SAUDER
School of Business

Real Estate Division

LESSON 2

Statistical Software Applications for Real Estate Analysis

Note: Selected readings can be found under "Online Readings" on your Course Resources webpage

Assigned Reading

1. UBC Real Estate Division. 2014. *BUSI 344 Course Workbook*. Vancouver: UBC Real Estate Division.
   Lesson 2: Statistical Software Applications for Real Estate Analysis

Recommended Reading

1. UBC Real Estate Division. 2009. *Advanced Computer-Assisted Mass Appraisal*. Vancouver: UBC Real Estate Division.
   Chapter 1: Univariate Statistical Analysis
   Chapter 2: Multivariate Analysis
   Chapter 12: Statistical Procedures and Performance Evaluation I
   Appendix B: Statistical Tables (including chi-square)

Learning Objectives

After completing this lesson, the student should be able to:

1. carry out basic procedures for a database including opening and saving, producing reports, and printing reports;
2. examine data using descriptive statistics;
3. examine data using graphic analysis, including histograms, scatterplots, and boxplots;
4. examine data using crosstabulation tables;
5. evaluate the correlation between variables;
6. identify variables as multiplicative or additive in basic form;
7. identify whether variables are continuous or discrete; and
8. determine the need for and type of transformations of variables, and carry out basic transformations.

Instructor's Comments

Salesperson: This software will cut your workload by 50%.
Appraiser: That's great, I'll take two of them.

Lesson 1 introduced the statistics that will be used throughout this course in a variety of practical applications. The statistics serve as a management tool: a way to interpret large volumes of data, and manage it to serve a variety of useful purposes. All statistics can be calculated manually using long and convoluted mathematical equations, but luckily for us, the age of computers has negated the need for these, and most statistics can be easily calculated at the push of a button. So, today, statistics is no longer the secret art of math geniuses, but is accessible by pretty much everybody. Of course, with increasingly easy access comes the danger of applying statistics thoughtlessly and uncritically, with the potential for alarming results – but we'll leave that discussion for later in the course.

The statistical calculations in this course can be easily performed using business calculators, business software (Excel), or high powered statistics software. This lesson will provide a guide to the use of two software tools that learners in this course will find of value:

- Microsoft Excel; and
- SPSS – Statistical Package for the Social Sciences.

This is not meant to be an exhaustive list of software that could be used for data analysis, but we have had to narrow down the vast list of possible software to just a few specific tools to illustrate in the course. For example, we have recommended Microsoft Excel, because it is included in the Microsoft Office Suite and thus is widely available to most students.[1] Microsoft Excel is a spreadsheet tool that can be referred to as a number processor – in relative terms to word-processors. It is good for general data manipulation and analysis. However, while it can handle a certain depth of statistical analysis, it cannot do all that we require in this course. As well, there are some things that Excel can do, but they are more easily done in a statistical software package. Therefore, we have also recommended SPSS, a statistical program commonly used in real estate practice (in particular in property tax assessment).[2]

SPSS is also a number processor, but is more focussed on complex statistical manipulation. While Excel can carry out some statistical operations (and more with an add-in program called "SSC-Stat"), SPSS is better suited for multivariate analysis, its intended purpose.

In the illustrations in this lesson and throughout the course, we will generally show the steps in Excel and in SPSS. Where the capabilities of Excel stop, we will make it clear that we are progressing to the specialized statistical tools.

It is important to bear in mind that Excel and SPSS are only tools for data analysis. Like many tools, they can make work a lot easier, but they can be dangerous in the wrong hands! These tools will not make mistakes in calculation, incorrect assumptions, or draw incorrect conclusions. These are all mistakes made by the *user*. The applicable computer programming principle is called GIGO: "garbage in, garbage out". The software can only

[1] We have focused on Microsoft Excel because of its market dominance in spreadsheet applications. However, if you do not already have access to Excel, you may want to investigate a free, open-source alternative called Calc, which is included in the free OpenOffice suite. OpenOffice has much the same functionality as Microsoft Office (and is fully compatible, so editing a spreadsheet with both applications is fine). Those interested in reviewing this alternative may download the full office suite or the individual programs here: *www.openoffice.org*.

[2] We have focused on SPSS because of its market dominance in statistical software applications. An alternative program is NCSS, Number Cruncher Statistical Software. This does much the same calculations as SPSS, although with different commands and output. NCSS is less expensive than SPSS. For students who choose to use NCSS, **it is the student's responsibility** to reconcile the differences in commands. Supplements with alternative instructions are provided under "Online Readings" on the course website.

do the mathematical calculations it is programmed to do – it is up to the user to ensure that the problem is structured properly and entries made accurately.

In proceeding through the course and in real estate practice, it is vital to verify that your outcomes are what you expected to see and then afterwards double-check the results. Whether for a class project or a report to hand to your boss for a multi-million dollar project, the level of certainty and reliance you may place on your results is directly related to the time spent ensuring they are correct. If your process has many steps in the analysis, then each step should be checked carefully – if an error is discovered later, you may be able to isolate the problem and not have to re-check everything. The lessons in this course will offer illustrations of what you might reasonably expect to see within various data analysis processes. We will point out things to look for and common mistakes, with the hope that problems can be rectified before they happen.

NOTE FROM THE TUTOR

This lesson is an immersion into statistical analysis, which is a foreign area for most students. Like learning a language, you should not expect to get it all perfectly on the first pass through. Don't panic, that's normal! In our experience, the best way to learn this is to first follow along by rote, slowly and methodically carrying out each step as shown. Afterwards, you should re read the lesson a second time, and hopefully the meaning will become clearer. Keep in mind that later lessons will use these same techniques, so you will be able to continue practicing them as the course progresses. Students in past courses often report that by the time they are reading Lessons 7 and 8, they find much of this material has become considerably simpler than when they first encountered it – in other words, don't get discouraged if you are finding this difficult now!

Software and Database Orientation

Basic Operations

In order to complete this lesson, you will need to have the necessary software installed and be familiar with its basic operation. You will need Microsoft Excel and one of SPSS or NCSS. If you need help with the basic tasks in these programs, you should review the "Orientation" documents found on the "Course Resources" webpage, found on the "Online Readings" page under "Pre-Reading".

For example, in Excel, you should know how to:

- open, save, and close spreadsheet files
- navigate through data
- print all or portions of files
- create charts
- use basic arithmetic to calculate new values (+, -, *, and /). That is create a basic formula
- use functions such as SUM and AVERAGE to calculate new values in the spreadsheet

You may find it helpful to install the SSC-Stat Add-In and Data Analysis Add-In. These free Excel Data Analysis add-ins allow you to carry out much more comprehensive statistical analysis. Installation instructions are found in the Excel help documents found under Pre-Reading & Reference on the Online Readings page of the Course Resources webpage.

In SPSS, you should know how to:

- open, save, and close database files and output files
- navigate the "Data View" and "Variable View" tabs
- carry out basic transformations
- save and load syntax files

Please review the "Orientation" documents found on the "Course Resources" webpage if you need further assistance on these basic functions. The more comfortable you are with the software, the easier this lesson will be.

COMPUTER HELP

In an effort to assist students using different statistical software (or versions), we have provided several tools on your Course Resources webpage. These tools have been developed to assist students using different versions of both SPSS and NCSS.

What Tools Are Available:

- Orientation: Each software package and version has an Introduction or Orientation document, which explains how to install the program and a review of its basic features.
- Supplement: Each software package has various supplements which may include short videos. These supplements explain (or demonstrate) the operating procedures for the different software packages in use by students.

Where to Find the Tools:

These documents can be found under the "Computer Help" link in Online Readings. We have also posted these in a table on the main BUSI 344 Course Resources webpage.

How to Use the Tools:

You should have a printout of the appropriate orientation and supplement for your software (if you are unable to print these out, please contact our office and we will send you a printout). You should immediately go through your supplement and put a note in your workbook on all pages referenced, e.g., "see supplement", so you'll be sure not to miss important points. As you proceed through the 344 lessons, any time there are SPSS instructions provided, you should look at the supplement to see the equivalent instructions for your software package or any notes specific to your package.

How to Get More Help:

If you still have difficulties after using these supplements, you should contact your course tutor for further assistance. Use the Tutorial Assistance link on the Course Resources webpage for your tutor's contact information.

Database Introduction

For the remainder of this lesson, we will review the statistical and graphic capabilities of these programs using a small database of home sales from Burnaby, British Columbia. You may download this database from the "Online Readings" webpage, titled "Burnaby". Depending on the program, the files will have the format: .sav (SPSS) or .xlsx (Excel).

The database provides information on 134 home sales. It provides information on the characteristics of each provided through the following variables:

- NBHD – neighbourhood number; coded as 9, 10, 22, 24, 25, 28, 30, or 31.
- S_PRICE – sale price; all properties sold within a four month period.
- QUALITY – a linear variable measuring quality of construction; .55, .671, .808, .94, 1.0 or 1.2.
- EFFYRBLT – effective year built; measure of effective age ranging from 1933 to 2005.

- BEDROOMS – number of bedrooms.
- BATHS – number of bathrooms; 1.0, 1.5, 1.75, 2.0, 2.25, 2.5, etc.
- FAMROOM# – number of family rooms (living areas) in addition to primary living room.
- TOTAREA – total finished area; includes finished basement area.
- FINAREA – total finished area above basement level; does not include finished basement area.
- LOT_SQFT – square foot of lot size.
- TESTVAL – test values; an estimated value for each property, which will be tested against its actual sale price later in this lesson.
- ONE – set to 1 for all sales; a variable that will be used in the SPSS Ratio and NCSS Appraisal Ratios modules, explained in more detail later in this lesson.

