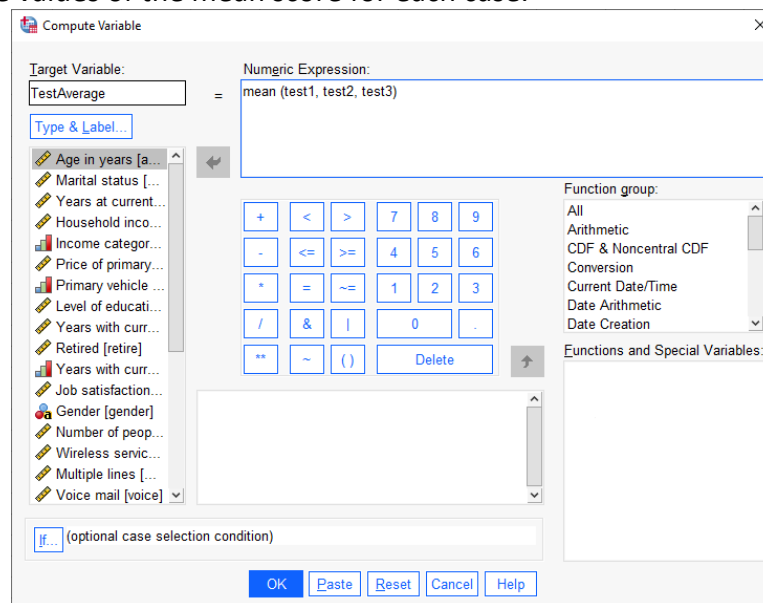
Procedures

- The **Frequencies** procedure (Analyze - > Descriptive Statistics -> Frequencies) is a useful tool for summarizing categorical and scale variables, one variable at a time.
- The **Descriptives** procedure (Analyze -> Descriptive Statistics -> Descriptives…) focuses specifically on scale variables and provides the ability to save standardized values (z scores) of your variables.
- The **Explore** procedure (Analyze -> Descriptive Statistics -> Explore) is useful for visually and numerically comparing groups, summarizing distributions examining the assumption of normality, and looking for outlying observations. It is easy to assume without looking that your data have no outliers, extreme values or distributional problems. If your dependent variable is categorical, try Crosstabs procedure.
- The **Crosstabs** procedure (Analyze -> Descriptive Statistics -> Crosstabs) allows you to obtain summaries for the relationship between two categorical (nominal or ordinal) variables.
- The **Means** procedure (Analyze-> Compare Means -> Means…) provides descriptive statistics and an ANOVA table for studying relationships between scale and categorical variables.
- The **Case Summaries** procedure (Analyze -> Reports -> Case Summaries) provides descriptive statistics and case summaries for studying relationships between scale and categorical variables.
- The **Correlate** procedure (Analyze -> Correlate) provides correlations between two scale variables (continuous)

Descriptives Procedure Options

- **Mean and Sum** – The mean, or arithmetic average is displayed by default.
- **Std. Deviation** – A measure of dispersion around the mean. In a normal distribution 68% of cases will fall within one standard deviation of the mean and 95% of cases will fall within two standard deviations.
- **Variance** – A measure of dispersion around the mean.
- **Range** – The difference between the largest and smallest values of a numeric variable.
- **Minimum** – The smallest value a of a numeric variable.
- **Maximum** – The largest value of a numeric variable.
- **S.E. Mean** – A measure of how much the value of the mean may vary from sample to sample taken from the same distribution.
- **Kurtosis** – Statistics that characterize the shape of the distribution (peak), For a normal distribution, the value of the kurtosis statistic is zero. Positive kurtosis indicates that, the observations are more clustered around the center distribution and have thinner tails.
- **Skewness** – A measure of the asymmetry of a distribution. The normal distribution is symmetric and has a skewness value of 0.

Obtaining Descriptive Statistics for Selected Variables

To quickly obtain descriptive statistics for selected variables:

- Right click the selected variable in either data or variable view.

- From the context menu select **Descriptive Statistics**.

By default, frequency tables are displayed for all variables with 24 or fewer unique values. Summary statistics are determined by variables measurement level and data type (numeric & string):

- **String**: No summary statistics are calculated for string variables.
- **Numeric, nominal**, **or unknown measurement level**: Range, minimum, maximum, and mode.
- **Numeric, ordinal measurement level:** Range, minimum, maximum, mode, mean and median.
- **Numeric, continuous (scale) measurement level:** Range, minimum, maximum, mode, mean median and standard deviation.

One click-descriptives

- To quickly run descriptive statistics, select variables in the **Data Editor** and get summary descriptives statistics (e.g., mean, median, frequency counts). Appropriate statistics are automatically determined based on the measurement level.

Exercise #4: Run Descriptive Statistics

Run descriptives procedure on the four continuous variables in the **Grades** data set that you created in **Exercise 2**.

- From the **Analyze** menu, select **Descriptive Statistics** -> **Descriptives…**
- Select the variables you would like to analyze, **Test1, Test2, Test3, and TestAverage** and use the arrow button to move them over to the box labelled "variable(s)".
- Click the **Options…** button, and select other statistics (e.g., variance) and display order (e.g., alphabetical), and click **continue**.
- Click **OK**.
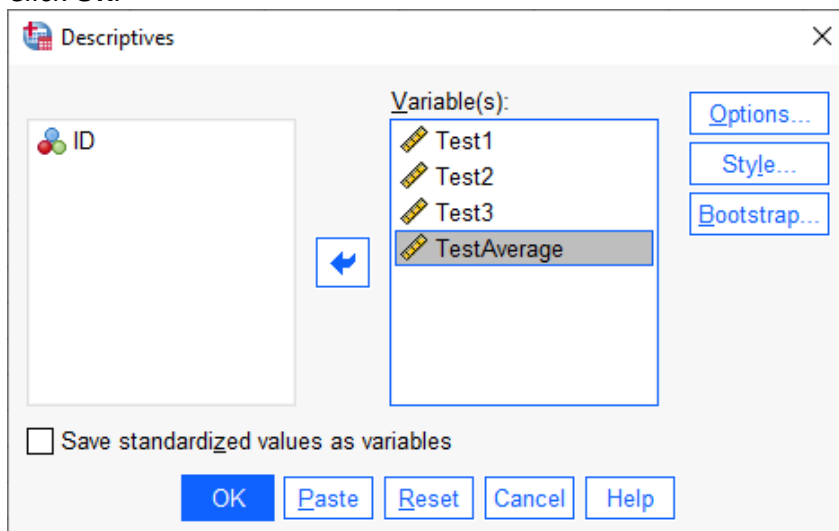
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
Test1	10	65	99	85.50	10.459	109.389
Test2	10	75	99	86.40	7.321	53.600
Test3	10	75	92	84.50	5.778	33.389
TestAverage	10	73.33	94.67	85.4667	6.73520	45.363
Valid N (listwise)	10					

Exercise #5 Compare Means

Run the **Means** procedure, using the **Grades** data set that you created in **Exercise #2.**