Listing and Printing Data

In Excel, you may print the database by selecting the Print Area and selecting File → Print. If you press Print Preview, you will be able to first view the data to ensure it is what you wanted to print. You may also click on Setup in order to modify print options, such as Portrait or Landscape, or "Scale to Fit" – e.g., to fit entire report onto one page.

SPSS offers a number of options for printing the data, one of which is Case Summaries. You can select some or all of the data as well as a number of basic statistics.

- Click Analyze → Reports → Case Summaries, Brings up the Summarize Cases submenu.
- Select the six variables:[3] BATHS, BEDROOMS, EFFYRBLT, FAMROOM#, FINAREA, and LOT_SQFT.[4]
- Click on the top arrow, moving selected variables to Variables window.
- Ensure that "Display Cases" is checked, and ensure that none of the other boxes, such as "Limit cases to first", are checked.
- Click on Statistics, and ensure that only Number of Cases is selected.
- Click Continue → OK.

The results are displayed in the output file. The first block contains a summary of the data used in the report. The second block labelled "Case Summaries" contains the summary data for the first six variables. The first ten lines of this report are shown below.

Please note: The data displayed in your Case Summaries window may not be the same as that shown here. A feature of SPSS is that when certain statistical procedures are run, the database automatically re-sorts the data according to the order of a given variable. The Burnaby.sav database originally supplied to you may have been sorted slightly differently than the database used to create this table. If so, do not be concerned – the data in your database is exactly the same as the one used in the Course Workbook!

[3] Selecting a group of items such as variable names or files is a common Windows procedure and will generally not be explained in detail in this Workbook. To select a list of files, you should click on the top file, press and hold down the Shift key and then click on the bottom file. This should cause all of the files to be highlighted to indicate they have been selected.

[4] Variables may be shown in alphabetical order or in the order in which they were entered into the database file. Similarly, output may show either the variable names or labels. Both of these depend on what you have specified in the SPSS Options – see "Customizing SPSS" in the SPSS Orientation document on the course website for more information.

Case Summaries

	Bathrooms	Bedrooms	Effective Year Built	Family Room	Finished Area above the basement	Lot Size in square feet
1	1.00	2	1939	0	669	6,230
2	1.00	3	1945	0	870	3,861
3	1.00	2	1972	0	1,002	4,029
4	1.00	2	1969	0	827	6,950
5	1.00	3	1967	0	1,044	3,960
6	1.00	2	1945	1	1,151	4,043
7	2.00	2	1970	0	860	4,738
8	1.00	3	1977	1	1,114	4,125
9	2.00	3	1933	1	1,536	4,113
10	2.00	2	1969	0	930	7,150

The format of the data can be changed to include things like commas, dollar signs, and fewer or more decimal places by using the "Variable View" tab. This is important to remember: if you are not getting the correct number of decimal places displayed for a variable in SPSS – go to the Variable View tab and change the Decimals for that variable to an appropriate number.

The Case Summaries report can be used to display data on all cases or the first selected number of cases, and on any number of selected variables. The title of the report can be changed by clicking on the Options button on the Summarize Cases window. The report can be printed by clicking on Case Summaries in the left column of the output screen, and then clicking on the printer icon. This report is a quick method for getting a printout of the data you are analyzing.

Computing Basic Statistics

Descriptive Statistics

Descriptive statistics describe the distribution of a data set in terms of both central tendency and dispersion. As seen in Lesson 1, the three most common measures of central tendency are the median (or "middle" value), mean (or "average" value), and mode (most "common" value). Measures of dispersion include the standard deviation, coefficient of variation (standard deviation divided by the mean), coefficient of dispersion, and range.

In Excel, the following functions are used to calculate descriptive statistics:

- COUNT: how many are there?
- MAX: what is the maximum value?
- MIN: what is the minimum value?
- SUM: what is the sum of all the values?
- AVERAGE: this is the mean and in non-mathematical language gives the average value. It is the sum divided by the count.
- MODE: what is the most common value?
- MEDIAN: what is the middle value? (or average of the two middle values if there are an even number of them)
- STDEV: what is the sample standard deviation of the numbers?
- STDEVP: what is the population standard deviation of the numbers?

Each of these functions in Excel use the structure

=FUNCTION(list of range of cells)

If you wanted the total of the values in cells B2 through B15 in cell B17 you would type (in cell B17) the following:

=SUM(b2:b15)

The SSC-Stat Add-in calculates these statistics through an easy to use interface. Simply block the variables of interest (click on the column, e.g., "B" for S_PRICE), then select SSCstat → Analysis → Descriptive Statistics.

All of these statistics (and many more) can be easily generated in SPSS.

In SPSS use:

- Analyze → Descriptive Statistics → Descriptives… OR
- Analyze → Descriptive Statistics → Frequencies…

In both cases, the Options button will give you many choices for statistics.

Following is the SPSS report showing the mean, standard deviation, minimum, and maximum for all 12 variables in the Burnaby database (in alphabetical order), using the Descriptives module.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
BATHS	134	.75	4.50	1.7164	.81676
BEDROOMS	134	1	5	2.82	.692
EFFYRBLT	134	1933	2005	1975.75	17.162
FAMROOM#	134	0	3	.93	.758
FINAREA	134	574	3235	1324.10	527.068
LOT_SQFT	134	3861	12540	6540.92	2012.004
NBHD	134	9	31	20.64	8.322
ONE	134	1	1	1	0
QUALITY	134	.55	1.20	.9533	.12542
S_PRICE	134	135000	463000	243586.78	72465.610
TESTVAL	134	148800.00	459500.00	242900.0000	65049.62247
TOTAREA	134	660	4428	1897.37	774.543
Valid N (listwise)	134				

The following report shows the SPSS Frequencies module calculating statistics for S_PRICE (statistics selected include quartiles, mean, median, mode, standard deviation, range, minimum, maximum, and S.E. mean). Note that "Display Frequency Tables" was left unchecked (see Helpful Hint following the table):

S_PRICE

N	Valid	134
	Missing	0
Mean		243586.78
Std. Error of Mean		6260.08
Median		230000
Mode		210000
Std. Deviation		72465.61
Range		328000
Minimum		135000
Maximum		463000
Percentiles	25	186750
	50	230000
	75	278500

The median is the 50th percentile or midpoint in the distribution, with half of the sales prices less than it and half greater than it. The median is a frequently used measure of central tendency in both assessment and single property appraisals. This is because the mean can be significantly influenced by outliers; for S_PRICE, the influence of some very high priced sales has resulted in a mean that is larger than the median. The 25th percentile is the first quartile and the 75th percentile is the third quartile. These represent the cut-off points for the lowest one-fourth and lowest three-fourths of the data, respectively. Although not shown, the coefficient of variation (COV) can be computed by dividing the standard deviation by the mean. Calculation of the COV and the coefficient of dispersion (COD) will be discussed in the next section.

As shown above, SPSS provides a full range of statistics within its analysis menu. Excel on the other hand does not. A few statistics depend on the results of the other Excel functions such as the Range which is the difference between the MAX and the MIN, or the Coefficient of Variation which is the STDEV divided by the AVERAGE times 100. However, these can be calculated in SSC-Stat using Analysis → Descriptive Statistics, and clicking on "Additional Statistics".

Helpful Hint!

SPSS: the Frequency Tables are highly useful to analyze variables with limited values (termed "discrete" variables), such as BEDROOMS and BATHS, showing how many observations are within each group. For these variable types, you may wish to leave the frequency distribution tables box checked.

Excel: the FREQUENCY function is an array function, where you specify the data you want arrayed and then the cells you want them summarized into (specifying the number of boxes). In SSC-Stat, you can use Analysis → Summary Statistics for a similar function, specifying the variable of interest in both the "Variable" and "Factor By" boxes.

Helpful Hint!

SPSS: You can adjust the look of output tables produced by SPSS in a variety of ways. By double clicking on a table you will open the table formatting mode. Column widths can be changed by positioning your cursor on a vertical line and dragging the line left (to narrow the column) or right (to widen the column). You will want to widen columns if the values appear as ****** in your output. You can also experiment with other formatting options such as removing unwanted rows (click on the row, use Edit → Select → Data and Label Cells, then Edit → Clear) or adding or removing decimal places (select the cell or cells, click Format → Cell Properties...→ Under the Format Value tab, change the number of decimal places). Clicking outside of the table will exit the table formatting mode.

Standard Error and Confidence Intervals for the Mean and Median

After finding the mean for a sample, it is a good idea to ask whether the figure is representative of the whole population. This involves calculating something called the standard error of the mean,[5] then determining confidence intervals around that number.

The standard error of the mean, often referred to as standard error or SE, is a measure of how well the mean for a particular sample estimates the mean for the whole population. Probably the most accurate technique that could be used to estimate the standard error of the mean (although prohibitively cumbersome) would involve

[5] Because *standard deviation* and *standard error* sound similar and are often confused, it is important that you keep them separate in your mind. Standard deviation is an expression of how individual data points are distributed around the mean of those data points, while standard error is an expression of how a group of means are distributed around the mean of those means.

taking hundreds of samples, calculating the mean of each, finding the mean of all the means, and then calculating the standard deviation of those means (how the individual means are distributed around the mean of the means).[6] Fortunately, there is an easier method of estimating whether the first mean you calculated was representative of the whole population, involving the following formula:

$$\overline{SE}_x = \frac{s}{\sqrt{n}}$$

Where:

$\overline{SE}_x$ = Estimate of the standard error of the mean[7]
s = standard deviation of the sample
$\sqrt{n}$ = the square root of the sample size

The standard error of the mean can be used to form confidence intervals around the mean. Adding and subtracting one standard error to and from the mean produces a range of values that typically encompasses approximately 68% of the possible means for the population overall. In other words, if one were to calculate the means of 100 samples, around 68 of them would fall into the range defined by adding and subtracting one standard error from one of the means (see Figure 2.1 in following pages for an illustration of this concept). Adding and subtracting two standard errors typically encompasses approximately 95% of the possible means; this "plus or minus" two standard errors is referred to as a 95% confidence interval, which is an important statistical concept that will be discussed further.

In SPSS, exact confidence intervals for the mean can be found in Analyze → Descriptive Statistics → Explore:

- Analyze → Descriptive Statistics → Explore.
- Click S_PRICE then click on the top arrow to move S_PRICE to the Dependent List.
- Under Display click Both, which ensures that statistics and histogram plots will be produced.
- Click the Plots… button.
- Under Boxplots click None.
- Under Descriptive Statistics select Histogram, unselect Stem-and-leaf.
- Select Normality plots with tests.
- Continue → OK.

The descriptive statistics should appear as follows (note that the Case Processing Summary has been omitted below):

Descriptives

			Statistic	Std. Error
S_PRICE	Mean		243586.78	6260.075
	95% Confidence Interval for Mean	Lower Bound	231204.59	
		Upper Bound	255968.96	
	5% Trimmed Mean		238282.57	
	Median		230000.00	
	Variance		5.3E+09	
	Std. Deviation		72465.610	
	Minimum		135000	
	Maximum		463000	
	Range		328000	
	Interquartile Range		91750.00	
	Skewness		1.055	.209
	Kurtosis		.846	.416

[6] Streiner, David L. 1996. "Maintaining Standards: Differences between the Standard Deviation and Standard Error, and When to Use Each". *Canadian Journal of Psychiatry*.

[7] This is an estimate of the standard deviation of the mean, rather than the *true* standard deviation of the mean, because we have to use the *sample* standard deviation in the numerator rather than the *population* standard deviation. In most cases, the population standard deviation (the true standard deviation) is unknowable because of practical constraints in data collection.

Note that the 95% confidence interval for the mean sale price is 231205 to 255969. Thus, in the example above, the Mean is 243586.78 with a 95% confidence interval of 231205 to 255969. That is, given a sample size of 134, we can be 95% confident that the true mean value of the whole population of residential properties falls between $231,205 and $255,969. Or alternatively, we can conclude that 95% of samples (or 19 out of 20) drawn from this population would have a mean in this range. As will be noted later, confidence intervals are very important in making decisions related to sales samples. However, keep in mind that this procedure above has calculated the confidence intervals for the mean and not the median.[8]

Helpful Hint!

In the Analyze → Descriptive Statistics → Explore module, you can vary the confidence interval of the mean by clicking on the Statistics... button and changing the Confidence Interval value from the 95% default value to any other of interest, e.g., 90% or 99%.

As stated above, the 95% confidence interval for the mean is closely approximated by adding and subtracting the value of two standard errors of the mean from the mean value. The mean will always be in the centre of the range of the 95% confidence interval, but this is not true for the median. While the median value will always be within its 95% confidence interval, the median will not always be at the centre of the confidence interval range. This is because the confidence interval for the median is found by counting values greater than and less than the median rather than by addition and subtraction of a value.

The confidence interval for the median of a sample can be important in inferential statistical analysis.

Ratio Analysis

The SPSS Ratio module was developed specifically for real estate analysis. Its main function is to calculate the assessment-sales ratio (ASR), which is a very useful statistic in property assessment or in any mass appraisal application (where market values are estimated and need to be compared to actual sale prices). The ASR shows how accurately assessed values relate to actual sale prices by dividing the assessed value (the value predicted by the model) by the actual sales price. The Ratio module in SPSS produces numerous helpful statistics including the confidence interval for the median, the coefficient of dispersion (COD), and the coefficient of variation (COV).

Because the Ratio module is designed to first create an ASR and then produce statistics for it, we will illustrate its use for analyzing an ASR for the TESTVAL and S_PRICE variables. In SPSS:

- Analyze → Descriptive Statistics → Ratio…
- Select TESTVAL as numerator and S_PRICE as denominator.
- Click on Statistics and select Median, Mean, Weighted Mean, Confidence Intervals (95%), COD, Mean centred COV, Range, Minimum, Maximum.
- Continue → OK.

[8] Confidence interval calculation and use are described in more detail in the excerpts from *Advanced Computer-Assisted Mass Appraisal*, found under "Online Readings" on the "Course Resources" webpage. Confidence intervals for the mean are described in Chapter 1 in the "Sample Theory" and "Statistical Inference" sections; confidence intervals for the median are described in Chapter 12.

The report will appear as shown below:

Ratio Statistics for TESTVAL / S_PRICE

Mean		1.013
95% Confidence Interval for Mean	Lower Bound	.991
	Upper Bound	1.036
Median		.999
95% Confidence Interval for Median	Lower Bound	.967
	Upper Bound	1.022
	Actual Coverage	95.3%
Weighted Mean		.997
Minimum		.712
Maximum		1.350
Range		.638
Coefficient of Dispersion		.102
Coefficient of Variation	Mean Centred	12.9%

The Ratio module first creates assessment-sales ratios (ASRs) for each row and then analyzes these results statistically. For example, a property assessed at $155,500 which sold for $157,000, would result in an ASR of 0.9904 or 99.04%[9] (155,500 ÷ 157,000). ASRs are used to test how well assessed values reflect market value with 1.00 (or 100%) meaning that assessed value is exactly equal to sale price; in this one example, an ASR of 99.04% is very close to perfect.

The median and mean ASR both appear close to 1.00 (100%). However, the range is quite high, which implies that there is a wide variance in ASRs, meaning that some properties are being assessed too high or too low. However, the lower and upper confidence limits of both the mean and median overlap 1.00, which means that there is less than 5% probability that the mean/median ASR is not equal to 1.00 and thus we can be 95% confident that assessed values are statistically equal to market values for the properties in this database. Finally, the coefficient of dispersion (COD) of 10.2% is reasonably low (below 15%). This shows the assessed values in TESTVAL are of acceptable quality. The analysis of ASRs to test assessments will be covered in much more detail in later lessons.

The Ratio and Appraisal Ratio modules were designed to analyze assessment-sales ratios (ASRs), which are expressed as percentages, but they can be used to analyze non-ratio or non-percent variables as well. Before Ratio can be used to compute statistics for non-ratio variables, you need a variable equal to one for all observations (called "ONE" in the Burnaby database).[10] The variable ONE is used as the denominator which means that the "ratio" being tested is just the variable divided by one (1) or simply the variable itself.

Following are the instructions for using the SPSS Ratio module for non-ratio variables:

- Analyze → Descriptive Statistics → Ratio...
- Select S_PRICE as numerator (variable under analysis).
- Select ONE as denominator (divides S_PRICE by 1, allowing Ratio to analyze this single variable, rather than creating a ratio).

[9] The Ratio Statistics module in SPSS computes ratios in decimal format. Readers will note that a decimal ratio of 0.9904 is equivalent to a percentage ratio of 99.04%. The decimal and percentage notation for ratios will be used interchangeably throughout this workbook.

[10] To use Ratio for non-ratio variables, the One variable will have to be created in the database. The following transformation can be used:
Select Transform → Compute Variable (transformations are explained in more detail later in this lesson).
Enter One as Target Variable.
Enter 1 in Numeric Expression; this creates a variable named "one" to be equal to 1 for all observations.
OK, runs transformation.

- Click on Statistics and select Median, Mean, COD, Mean centred COV, Range, Minimum, Maximum, and Confidence Intervals.
- Continue → OK.

By using the procedure above, Ratio can be used to calculate statistics for any variable, not just ratio variables. Note, however, that when using this method the weighted mean will equal the mean and the PRD will be equal to 1.000. It is important to note that if you do have a ratio you are interested in testing like assessed value divided by sale price, you do not need a ONE variable.

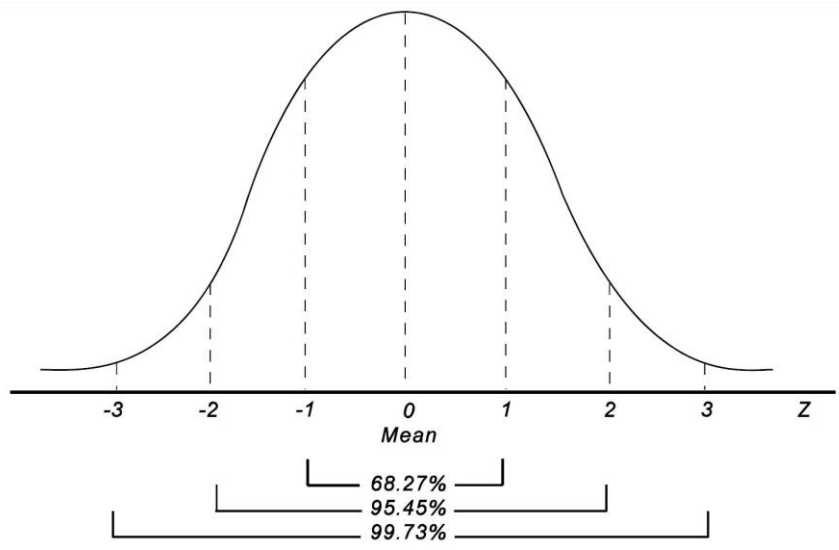
In Excel, calculating some of these advanced statistics is possible, using regular "Functions" or those in SSC-Stat. Others may require a little more work, manually creating a formula to calculate the statistic.