- From the **Analyze** menu, choose **Compare Means** -> **Means...**
- Select **Test1, Test2, Test3,** and the **TestAverage** as the **dependent variables**.
- Select **Gender** as the **independent variable**.
- Click **Options…**
- Mean, Number of cases and Standard Deviation are selected as statistics.
- Click **Continue**.
- Click **OK**.
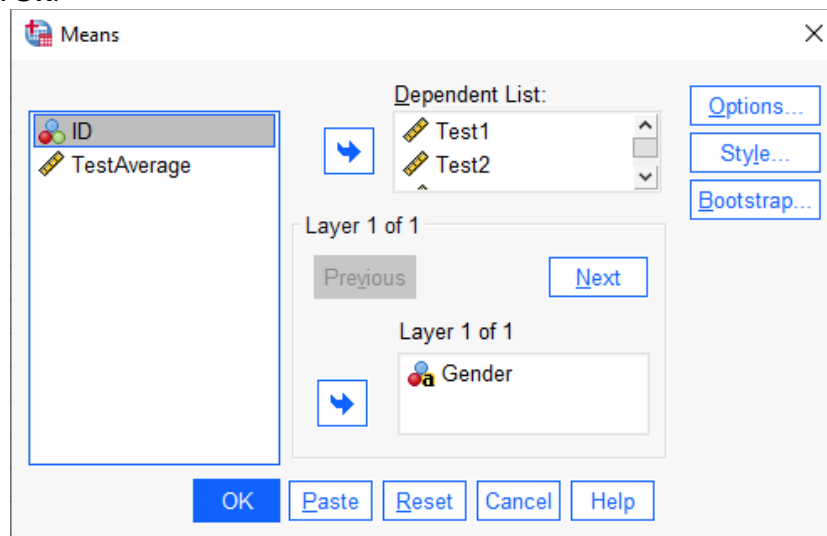
Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
Test1 * Gender	10	100.0%	0	0.0%	10	100.0%
Test2 * Gender	10	100.0%	0	0.0%	10	100.0%
Test3 * Gender	10	100.0%	0	0.0%	10	100.0%

Report

Gender		Test1	Test2	Test3
Female	Mean	83.00	84.00	83.33
	N	6	6	6
	Std. Deviation	10.564	5.550	6.439
Male	Mean	89.25	90.00	86.25
	N	4	4	4
	Std. Deviation	10.532	8.981	4.924
Total	Mean	85.50	86.40	84.50
	N	10	10	10
	Std. Deviation	10.459	7.321	5.778

Setting Viewer Options
- From the **Edit** menu, select **Options…**
- Click the **Viewer** tab.
- Select the settings that you want.
- Click **OK** or **Apply**

Note: Viewer output display options affect only new output produced after you make changes to viewer settings.

Editing Results Tables
- Output table objects can be opened for editing before you use them elsewhere (e.g., export to PDF).
  - Right-click on the table and then select **Edit…** you can then do use the following menu options in the window that appears:
    - Table looks…
    - Pivoting Trays
    - Edit Content
    - Text can be formatted, rows or columns removed, footnotes added, etc.
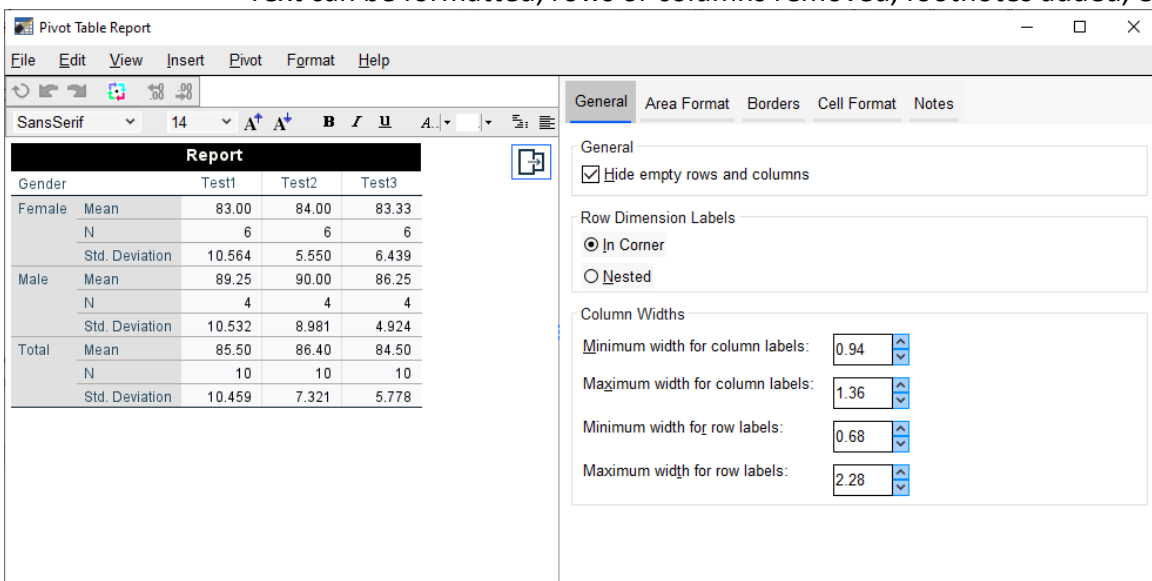
Saving the Output

The content of the viewer can be saved to a viewer document. The saved document includes both panes of the viewer window (the outline on the left and the contents on the right).

To Save a Viewer Document

- From the **Viewer** window menus choose **File** -> **Save As…**
- Enter the name of the document, and then click **Save.**
- To save the results in external formats (e.g., **HTML, PDF, or text**) use **Export** on the **File** menu.
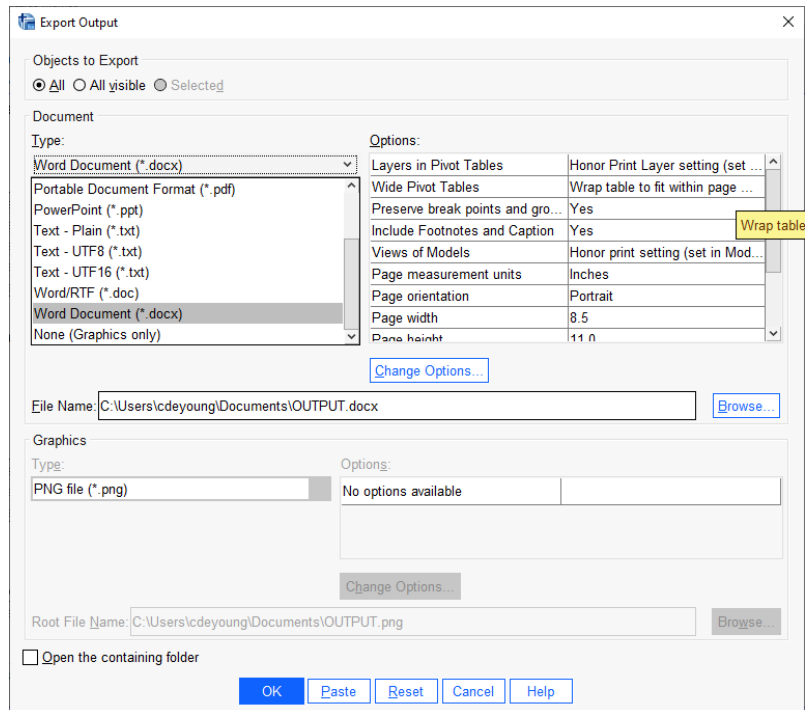
Exercise #6 Creating a Frequency Table

The frequencies procedure provides statistics and graphical displays that are useful for describing many types of variables.