Tests for Normality

A "normal" distribution means that the data is evenly spread out on either side of the mean, with the bulk of the observations near the mean and trailing off on either side. This forms the commonly described "bell curve", sought after in course grades and other applications. If the data is bulked below the mean, but with a few high outliers, then the data is said to be skewed right (and the median will be lower than the mean). Alternatively, if the data is bulked above the mean, but with a few low outliers, then the data is said to be skewed to the left (and the median will be higher than the mean). Figures 1.8, 1.9, and 1.10 in Lesson 1 illustrate the normal curve and skewed curves.

A normal distribution is required in order to accurately estimate confidence intervals of the mean and to carry out probability estimates for sample data: e.g., "68% of the data will fall between the points ___ and ____, which are one standard deviation on either side of the mean". See Figure 2.1 for an illustration. However, if the data is not normal, these measures will not be completely reliable. Quite often, real estate data is not normally distributed, especially residential sales data, as it tends to have a large number of lower-priced sales and then a few high priced sales that skew the curve. When this is the case, the median is a better indicator of central tendency than the mean and the confidence intervals of the median are preferable to those around the mean.

Figure 2.1: Standard Deviation Ranges for a Normal Distribution



In SPSS, the Explore module shown earlier (Analyze → Descriptive Statistics → Explore...) calculates two tests for normality:

- the Kolmogorov-Smirnov (K-S) and
- Shapiro-Wilk (S-W) tests (under the Plots… button, select Normality Plots with Tests).

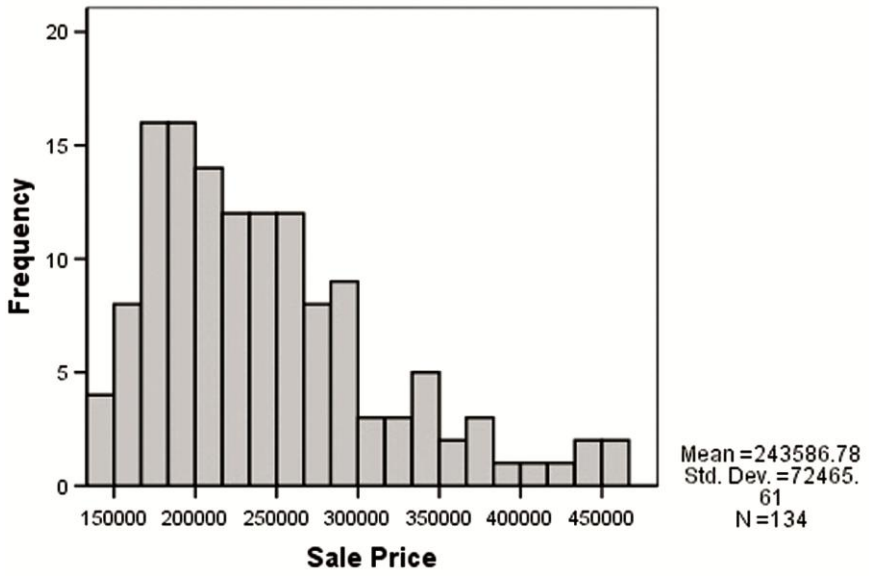
The statistics below show that the significance (Sig) is less than .05, indicating that the distribution for S_PRICE fails this test for normality. The K-S and S-W tests tend to place less weight on distributions which have a greater number of observations near the median than would be expected in a normal distribution. For most real estate applications, we want to have a concentration of values close to the median, so this type of test is well-suited here.

Tests of Normality

	Kolmogorov-Smirnov(a)			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
S_PRICE	.097	134	.004	.921	134	.000

a Lilliefors Significance Correction

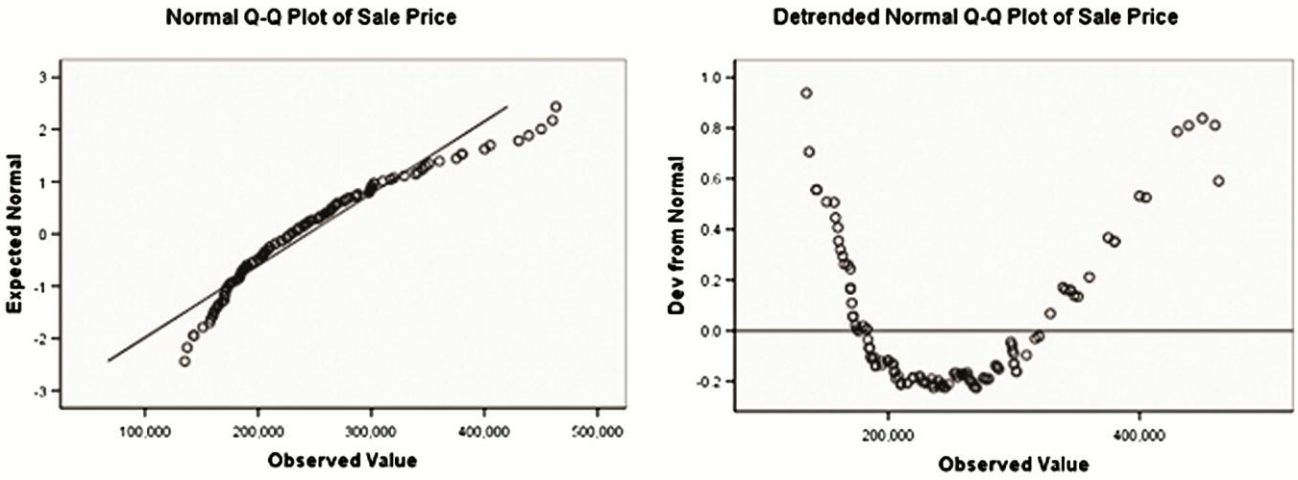
The SPSS output also shows a frequency graph of the data, known as a histogram (under the Plots… button select Histogram). The histogram below shows how the data for S_PRICE is distributed. You can see that the sales are highly concentrated towards the lower values with a wide scattering of sales occurring above the 300000 value.



Helpful Hint!

You can edit a chart by double clicking on the chart with the left mouse button. This puts the chart in the SPSS chart editor which has various editing options, such as altering the axes or bars, and so on. For example, to add a title, after entering the editor mode, click Options → Titles.

The SPSS output from Explore also includes a normal probability plot, which should appear as follows (SPSS also includes a second "detrended" plot):



The first plot shows the data represented by the small circles, compared to the expected 45 degree line that would exist if the data were normally distributed. The second plot changes the expected representation to a horizontal line. In both cases it is clear that there is considerable deviation from the expected "normal" result at both the high and low ends. There is also some moderate deviation in the centre portion of the curve. In this case, the distribution does not appear normal, as it is asymmetrical, truncated on the left, and skewed to the right.

Excel does not offer these normality statistics, but SSC-Stat provides a normal probability plot under SSCstat → Visualization → Normal Probability Plot – note that the axes are reversed though, so the curve is under the fit line. As well, a histogram can be created by first using the FREQUENCY function to create the desired categories and then use the Insert → Column in the Chart Group to create a histogram for these frequencies.

Break (Group) Variables

It is often helpful to calculate descriptive statistics for a variable broken down into various subgroups of properties, such as age groups or neighbourhoods.

In SPSS, you can do this in at least two ways. The first offers a quick comparison of means and standard deviations by a break variable. We will examine S_PRICE using NBHD as a break variable:

- Analyze → Compare Means → Means…
- Click S_PRICE, top arrow (Enter S_PRICE into the Dependent list).
- Click NBHD, bottom arrow (Enter NBHD into the Independent list).
- Click on Options…
- Select Median, Minimum, Maximum to add these statistics to the default list.
- Continue → OK to produce Means report.

The results should appear as follows:

S_PRICE

NBHD	Mean	N	Std. Deviation	Median	Minimum	Maximum
9	221409.09	22	57956.40	222500.00	143000	360000
10	255991.25	24	73539.11	225000.00	172000	430000
22	189980.00	5	14611.71	190000.00	169900	210000
24	261212.52	23	78707.62	235000.00	172000	463000
25	272128.57	14	53556.45	277000.00	168000	348000
28	227431.43	35	80198.85	194500.00	135000	450000
30	238600.00	5	31816.66	243000.00	200000	280000
31	284191.67	6	89163.56	306000.00	137150	380000
Total	243586.78	134	72465.61	230000.00	135000	463000

These results show that there are very few sales in NBHDs 22, 30, and 31. Of the remaining five neighbourhoods, NBHD 25 has the lowest dispersion of sale prices suggesting a homogeneous group of homes in that neighbourhood. It also has the highest mean and median sale prices, although the range is quite wide, at 180,000. NBHD 28 has the most dispersed range of prices, the largest standard deviation, and lowest median price. The mean is considerably greater than the median showing the influence of several very high valued sales, which do not seem consistent with the remainder of the neighbourhood.

Helpful Hint!

Note that you can specify multiple dependent or independent (break/group) variables. Also, the break variables may be layered, for example, you could compute average sales prices by quality class within neighbourhood.

In Excel, you could calculate similar output in SSC-Stat by specifying a "By Factor" variable in either of the Descriptive or Summary Statistics modules.