- Open the **Demo.sav** sample file if it is not open.
- From the **Analyze** menu, choose **Descriptive Statistics** -> **Frequencies…**
- In the frequencies dialog box, click the Statistics button and then select the statistics you want (e.g., Mean, Median, Std. Deviation, Variance), click continue.
- Select **Income category in thousands (inccat)** and use the arrow button to move it into the **Variable(s)** Box.
- Click **OK.** Results will be displayed in the Output viewer.
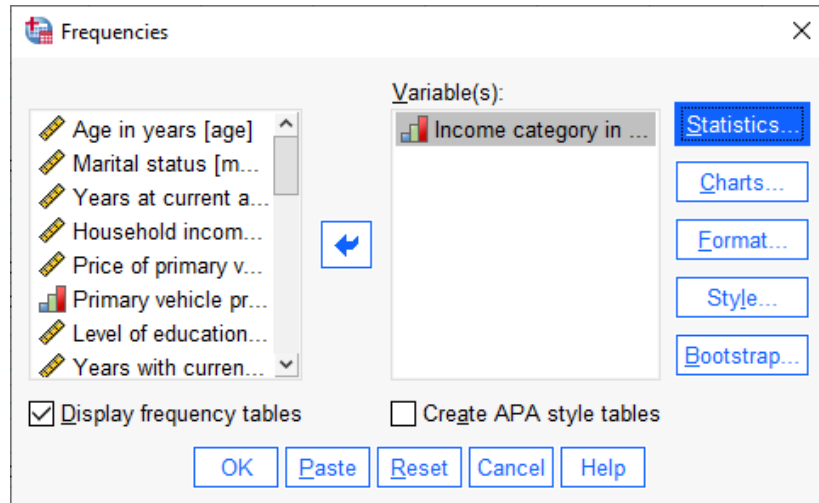
Exercise #7 Create a Chart

Using the Demo.sav sample file create a chart that shows the relationship between wireless telephone service and PDA (personal digital assistant) ownership.

- From the **Graphs** menu, click **Chart Builder…**
- Click **OK**.
- On the **Gallery** tab, choose the **Bar.**
- Drag the **Clustered Bar** icon onto the canvas (Large area above the chart).
  Note: without clustering, one bar would be drawn for each category
    - Right-click on the variable **Wireless** (categorical: nominal or ordinal data), and then select **Nominal** as its measurement level, and the X-axis
- Right-click on the **Owns PDA** (ownpda) variable and choose **Nominal** as its measurement level, and then
    - Drage the Owns PDA variable to the cluster drop zone in the upper right corner of the canvas, so that each bar is distinguished by color or pattern. To Change the default, choose from **Edit** menu -> **Options…**
- Click **OK.**

Carlee DeYoung, M.A.
Carlee.M.DeYoung@ttu.edu

Technology Support
Texas Tech University

Page 20
Updated 8/24/2022

Clustered Bar Count of Wireless service by Owns PDA
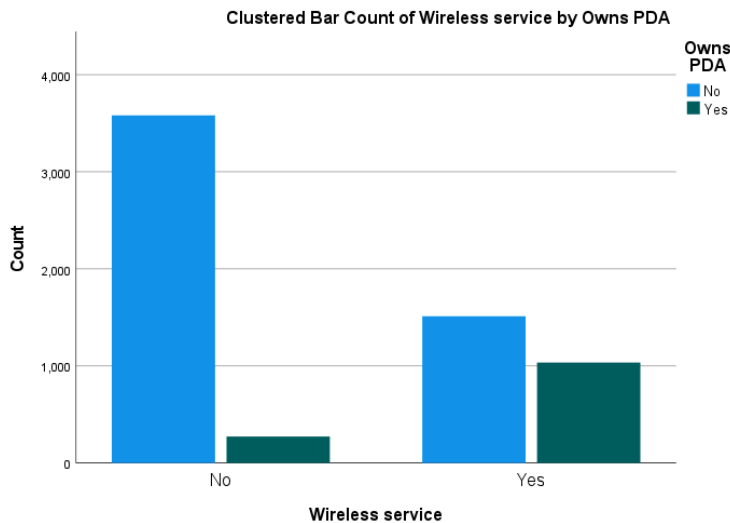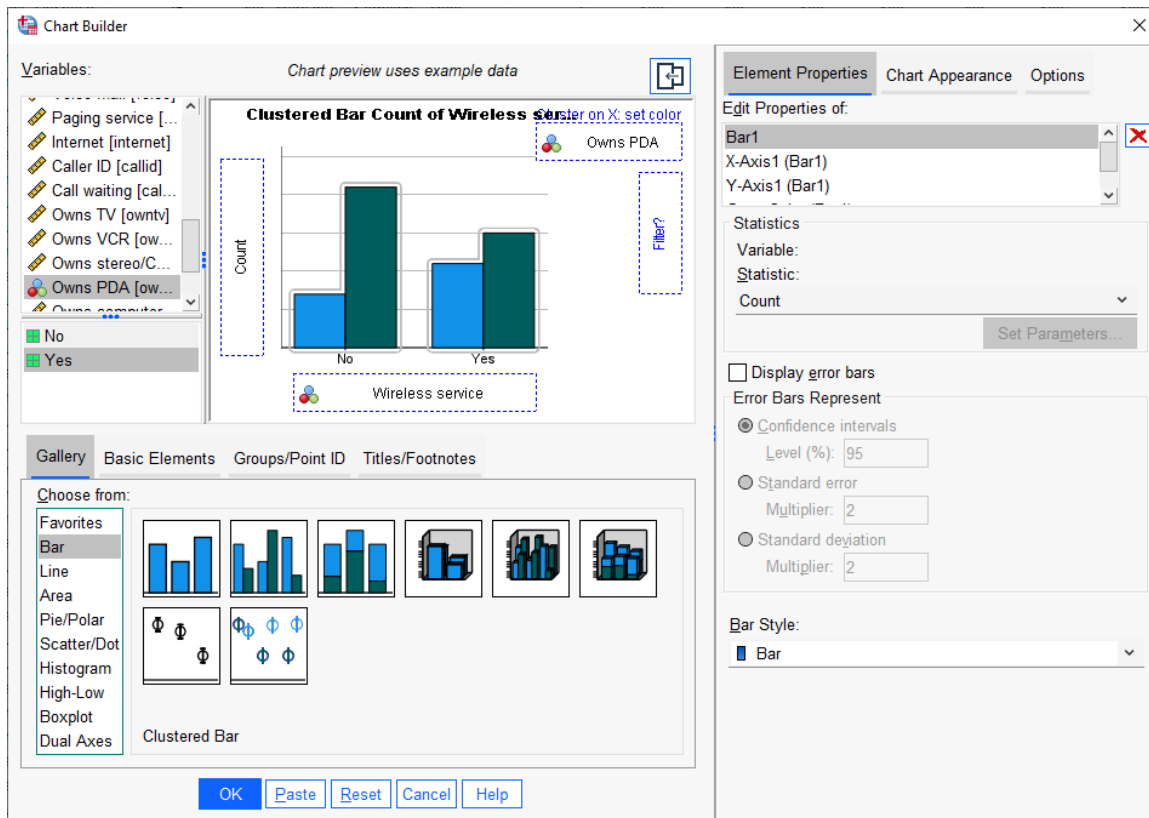
Exercise #8 Perform Descriptive Analysis

A team sells computer hardware to software development companies. At each company, representatives have a primary contact. Contacts are categorized by the department or the company in which they work (*Development, Computer Services, Finance, Other, Don't Know*). This information is collected in the **contacts.sav** file.

Use Frequencies to study department distributions

- Open the **contacts.sav** file (C:\ Program Files\IBM\SPSS Statistics\Samples\English\contacts.sav)

- From the **Analyze** menu, **Descriptive Statistics** -> **Frequencies…**
- Select the Department as an analysis variable, and then click **Charts…**
- Select **Pie charts** (for assessing the relative frequencies of each category), and then click **Continue.**
- Click **OK.**
- Right-click on the chart, select Edit, in the chart editing window menu select **Elements**, and click on the **Show Data Labels** option.
- **Close** the Chart Editor.
- Minimize the Output Window.
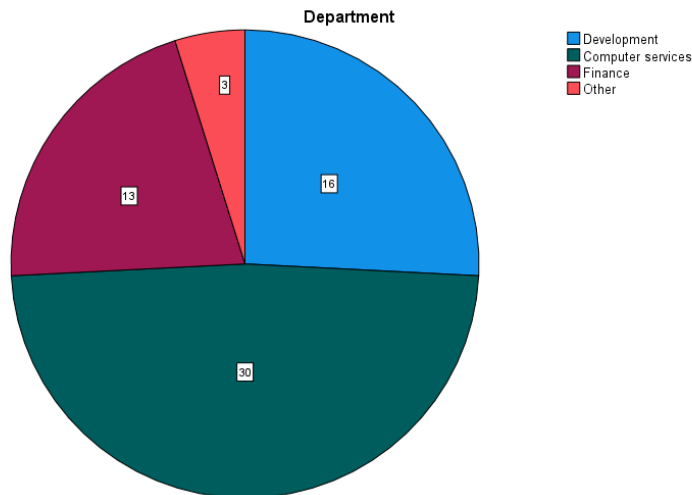
Exercise #9 Create Bar Chart

Create a Bar Chart for the Departments variable, using contacts.sav file.

- Click the recently used dialog button [button] and select **Frequencies**.
- Click **Charts…**
- Select **Bar charts** and click continue.
- Click the **Format…** button, select **Descending counts,** click **Continue.**
- Click, **OK.**

Carlee DeYoung, M.A.
Carlee.M.DeYoung@ttu.edu
Technology Support
Texas Tech University
Page 22
Updated 8/24/2022
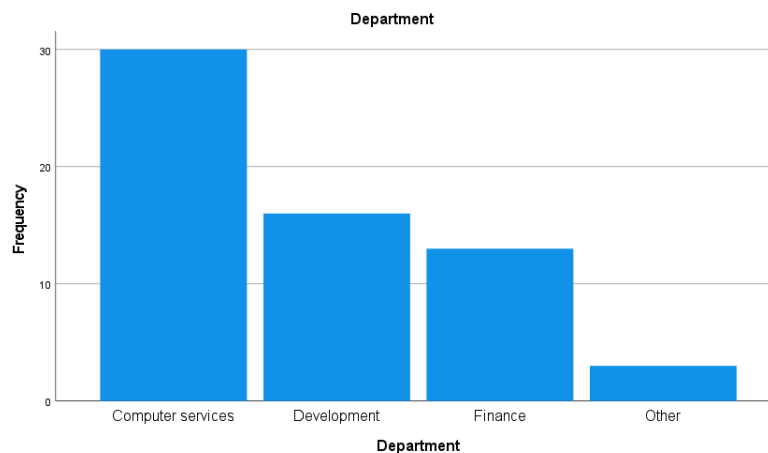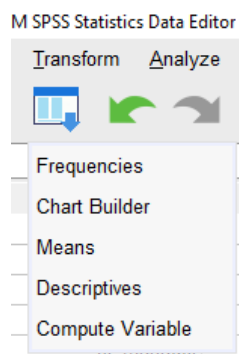
Exercise #10 Summarize Company Ranks

Summarize the company ranks, using the contacts.sav file.

- From the **Analyze** menu, select **Descriptive Statistics** -> **Frequencies…**
- Click **Reset** to restore the default settings.
- Select **Company Rank** as an analysis variable, and click **Charts…**
- Selecty **Bar charts** and click **Continue.**
- Click **Format---**, select **Descending values**, and click **Continue.**
- Click **OK** in the frequencies dialog box. These selections produce a frequency table and bar chart with the categories ordered by **descending value.**

Exercise #11 Summarize Last Sales Amounts

Using the **contacts.sav** file summarize the amounts of the last sale.

- From the **Analyze** menu, choose **Descriptive Statistics,** and then select **Frequencies…** and click **Reset** to restore default settings.
- Select **Amount of Last Sale** as an analysis variable.
- Deselect **Display frequency tables** box.
  Note: it is a good idea to turn off the display of frequency tables for scale data because this type of data usually has many different values.
- Click **Statistics** in the Frequencies dialog box.
  Check **Quartiles**, **std. deviation, Minimum, Maximum, Mean, Median, Skewness,** and **Kurtosis** and click **Continue.**
- Click **Charts…** and select **Histograms,** select **Show normal curve** on histogram and click **Continue**.
- Click **OK.**

Histogram

Mean = 55.45
Std. Dev. = 103.939
N = 70

Amount of last sale
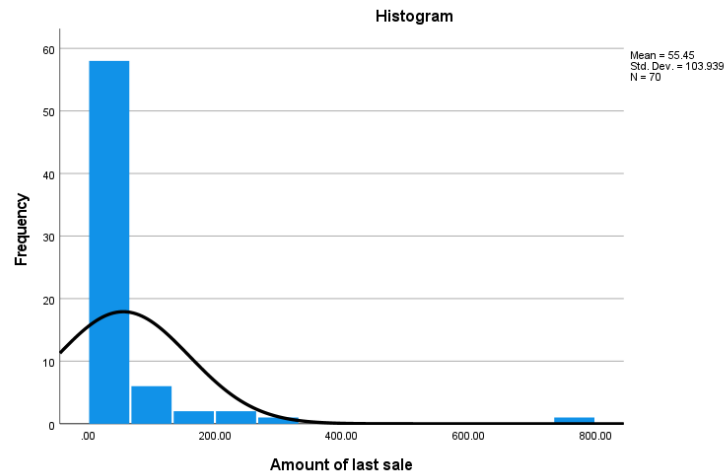
Results:

Notice that the distribution is asymmetric, with some distant values in a positive direction from the center of the distribution. Most variables with finite lower limit ($0) but no fixed upper limit tend to be positively skewed. The histogram is a visual summary of the distribution of values. The overlay of the normal curve helps you assess the skewness and kurtosis.

Exercise #12 Transformation of a Variable

Transformation of a variable using the **contacts.sav** file.

Sometimes, a transformation of the variable can bring the distribution of values closer to normal.

- To transform the variable **sale**, from the **Transform** menu, choose **Compute Variable**…
- Type **logsale** as the **Target Variable**.
- Type **LN(sale)** as the **Numeric Expression** for log transformation of Sale variable.
- Click **OK**. The log transformation is a good choice because **Amount of Last Sale** takes only positive values and is right skewed.

- **Recall** the Frequencies dialog box **,** deselect **amount of last sale** as an analysis variable, and select **logsale** as an analysis variable, and then click **OK.**
  Note: From the statistics table, it appears that the transformation has brought the distribution closer to normal. The skewness and kurtosis are greatly reduced, and the mean and median are much closer together. The histogram is also much more aligned with the normal curve displayed over it.
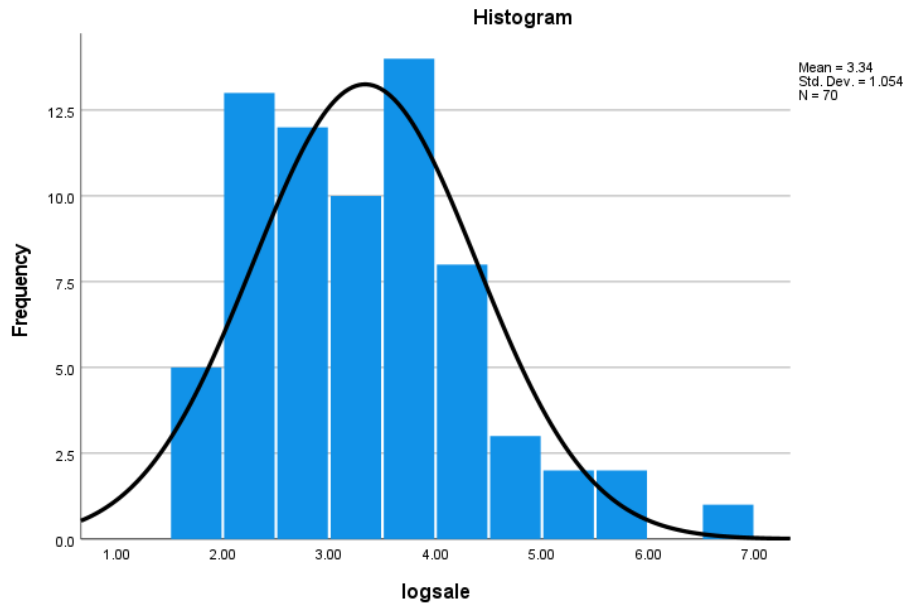
Exercise #13 Study Customer Spending

A telecommunication company maintains a customer database that includes information on how much each customer spend on long distance, toll-free, equipment rental, calling card, and wireless services in the previous month. This information is collected in **telco_missing.sav**. Study customer spending to determine which services are most profitable.