The second method in SPSS for calculating descriptive statistics with break variables is to use the Split File facility located under Data on the main menu. Split File divides the data file into subsets based on a break variable (or variables). Whatever statistics or charts are produced are repeated for each group. The Split File option stays in effect until it is manually removed. By using this method, you can produce more detailed and varied output than by using the previous procedure. For example, you could break down S_PRICE by NBHD, but also produce outputs such as histograms and probability analysis. In some detailed analyses, histograms can be useful to show data concentrations in groups which assist in the determination of errors or the need for additional adjustments. However, in most cases, histograms are not necessary.

- Data → Split File.
- Click "Organize output by groups".
- Click on NBHD, arrow, to place NBHD into the "Groups Based on" box.
- Click "Sort the file by grouping variables".
- Click OK. The "Split NBHD" message should appear in the lower right hand corner of your Data Editor window. The data are sorted by NBHD.
- Analyze → Descriptive Statistics → Explore.
- Click S_PRICE, top arrow, to move S_PRICE to the Dependent List.
- Click on Plots button, confirm that Boxplots shows "None" and Histograms and Normality plots are checked.
- Continue → OK.

Results for the first neighbourhood are displayed below.

Descriptives [a]

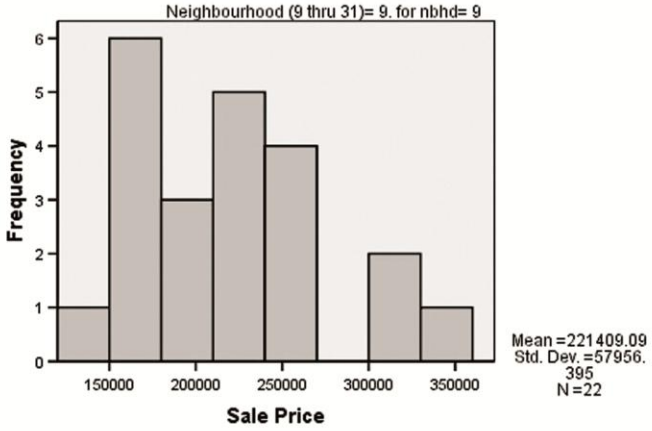
	NBHD			Statistic	Std. Error
S_PRICE	9	Mean		221409.09	12356.345
		95% Confidence Interval for Mean	Lower Bound	195712.66	
			Upper Bound	247105.52	
		5% Trimmed Mean		218227.27	
		Median		222500.00	
		Variance		3.4E+09	
		Std. Deviation		57956.40	
		Minimum		143000	
		Maximum		360000	
		Range		217000	
		Interquartile Range		87500.00	
		Skewness		.713	.491
		Kurtosis		.054	.953

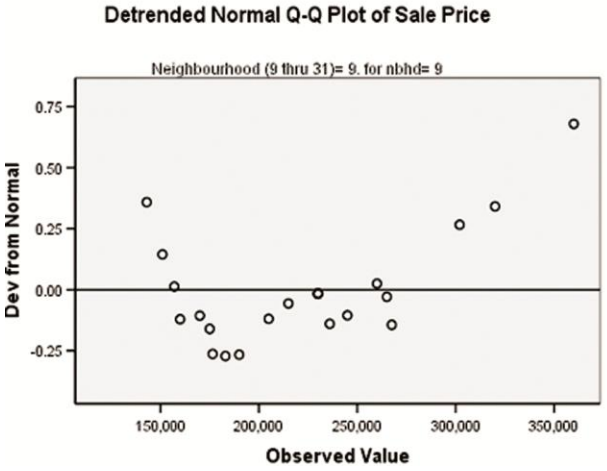
a NBHD = 9

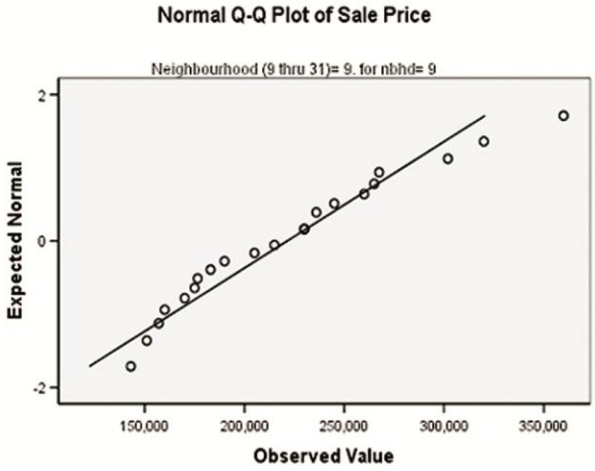
Tests of Normality [b]

		Kolmogorov-Smirnov [a]			Shapiro-Wilk		
	NBHD	Statistic	df	Sig.	Statistic	df	Sig.
S_PRICE	9	.115	22	.200*	.945	22	.254

* This is a lower bound of the true significance.
a Lilliefors Significance Correction
b NBHD = 9







You can then proceed to examine each neighbourhood in turn and compare their results. Before proceeding to further sections, ensure you turn off Split File by returning to Data → Split File, and selecting Analyze all cases, do not create groups, then click OK.

In Excel, you could manually carry out a similar method by first sorting the database according to the break variable of interest (Data → Sort), and then running the desired statistics only on the appropriate ranges (e.g., only for the entries where NBHD = 9), using either the Function commands or SSC-Stat.

Transformations

Transformations allow you to use mathematical and logical operations to create new variables from existing ones. For example, a database may include information on sale price and house size (square footage). If you were interested in creating a variable that represented sale price per square foot, you would create a transformation for such a variable.

In Excel, transformations may be carried out manually using Functions and Formulas. For example, you could create a new column that divides S_PRICE by TOTAREA for all rows of data. Commonly used transformations include the standard arithmetic operations of addition (+), subtraction (-), multiplication (*), division (/), and exponentiation (^). Other mathematical functions are shown in the Functions menu.

The syntax of transformations is often critical to obtaining a correct calculation. In other words, the correct placement of commas, semicolons, and brackets as well as the order of the operations can be crucial for achieving meaningful results. The software will give an error message when some aspects of the transformation are incorrect, for example, if there is an unrecognized variable name in the formula or if there is not the same number of right and left brackets. Perhaps even more critically, it is very easy to create a transformation that will be accepted by the software as mathematically correct, but which will not give the desired result. You should form a habit of always manually checking the resulting variables to ensure your transformation was correct.

SPSS has a separate module specifically created for carrying out transformations, which you can find on the main menu bar. You can view some of the transformations that will commonly be used in this course under Transform → Compute Variable.

The Compute Variable window is where you will create most of the new variables needed for carrying out mass appraisal model building. The mathematical transformations are similar to those in Excel – one notable difference is that exponentiation is denoted with double asterisks [**], not the caret [^] as in Excel. Additional functions are listed in the Function window. A frequently used transformation is RECODE, which converts an existing variable into a new variable.

SPSS transformations can be saved into syntax files, allowing them to be saved and then reviewed, re-run, or modified as needed. Syntax files will be explained in more detail later in the course.

Simple Mathematic Formulae

Suppose you wanted to examine sale price per square foot. You could use a transformation to create a new variable, sale price per square foot, by dividing the S_PRICE variable by another variable which represents the square footage of the building. This database contains two building size variables, total finished area including the basement (TOTAREA) and finished area above the basement (FINAREA). As the real estate market in most parts of the country now uses total finished area when advertising residential property for sale, we will use TOTAREA for this calculation.

In SPSS, we will create the new variable, SP_SQFT, as follows:

- Transform → Compute Variable.
- Enter SP_SQFT (new variable name) as the Target Variable.
- Click into the Numeric Expression box and enter S_PRICE/TOTAREA
- OK, creates and adds the new variable to the data file.

Helpful Hint!

In SPSS, you may type in the transformation formula in Numeric Expression box or you may instead complete this procedure by pointing and clicking the mouse. This saves time and avoids the possibility of typing errors (which result in an error message in SPSS because the variable names in the Numeric Expression box must be typed exactly right). Also note that you can enter a label (e.g., SALE PRICE PER SQUARE FOOT) by pressing Type&Label... when creating the variable.

To view the new variable, you can scroll to the far right on the Data Editor screen (or you could list these three variables using Analyze → Reports → Case Summaries). You can check the transformation by reviewing the new variable and the variables from which it was created. You should use a calculator or manually calculate the value of SP_SQFT for the first few cases to ensure the transformation achieved the desired result.

You may now examine the summary statistics for this new variable. SPSS Descriptives shows the following results:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
SP_SQFT	134	62.97	259.98	137.7963	35.4684
Valid N (listwise)	134				

Helpful Hint: Syntax Files

A syntax file is a list of a SPSS commands. This can include transformations, regression commands, or pretty much any command SPSS offers.

There are two main methods of creating syntax files for transformations. The first is after creating a transformation, you can save it directly to a syntax file. Doing this for a group of transformations allows you to easily run them again later, should you need to re-create the new variables in a new database or in restoring a backup database (e.g., if you need to start over from scratch).

We will do this for the SP_SQFT transformation now:

- Transform → Compute Variable...
- Enter SP_SQFT (new variable name) as the Target Variable.
- Click into the Numeric Expression box and enter S_PRICE/TOTAREA
- Click the Paste button – this opens a new syntax file window with the SPSS instructions to perform the SP_SQFT transformation.

You can now save this file (File → Save) and then keep it open so that any new transformations you create and Paste in will be added to this file. The syntax file now shows:

```
COMPUTE SP_SQFT = S_PRICE/TOTAREA.
EXECUTE.
```

continues on next page

When you are ready to run any portion of the syntax file, select the lines you wish to run using your mouse (be sure to select the final period), and click Run → Selection or press the Play arrow icon in the tool bar. The advantage of this, over the transform module, is that the whole file can be saved for later as a record of what transformations were carried out on the database. Syntax files are saved with a file extension of .SPS.

When you are completing your projects, you will find syntax files to be very helpful.

This second method of creating syntax files for transformations is by typing them directly into the file. Say, for example, you wanted to create an "Age" variable, by subtracting effective year built from 2007 (please note: this is not needed for this lesson and is provided for illustration only).

If you do not yet have a syntax file open, select File → New → Syntax. In the syntax window, in the first available line, type the following, exactly as shown, including the period at the end. Exact "syntax" is crucial in these commands: one missing comma, bracket, or period will render them useless.

```
COMPUTE Age = 2007 – Effyrblt.
```

Using your mouse, select the entire command and click the Play icon, or select Run → Selection. Now return to the Data View window and you will see an Age variable has been created, but with no values. Select Transform → Run Pending Transformations, and you will now see values in the Age variable. You should check one or two manually to ensure they are correctly calculating 2007 less the effective year built.

Note that you could avoid the "Run Pending Transformations" step by specifying the transformation as follows:

```
COMPUTE Age = 2007 – Effyrblt.
EXECUTE.
```

Then select and run the two lines – the transformation values will be calculated. Notice that the transformation syntax created using the Paste method had the EXECUTE. command line included automatically.

The advantage of using syntax files for transformations is that you have a record of all transformations carried out. This can help you later in reviewing transformations made. Also, if you ever have a data problem or lose your data (e.g., computer theft, hard drive crash, fire), you can quickly and easily restore your data from the original database.

If you wish to save your syntax file, you may select File → Save As and name it something you will remember, e.g., "Lesson2transformations.SPS".

Recode

Apart from simple mathematical calculations, the Recode transformation is the most used transformation in this course. Recoding allows you to change the value of one or more variables according to specific rules which you supply. However, in all of these lessons, we will limit our transformations to recoding only one variable at a time. In this example, you want to recode the effective year built (EFFYRBLT) into five different classes so that you can examine the relationship between sale price and these age groups. The age groups will be as follows:

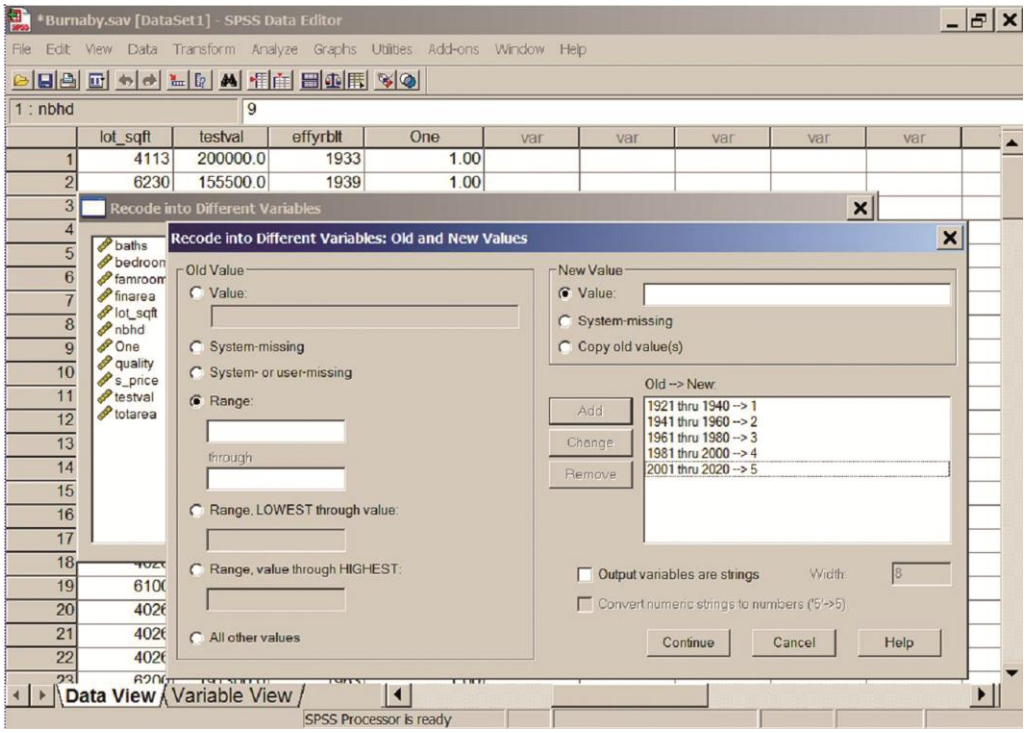
1921 - 1940 = Class 1;	lowest through 1940 = 1
1941 - 1960 = Class 2;	1941 through 1960 = 2
1961 - 1980 = Class 3;	1961 through 1980 = 3
1981 - 2000 = Class 4; and	1981 through 2000 = 4
2001 and above = Class 5	2001 through highest = 5

The new variable will be named YRCLASS.

In SPSS, YRCLASS is created as follows:

- Transform → Recode Into Different Variables.
- Enter EFFYRBLT as the Input Variable.
- Type YRCLASS into Name under Output Variable box.
- Click Old and New Values.
- Click the Range button and enter 1921 and 1940.
- Enter 1 into the Value box below New Value and click the Add button.
- Enter remaining ranges:
  1941 → 1960 → 2 → Add
  1961 → 1980 → 3 → Add
  1981 → 2000 → 4 → Add
  2001 → 2020 → 5 → Add (the RECODE window should appear as shown in the following screen shot)
- Continue → Change → OK.

This recode transformation has created a new variable, YRCLASS, with five values depending on the value range. The variable has been added to end of the data file and can be used in subsequent analyses. You can view the resulting variable by scrolling to the far right on the Data Editor window.



Helpful Hint!

This is an example of where you may wish to save your transformations into a syntax file. This allows you to easily run them again later, should you need to re-create new variables in a new database or in restoring a backup database. To capture the YRCLASS transformation, click the Dialog Recall button in the tool bar (fourth icon from left in main SPSS window, sixth from the left in output window) and select Recode into Different Variables. The transformation you just completed will be displayed. Click the Paste button in the transformation window and it will paste the transformation into a syntax file.

In Excel, this variable can be created manually by first sorting the database by EFFYRBLT (column K) and then typing in the YRCLASS numbers in the next available column. Alternatively, you could use the following function: =IF(K2>2000,"5",IF(K2>1980,"4", IF(K2>1960,"3",IF(K2>1940,"2","1")))) with this copied and pasted into each row for the new column.

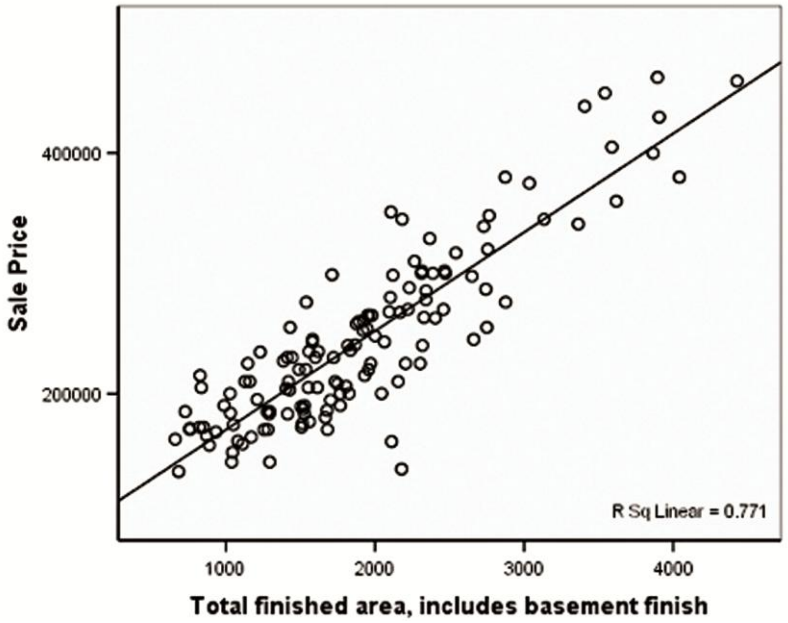
Graphics

Scatterplots

The scatterplot is a highly useful tool for showing the relationship between two quantitative variables, such as living area and sale price. A line of best fit or *regression line* can be displayed on the graph, as well as several other fit lines.

We will illustrate the SPSS procedure to produce a scatterplot of S_PRICE (dependent variable) with TOTAREA (independent variable) and add a regression line to the display:

- Graphs → Legacy Dialogs → Scatter/Dot → Simple Scatter → Define.
- Enter S_PRICE as the Y Axis variable.
- Enter TOTAREA as the X Axis variable.
- OK, displays the scatterplot.
- Double click on graph (allows you to enter the Chart Editor mode).
- Elements → Fit Line at Total – displays regression line and R-square value for regression line (ensure Linear is selected).
- Close "Properties" window, close the Chart Editor, and SPSS displays the following graph:



The RSq Linear (Rsq) is named the coefficient of determination. It measures the degree to which the dependent variable (sale price) is explained by the independent variable (total area). Rsq can take on values from 0 to 1. The closer to 1, the better the explanatory power. The result here, .771, is a high R-square, indicating that 77% of the variation in sale price can be explained by the size of living area. R-square will be explained in more detail in Lesson 6.

Examining the scattering of data in the plot, you can clearly see that sales prices generally increase with total living area, although there are some wide departures from the regression line.

Helpful Hint!