- Open the **telco_missing.sav** file (C:\ Program Files\IBM\SPSS Statistics\Samples\English\telco_missing.sav)
- From the **Analyze** menu, choose **Descriptive Statistics** -> **Descriptives…**
- Select **Long distance last month, Toll free last month, Equipment last month, Calling card last month,** and **Wireless last month** as analysis variables.
- Click **OK**. *Notice* that most customers don't have every service, so a lot of 0's are being counted. One solution to this issue is to treat 0's as missing values so that the analysis for each service becomes conditional on having that service.
- To **recode 0's** as **missing values**, from the **Transform** menu, choose **Recode Into Same Variables…**
    - Select all the variables, as analysis variables.
    - Click **Old and New Values…,** Type 0 as the Old Value.
    - Select **System-missing** as New Value.
    - Click **Add**, and then click **Continue.**
    - Click **OK**.
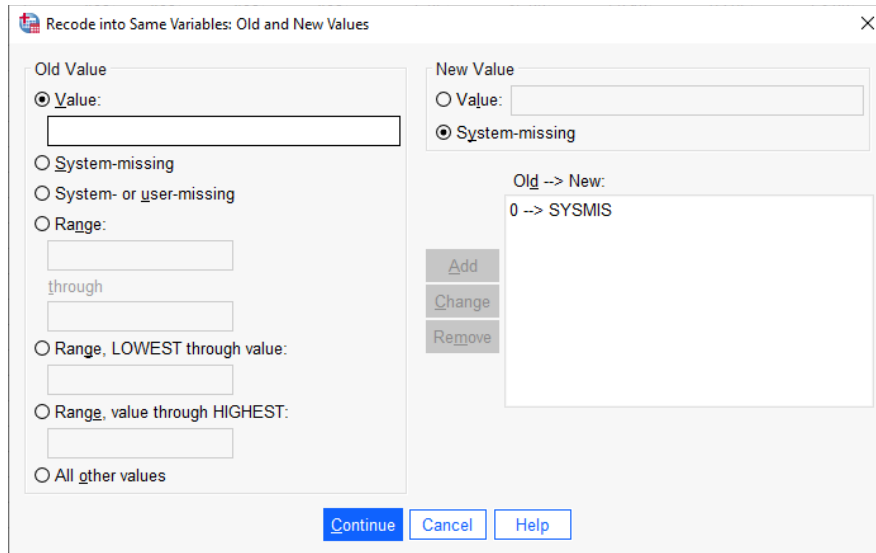
Exercise #14 Running Descriptive Analysis

Use recoded data from **Exercise #13**.

- Recall the **Descriptives** dialog box.
- Click **Options** in the Descriptive dialog box and deselect **Minimum** and **Maximum** -> select **Skewness** and **Kurtosis** -> Click **Continue** -> click **OK.**
  Note: when the analysis is conditional upon the customer's actually having the service, the results are dramatically different. This hasn't solved the problem of who purchases these services, but it does point you in the direction of which services deserve greater marketing.

Exercise #15 Exploring Distributions

A manufacturing firm uses silver nitride to create ceramic bearings, which must resist temperatures of 1500 degrees centigrade or higher. The heat resistance of a standard alloy is known to be normally distributed. However, a new premium alloy is under test, and its distribution is unknown. Explore the distribution of the file **ceramics.sav**.

- Open the **caramics.sav** file (C:\ Program Files\IBM\SPSS Statistics\Samples\English\ceramics.sav)
- From the **Analyze** menu, choose **Descriptive Statistics** and then select **Explore…**
- Select **Degrees Centigrade** as the dependent variable.
- Select **Alloy** as the Factor List.
- Label cases by **labrunid** variable.
- Click the **Statistics** button.
- Select **Outliers** and then click **Continue.**
- Click **Plots…** in the Explore dialog box, check the **Normality plots** with test box, and click **continue.**
- Click **OK.**
- Review and then minimize the output.