While still in the chart editor, carry out the following:

- File → Save Chart Template.
- Check Optional Lines and Fit Lines → Continue.
- Browse to an appropriate folder (this could be My Documents or perhaps you have created one for any files related to this course called, C:\Program Files\SPSS\UBC).
- Enter the file name RSQ1 for this template and save it.

This will allow you later to recall the "RSQ1.sgt" template on the scatterplot input screen and produce a chart with the regression line and R^2 value, without needing to use the chart editor.

In Excel, a similar scatterplot is shown as follows:

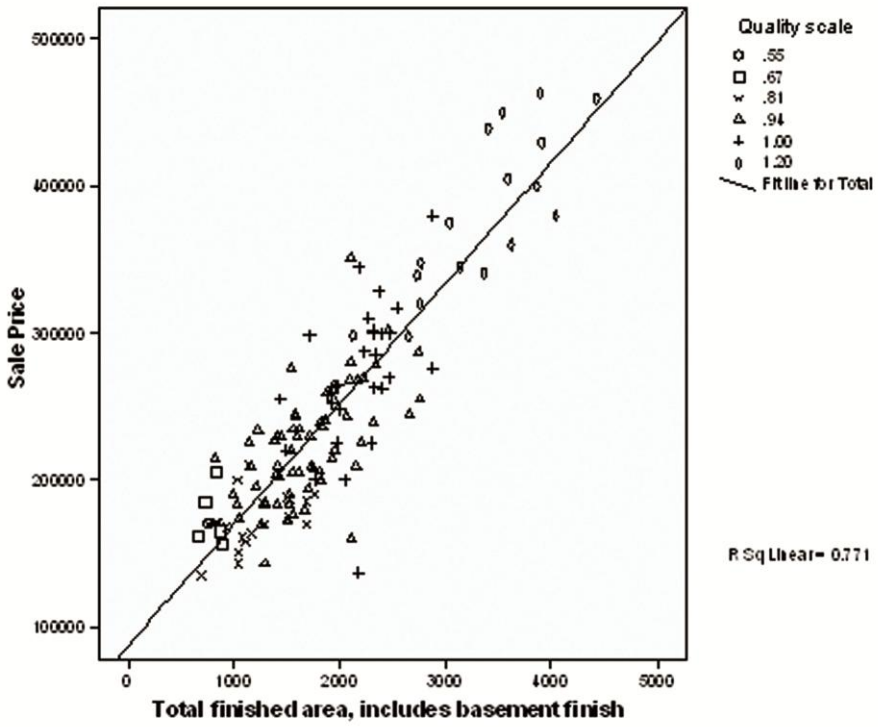
- Highlight the S_PRICE column by clicking on the "B" column header. Then, highlight the TOTAREA column by holding down the CTRL key and clicking on the "G" column header. The two columns should now be highlighted
- On the Insert tab, select Scatter in the Charts group
- Select the scatterplot sample that has no lines on it and the scatterplot will be super-imposed over the worksheet (you can use drag and drop to place the chart wherever you want it on the worksheet)
- You should notice that column B (sale price) is graphed on the X-axis and column G (total area) is graphed on the Y-axis. This is the default, since column B is to the left of column G on the spreadsheet. However, since column G is actually the independent variable and column B dependent, the order should be switched. To do this, Click Select Data in the Data group, then under Legend Entries (Series), Click Edit
- Change the Series name box so that it reads "=Burnaby!$B$1"; change the Series X values box so that it reads "=Burnaby!$G$2:$G$135" and the Series Y values box so that it reads "=Burnaby!$B$2:$B$135" (all without the quotes). Click OK, OK
- To change the chart title or X-axis or Y-axis labels, click on the Layout tab and in the Labels group use the appropriate Label or Title choice. When the text box appears on the chart, double click inside the text box to edit the text
- To add a trendline and get the R-square value, click Trendline in the Analysis Group and Click More Trendline Options… . Select Linear, and click the tick box beside Display R-squared value on chart. Click Close
- A trendline and its R-square value should appear on the graph.

In SSC-Stat:

- Block the variables of interest (click B column and then, while holding CTRL key down, click on column G; both S_PRICE and TOTAREA columns should be highlighted.
- Select SSCstat → Visualization → X-Y Scatter Plot.
- Select S_PRICE as "Y Variable" and TOTAREA as "X Variable".
- Under "Chart Type", select "Data Points Only".
- Click "Show Trend Line" → OK, to run the scatterplot. SSC-Stat does not provide the R-square statistic.

Moving on, it may also be informative to re-run the scatterplot using an optional marker for a third variable. For example, what is the relationship between S_PRICE, TOTAREA, and QUALITY? In SPSS:

- Graphs → Legacy Dialogs → Scatter/Dot → Simple Scatter → Define.
- Enter S_PRICE as the Y (dependent) variable.
- Enter TOTAREA as the X (independent) variable.
- Enter QUALITY into "Set Markers By".
- Select "Use chart specifications from:" and browse to the "RSQ1" template saved earlier.
- OK, produces the following graph:[11]



Helpful Hint!

You can shortcut the first three steps above by clicking the Dialog Recall button on the tool bar (fourth icon from left in main SPSS window, sixth from the left in output window) and select Simple Scatterplot. Then simply enter QUALITY as the "Set Markers By" variable.

The graph is reproduced with the data points of different shapes by quality class. It shows that higher-quality homes are associated with higher sales price and larger total areas. Lower quality homes tend to be smaller and sell for less. Note also that the chart above displays the total fit line and R-square value, as a result of using the "RSQ1" chart template.

In Excel SSC-Stat, similar output can be obtained by specifying Quality as the "By Factor" variable.

[11] The graph created by SPSS defaults to differentiate quality by colours. The graph shown in this lesson was created by leaving the output file back into Burnaby.sav, selecting Edit → Options → Charts, and under "Style Cycle Preferences" selecting "Cycle through patterns only". This will default SPSS to use patterns instead of colours. If you have SPSS options set to show colour output, it is difficult to identify the third variable (quality or neighbourhoods). When printing in black and white or submitting assignments by facsimile, the marker will not be able to see some important information in your printouts.

You should examine several combinations of variables with scatterplots to see which variables are closely related to S_PRICE and which are not (ensure that you edit the chart to include R-square statistic, so that the degree of explanation of S_PRICE is shown). You should also look for relationships between the independent variables themselves. Besides QUALITY, you may want to use NBHD and YRCLASS as marker variables to see if relationships vary by neighbourhood and age group.

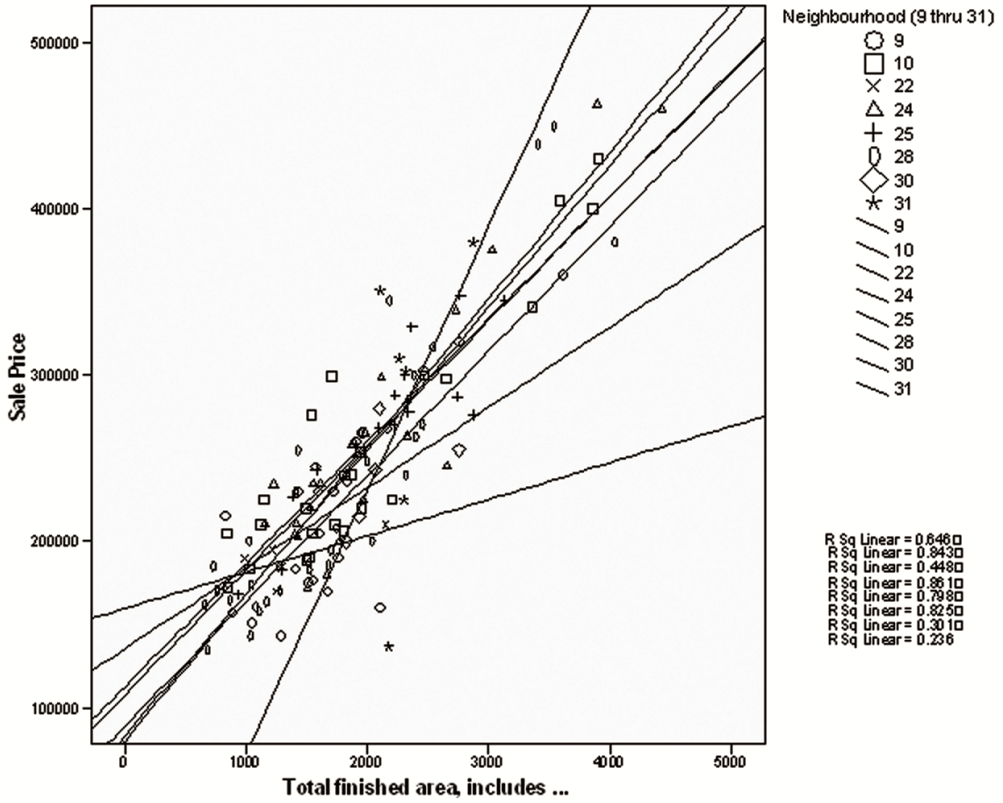
You will find that the linear relationships between S_PRICE, TOTAREA, and FINAREA are strong. Linear relationships between S_PRICE, NBHD, and LOT_SQFT are weak. Linear relationships between S_PRICE and the remaining variables are moderately strong.

Scatterplot with Multiple Trend Lines

It is also helpful in some cases to view a scatterplot with multiple trend lines, one for each value of another variable, such as quality class, year built, or neighbourhood.