Carlee DeYoung, M.A.
Carlee.M.DeYoung@ttu.edu

Technology Support
Texas Tech University

Page 26
Updated 8/24/2022

Normal Q-Q Plot of Degrees Centigrade
for batch= Premium

Normal Q-Q Plot of Degrees Centigrade
for batch= Standard

Detrended Normal Q-Q Plot of Degrees Centigrade
for batch= Premium

Detrended Normal Q-Q Plot of Degrees Centigrade
for batch= Standard
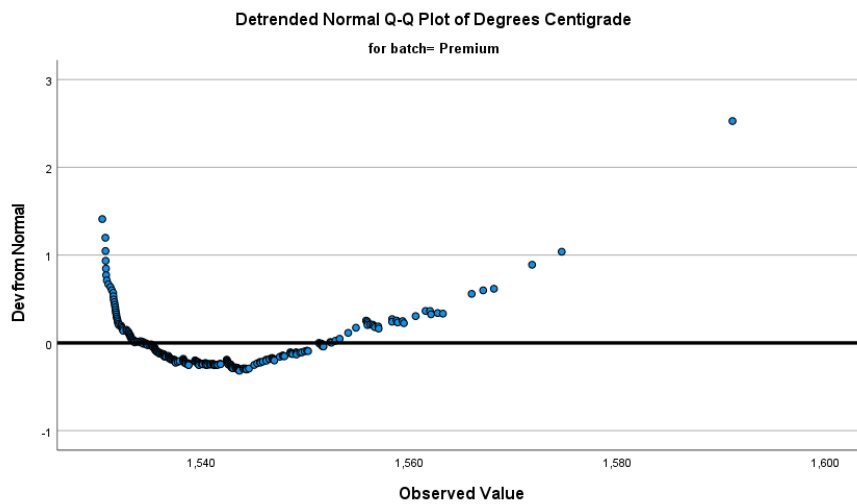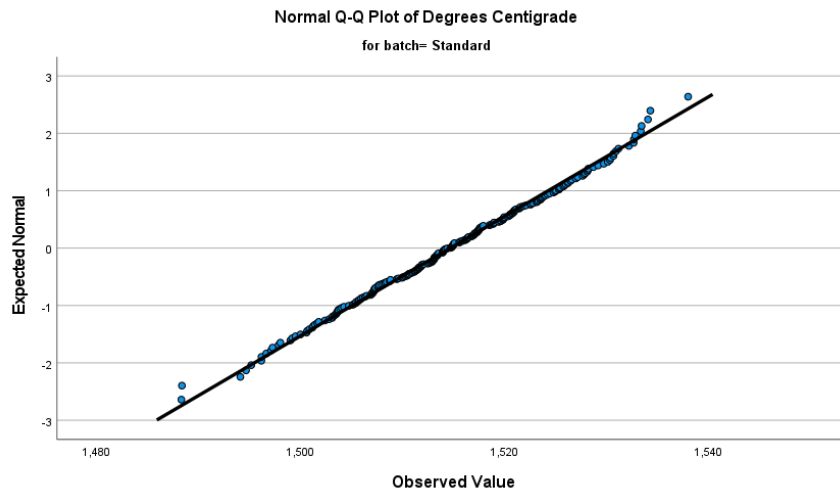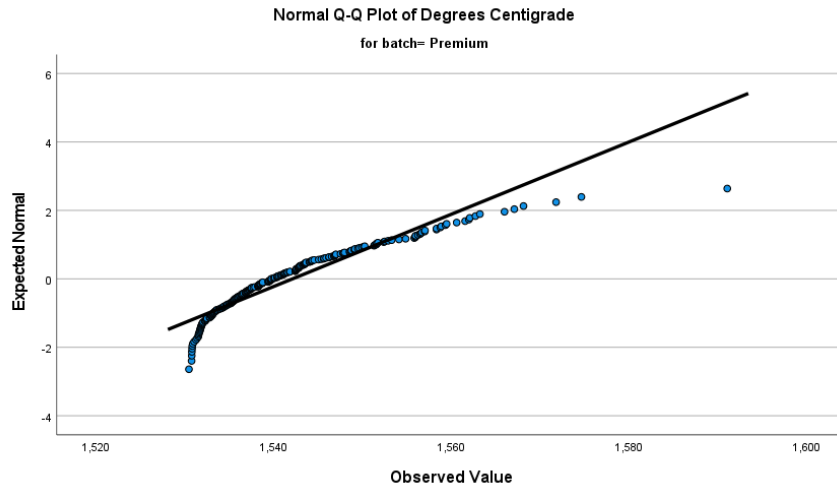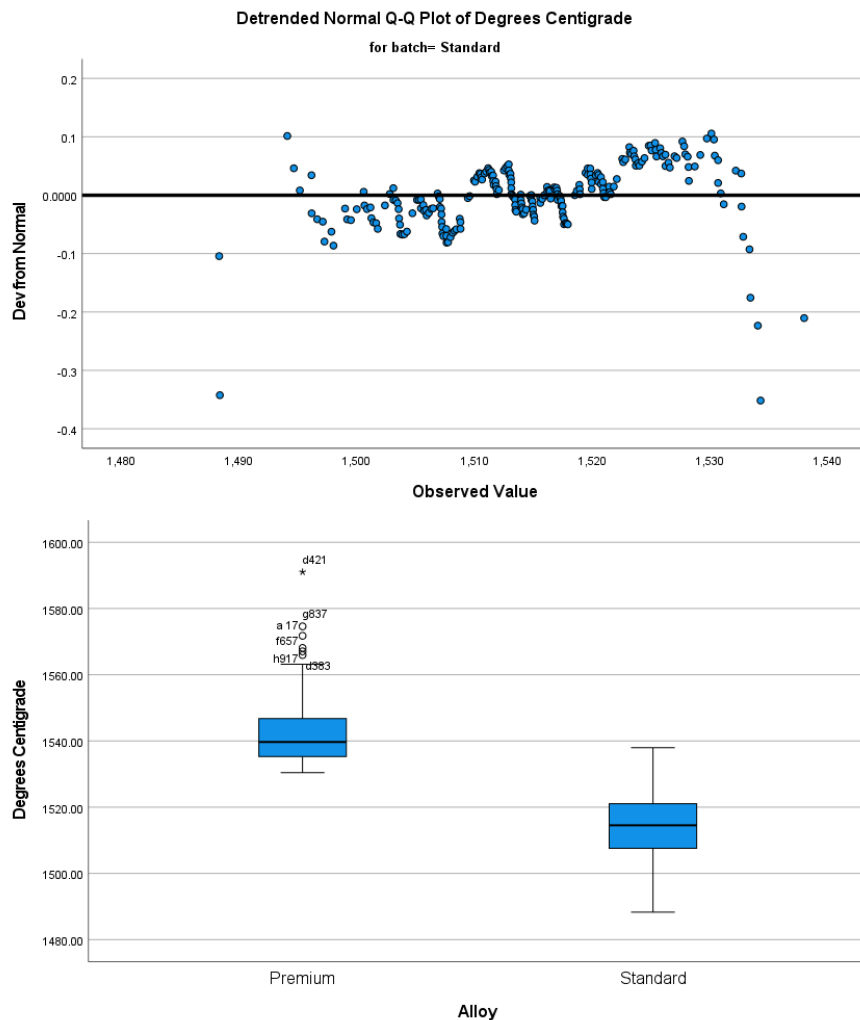
Results:

- The **tests of normality** overlay a normal curve on actual data, to assess the fit. A significant test means the fit is poor.
- For the standard alloy, the test is not significant; the fit the normal curve well. However, for the premium allot the test is significant; they fit the normal curve poorly.
- **Stem- and-leaf plots** -- for premium bearings visualizes the positive skew statistics seen in the descriptive table; the values cluster uniformly in a range of 1530 to 1543 degrees, then disperse gradually at the higher temperatures.
- **Q-Q plot** – Displays a straight line in the plot and represents expected values when the data are normally distributed. The observed premium bearing values deviate markedly from that line, as temperature increases.

Exercise #16 Use Crosstabs Analysis

To determine customer satisfaction rates, retail company conducted surveys of **582 customers at 4 store locations**. From the survey results, you found that the quality of customer service was the most important factor to a customer's overall satisfaction. Given this information, you

want to test whether each of the store locations provides a similar and adequate level of customer service. The results of the survey are stored in **satisf.sav.** Use crosstabs analysis (Chi-squared test) to test the hypothesis that the levels of service satisfaction are constant across stores (comparing observed and expected).

- Open the **satisfy.sav** file (C:\ Program Files\IBM\SPSS Statistics\Samples\English\satisf.sav)
- From the **Analyze** menu, choose **Descriptive Statistics.**
- Select **Crosstabs.**
- Select **Store** as the row variable and **Service satisfaction** as the column variable.
- Click **Statistics.**
- Select **Chi-square, contingency coefficient,** and click **continue.**
- Click **OK.** These selections produce a cross tabulation table, a chi square test, and a nominal-by-nominal measures of association for Store by Service satisfaction.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Store * Service satisfaction	582	100.0%	0	0.0%	582	100.0%

Store * Service satisfaction Crosstabulation

Count

		Service satisfaction					Total
		Strongly Negative	Somewhat Negative	Neutral	Somewhat Positive	Strongly Positive	
Store	Store 1	25	20	38	30	33	146
	Store 2	26	30	34	27	19	136
	Store 3	15	20	41	33	29	138
	Store 4	27	35	44	22	34	162
Total		93	105	157	112	115	582

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	16.293[a]	12	.178
Likelihood Ratio	17.012	12	.149
Linear-by-Linear Association	.084	1	.772
N of Valid Cases	582		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 21.73.

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Contingency Coefficient	.165	.178
N of Valid Cases		582	

Results:

- The cross tabulation shows the frequency of each response at each store location. If each store location provides a similar level of service, the patter of responses should be similar across stores.
- At each store, most responses occur in the middle.
- Store 2 appears to have fewer satisfied customers.
- From the cross tabulation alone, it is impossible to tell whether these differences are real (statistically significant) or due to chance variation. Check the chi-square test to be sure.
- The ci-square test measures the discrepancy between the observed cell counts and what you would expect if the rows and column were unrelated.
- The two-sided **asymptotic significance** of the chi-square statistics 0.178 is greater than 0.10, so we cannot reject the null hypothesis (no significant differences). Therefore, it is safe to say that the differences are due to chance variation, which implies that each store offers the same level of customer service. However, not all customers surveyed had contact with a service representative. The ratings from these customers will not reflect the actual quality of service at a store, so you further cross classify by whether they had contact with a service representative.

Exercise #17 Splitting Data

SPLIT FILE command splits the active dataset into subgroups that can eb analyzed separately. SPLIT FILE can specify both **numeric** and **string** split variables. SPLIT FILE can specify up to eight variables.

Split the data in the Cross-tabulation table, and then calculate Chi-square test for customers who has contact with store representative and a Chi-square test for customers who did not have contact with store representative.

- Recall the **Crosstabs** dialog box.
- Select **Contact with employees** as a layer variable
- Click **OK** in the crosstabs dialog box. The cross tabulation now splits the previous cross tabulation into two parts. Now that the customers who didn't have contact are sorted out, there appears to be a significant association between store 2 and low service satisfaction.
- Check the chi-square test to be sure.
- The Chi-square test is [performed separately for customers who did and did not have contact with a store representative.

- The significance value of the test for customers who did not have contact is 0.052. This is suggestive, but not conclusive, evidence of a relationship between Store and Service satisfaction for these customers.
- The significance value of the test for customers who had contact with an employee is 0.012. Since this value is less then 0.05, you can conclude that the relationship observed in the cross tabulation is real and not due to chance.

  Note: While the chi-square test is useful for determining whether there is a relationship, it doesn't tell you the strength of the relationship. Symmetric measures (below) attempt to quantify this.

Store * Service satisfaction * Contact with employee Crosstabulation

Count

Contact with employee			Strongly Negative	Somewhat Negative	Neutral	Somewhat Positive	Strongly Positive	Total
No	Store	Store 1	16	9	18	17	19	79
		Store 2	2	15	16	13	12	58
		Store 3	9	14	23	22	14	82
		Store 4	17	14	19	10	10	70
	Total		44	52	76	62	55	289
Yes	Store	Store 1	9	11	20	13	14	67
		Store 2	24	15	18	14	7	78
		Store 3	6	6	18	11	15	56
		Store 4	10	21	25	12	24	92
	Total		49	53	81	50	60	293
Total	Store	Store 1	25	20	38	30	33	146
		Store 2	26	30	34	27	19	136
		Store 3	15	20	41	33	29	138
		Store 4	27	35	44	22	34	162
	Total		93	105	157	112	115	582

Service satisfaction column headers span: Strongly Negative, Somewhat Negative, Neutral, Somewhat Positive, Strongly Positive

Chi-Square Tests

Contact with employee		Value	df	Asymptotic Significance (2-sided)
No	Pearson Chi-Square	20.898[b]	12	.052
	Likelihood Ratio	22.937	12	.028
	Linear-by-Linear Association	3.514	1	.061
	N of Valid Cases	289		
Yes	Pearson Chi-Square	25.726[c]	12	.012
	Likelihood Ratio	25.777	12	.012
	Linear-by-Linear Association	1.993	1	.158
	N of Valid Cases	293		
Total	Pearson Chi-Square	16.293[a]	12	.178
	Likelihood Ratio	17.012	12	.149
	Linear-by-Linear Association	.084	1	.772
	N of Valid Cases	582		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 21.73.

b. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 8.83.

c. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 9.37.

Symmetric Measures

Contact with employee			Value	Approximate Significance
No	Nominal by Nominal	Contingency Coefficient	.260	.052
	N of Valid Cases		289	
Yes	Nominal by Nominal	Contingency Coefficient	.284	.012
	N of Valid Cases		293	
Total	Nominal by Nominal	Contingency Coefficient	.165	.178
	N of Valid Cases		582	

Checking Chi-Square Assumptions

- **Independence of Observations:**
  - Each observed case is independent of each other (i.e., from a different person).
  - A Chi-square would be inappropriate if a person could have responses in more than one category or could have more than one contribution to a single category. You would not use a chi-spare for repeated-measure design.

Pearson Correlation Analysis

- The correlation coefficient is a measure of the **linear relationship** between two variables. The correlation coefficient is also known as the Pearson product-moment correlation denoted by **r** The value of **r** can range from -1 to +1 and is independent of the units of measurement.
- For perfect linearity, **r=±1**. If there is no linear relationship the value of r is close to 0.

- A correlation analysis is performed to quantify the **strength of association** between two **numeric variables.**
- The simplest way to find out about the relationship between two variables is to plot the data. Then use correlation analysis to determine the extent to which changes in the value of an attribute are associated with changes in another attribute.

Exercise # 18 Using Correlations to Study Associations

Study the associations between Motor Vehicles Sales and Fuel Efficiency using correlations. To increase sales, motor vehicles design engineers want to focus on aspects of the vehicle that are important to customers—for example how important is fuel efficiency with respect to sales? We can investigate this relationship by computing the correlation between past sales and fuel efficiency. Information concerning various makes of motor vehicles is collected in **car_sales.sav.**

- Open the **car_sales.sav** file (C:\ Program Files\IBM\SPSS Statistics\Samples\English\car_sales.sav)
- From the **Analyze** menu, choose Correlate -> Bivariate…
- Select **Sales in thousands** and **Fuel efficiency** as analysis variables.
- Click OK. The procedure produces a correlation matrix for sales and mpg. The Pearson correlation coefficient measures the linear association between two scale variables. The correlation reported in the table is negative (-0.017), Although not significantly different from 0 because the p-value of 0.837 is greater then 0.10. This suggests that designers should not focus their efforts on making cars more fuel efficient because there isn't an appreciation effect on sales.
  Note: The Pearson correlation coefficient works best when the variables are approximately normally distributed and have not outliers. A scatter plot can revel these possible patterns.

Exercise #19 Produce a Scatter Plot

To produce a scatter plot of Sales in thousands by Fuel efficiency, using the car_sales.sav file:

- From the **Graphs** menu, choose **Chart Builder… ->** click **OK.**
- Select the Scatter/ Dot gallery and choose Simple Scatter (double-click).
- Select Sales in thousands as the y variable and Fuel efficiency as the x variable.
- Click the Groups/ Point ID tab and select the Point ID Label Box.
- Select Model variable as the variable to label cases by (drag the Model variable to Variable box on the chart).
- Click OK. The resulting scatter plot shows two potential outliers, the metro in the lower tight of the plot and the F-series in the upper left. The F-series is found to be generally representative of the vehicles your design team is working on, so you decide to keep it in the data set for now. This point may appear to be an outlier because of the skew distribution of Sales in thousands, so try replacing it with **Log-transformed** sales (lnsales) in further analyses. The metro is not representative of vehicles that your design team is working on, so you can remove it from further analyses for now.

Scatter Plot of Sales in thousands by Fuel efficiency
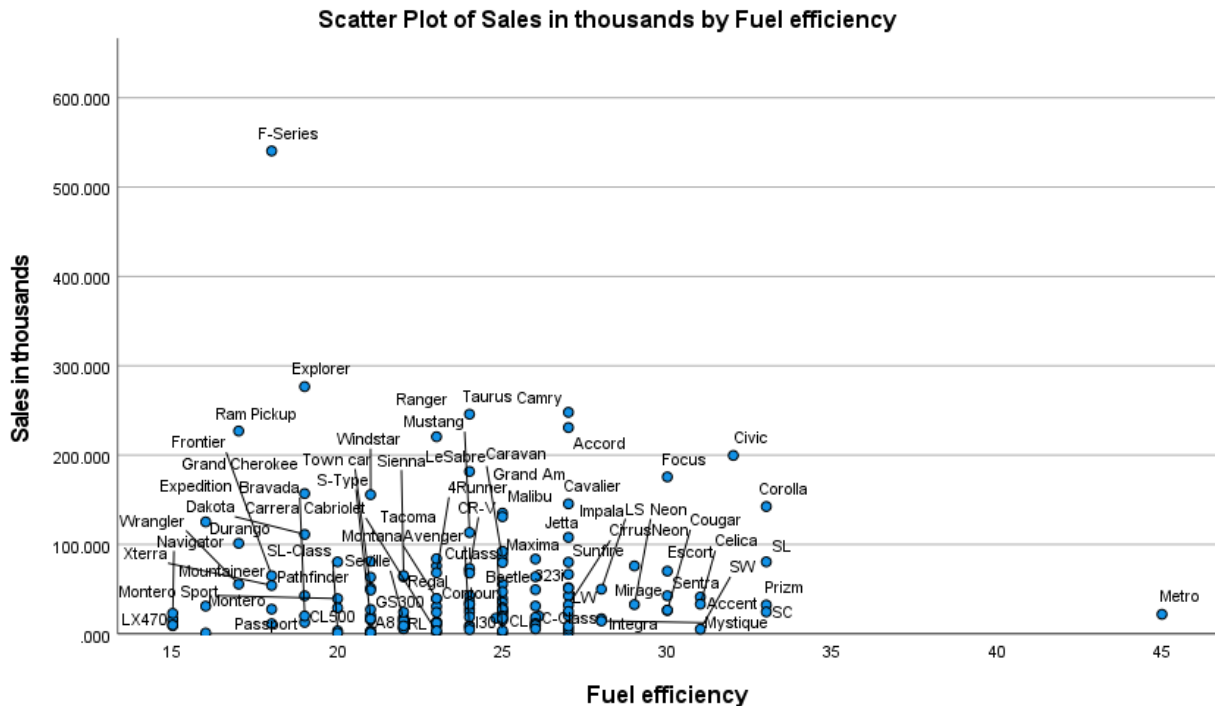
## Exercise #20 Remove Variable from Correlation Computations
- From the **Data** menu, click **Select Cases…**
- Select **If condition is satisfied** radio button and click **If…** button
  - Type **model ~= 'Metro'** in the text box. *Note:* Metro is a string constant and must be enclosed in quotation marks. Also, we must use the "~=", "**Not Equal**" conditional operator, since we want to **Exclude** the 'Metro' case in the sample. Most conditional expression use one or more of the six relational operators (<, >, <=, >=, =, and ~=) or calculator pad.
  - Click Continue, and then click OK in the select cases dialog box. A new variable has been created that uses all cases except for the Metro in further computations.
  - To analyze the filtered data, recall the Bivariate Correlations dialog box.
  - **Deselect** sales in thousands as an analysis variable.
  - **Select** Log-transformed sales (lnsales) as an analysis variable.
  - Click OK. After removing the outlier and looking at the log transformed sales, the correlation is now positive but still not significantly different from 0. However, the customer demographics for trucks and automobiles are different, and the reasons for buying a truck vs. car may differ. It's worthwhile to look at another scatter plot. This time marking trucks and cars separately.

Exercise #21 Control for Vehicle Type
To produce a scatter plot of Log-transformed sales by Fuel efficiency, controlling for Vehicle type, recall the Chart Builder.
- Select **Simple Scatter**, chart type (on the **gallery** tab).

- Select **Log-Transformed** sales as the**y-variable.**
- Select **Vehicle Type** as the variable to **set colors by**.
- Click OK. *Note:* The scatter plot shows the **trucks** and **automobiles** form distinctly different groups. By splitting the data file according to Vehicle type, you might get a more accurate view of the association. Also, that following the log transformation of sales, the potential outlier in the upper left has disappeared.
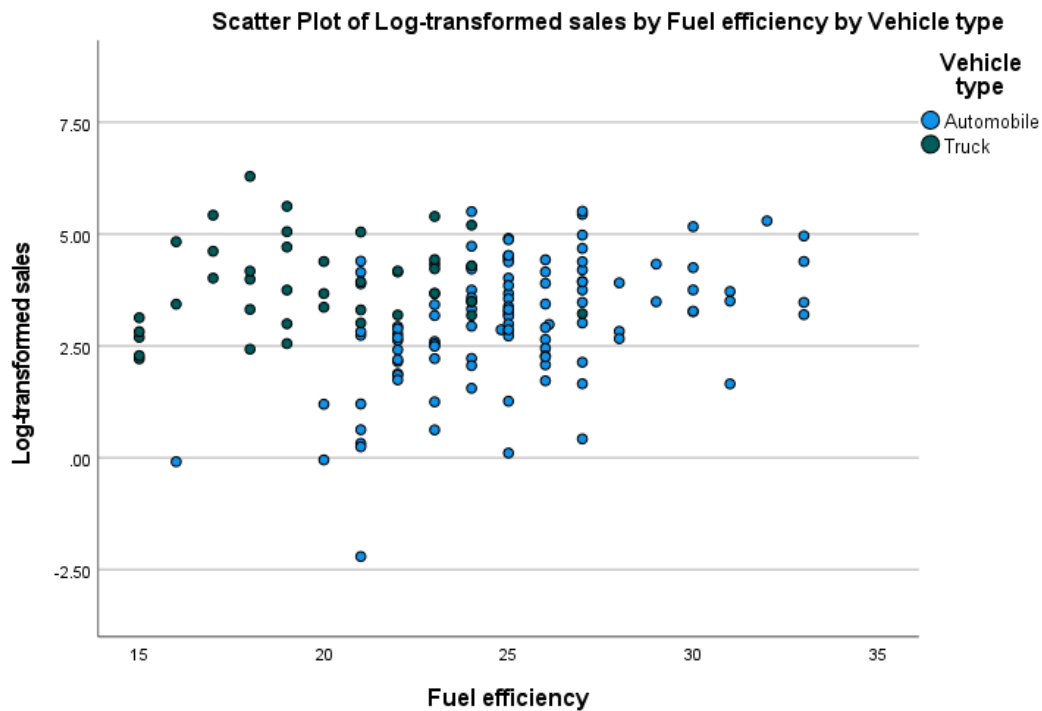
Exercise #22 Split Data File by Vehicle Type

To split the data file according to Vehicle type, from the Data menu, choose Split File…

- Select **Compare groups** radio button.
- Select **Vehicle type** as the variable on which groups should be based, and then click **OK**.
- To analyze the split file recall the Bivariate Correlations dialog box, and then click OK.
- Splitting the file on vehicle type has made the relationship between sales and fuel efficiency clearer.
- There is a significant and strong positive correlation between sales and fuel efficiency for automobiles.
- For trucks, the correlation is positive but not significantly different from zero.
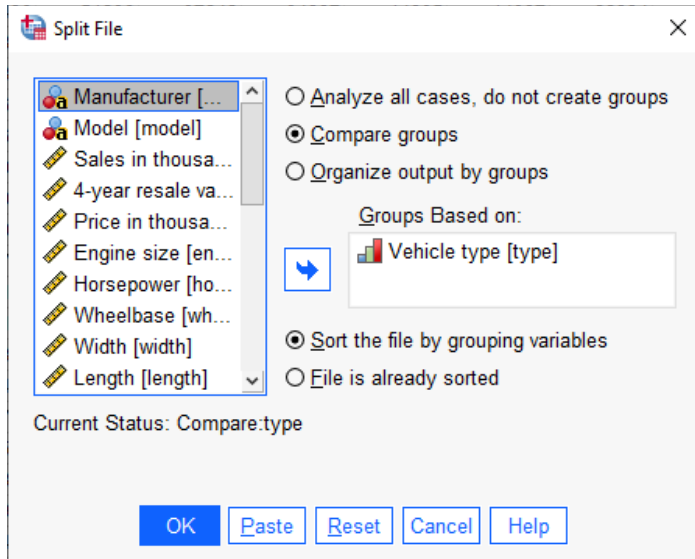
Correlations

Vehicle type			Sales in thousands	Fuel efficiency
Automobile	Sales in thousands	Pearson Correlation	1	.328**
		Sig. (2-tailed)		<.001
		N	115	113
	Fuel efficiency	Pearson Correlation	.328**	1
		Sig. (2-tailed)	<.001	
		N	113	113
Truck	Sales in thousands	Pearson Correlation	1	.011
		Sig. (2-tailed)		.944
		N	41	40
	Fuel efficiency	Pearson Correlation	.011	1
		Sig. (2-tailed)	.944	
		N	40	40

**. Correlation is significant at the 0.01 level (2-tailed).