In SPSS, you can accomplish this as follows:

- Click Dialog Recall button to recall Simple Scatterplot window. The Y axis should contain S_PRICE and the X axis should contain TOTAREA.
- Enter NBHD in the "Set Markers By" box. You will have to click the "Set Markers By" box, click the arrow to return QUALITY to the Variable list, then select the neighbourhood variable.
- De-select "Use chart specifications from: File → RSQ1.sgt (do not use template saved earlier).
- OK, runs the graph.
- Double click on chart to open the Chart Editor.
- Elements → Fit Line at Subgroups – displays regression line and R-square value for regression line for each NBHD.
- Close "Properties" window, produces a separate regression line for each NBHD.



From this graph, you can see that most neighbourhoods have a similar relationship between price and total area as exists for the total population.

In Excel SSC-Stat, you select neighbourhood as the "By Factor" variable and ensure "Show Trend Lines" is selected. You may need to expand the size of the resulting chart to be able to see the relationships.

Helpful Hint!

If you have SPSS options set to show colour output, it is difficult to identify specific neighbourhoods on a black and white printout. When printing in black and white or submitting assignments by facsimile, the person who reads your document may not be able to see some important information. You should use the following procedure so instead of colours SPSS will use various patterns or symbols in the output. In the SPSS Data Editor window, select Edit → Options → Charts. Then on the left hand side in the middle is a Style cycle preference drop down box. Change that to "Cycle through patterns only". Click OK.

Scatterplots with Split Files

It may also be helpful to produce individual scatterplots for each neighbourhood in order to examine these more closely. Earlier in this lesson we demonstrated the split file procedure for producing statistics or reports for various groups of properties, such as neighbourhoods. Now we will apply the same procedure to produce multiple scatterplots. We will reproduce the scatterplot of S_PRICE with TOTAREA with a regression line fit to the data. Then we will save this plot as a template which will be used to produce the same graph by neighbourhoods through the Split File facility.

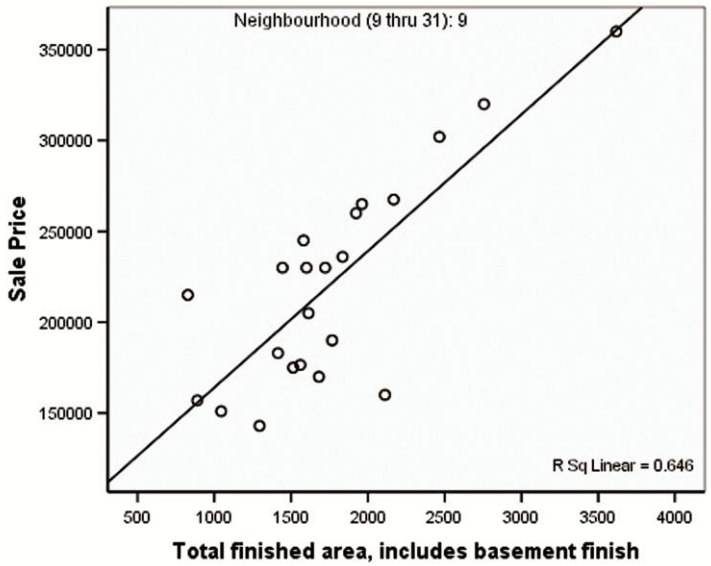
First, we reproduce the scatter graph with the regression line:

- Click Dialog Recall button to open the Simple Scatterplot window. The Y axis should contain S_PRICE and the X axis should contain TOTAREA.
- Click NBHD and left arrow to remove NBHD from "Set Markers by" window.
- Select "Use chart specifications from: File → RSQ1.sgt (use template saved earlier).
- OK, displays the scatterplot with the regression line and R-square statistic.

Next, we will split the data file by neighbourhood and reproduce the scatterplot for each neighbourhood:

- Data → Split File.
- Organize output by groups.
- Sort the file by grouping variable must be checked.
- Place NBHD in the Groups Based on box (to split the file by NBHD).
- OK (the "Split File On" message should appear in the bottom right corner of the SPSS window).
- Click the Dialog Recall box and select Simple Scatterplot.
- OK, produces individual scatterplots by neighbourhood.

The following is the scatterplot for the first neighbourhood (neighbourhood 9):



You can scroll through the output file to view the scatterplots for the other neighbourhoods. Use the down arrow to move to neighbourhood 10, 22, 25, etc. Use the up arrow to move in the opposite direction.

Helpful Hint!

In SPSS, you can double click the chart to edit a chart or view it more clearly. When in the Chart Editor mode, it is possible to identify the case or row number of any of the observations. This is especially useful where there are outliers. To identify the case number of points, click on the Data Label Mode button (looks like a square bulls-eye). This will change the cursor to a square with a cross in the centre. If you place this over a point and click, the case number will be displayed above the point. To examine the data for that case, right-click on it and select Go to Case.

Some of the charts are not very informative because of small sample size. The larger the samples, the more they approximate the plot for the jurisdiction, although some neighbourhoods could exhibit different patterns due to variations in market factors. Also, the slopes of the regression line or line of best fit seems different on some of these charts compared to the combined one. This is due to the different scales used on each chart for the horizontal and vertical axes.

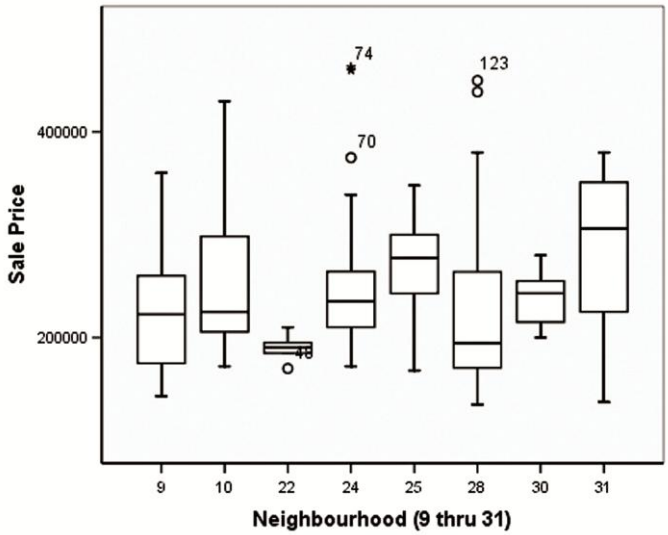
In Excel, you can get the same results manually by sorting the spreadsheet by NBHD and then creating a separate chart for each grouping of neighbourhoods.

Boxplots

Boxplots are useful for comparing the distribution of one variable by values of another variable. Normally the second variable will be a discrete variable with only a few distinct values. For example, you might be interested in comparing sales prices by neighbourhood. The boxplot provides some of the statistics that can be found in the descriptive statistics reports, but in a visual context. The "box" or rectangle of a boxplot contains the middle 50% of the cases. The lower and upper ends of the box represent the 25th and 75th percentiles. The dark line within the box represent the median (50th percentile). Points which are more than 1.5 box lengths from the edge of the box constitute *outliers* (values that lie outside the norm) and are represented by circles. Points which are more than three box lengths from the edge of the box are indicated by asterisks and are termed *extremes*.

In SPSS, use the following steps to create a boxplot of sales prices by neighbourhood:

- Data → Split File → Analyze all cases → OK, closes split file mode.
- Graphs → Legacy Dialogs → Boxplot → Simple → Define.
- Enter S_PRICE as the Variable.
- Enter NBHD as the Category Axis.
- OK, produces the following boxplot:



In Excel SSC-Stat:

- Select the columns for S_PRICE and NBHD.
- Click SSCstat → Visualization → Boxplot.
- Select S_PRICE as "Variable(s) to plot" and NBHD as the "By Factor" variable.
- Click "Display outliers" → OK.

A comparison of the boxes says much about both the level and consistency of sales in each neighbourhood. In this example, Neighbourhood 31 has the highest median price. Neighbourhoods 22 and 28 have the lowest. NBHD 22 has the least dispersion of sale prices and NBHD 31 has the greatest dispersion. Median prices are almost identical for NBHDs 9 and 10, but NBHD 10 has more dispersion at the high end and less dispersion at the low end. In all, there are five outliers and two extremes, both in neighbourhood 28 (properties 59 and 65, which sold for $463,000 and $460,000, respectively). The numbers displayed above NBHD numbers are the case counts for each neighbourhood, which is important information for interpreting the distribution of the boxplot.

Helpful Hint!

In SPSS, you can quickly examine the data on the outliers and extremes in the database. As the case numbers of the outliers are shown on the plot, determine the one you wish to view and return to the Data Editor window. Click the "Go to Case" icon (ruler with an arrow over top). In the Go to Case window, you can enter the case number of interest and SPSS takes you directly to that case in the Data Editor window. This is a handy way to locate and review outliers. Another useful shortcut is to open the Chart Editor for the boxplot, select one or all of the outliers, right click on your selection, then choosing "Go to Case". The relevant case(s) will then be highlighted in the Data Editor window.

Boxplots are a very useful tool for examining a continuous variable based on the characteristics of discrete variables. For example, they can be used to determine if there is a difference in the value of one variable for properties that are in different neighbourhoods, as above, or have a different number of bathrooms, bedrooms, or fireplaces for example. However, boxplots only provide a graphic presentation of these relationships and should not be relied upon to determine the precise value of the median or other statistics. This is especially true if there are a number of *outliers* or *extremes*, as the expanded scale may cause the boxes to be compressed, which may mask differences. The only absolutely accurate measures of the statistics illustrated in the boxplots are from the statistics reports produced earlier in this lesson (e.g., confidence interval for the median and compare means test).

Boxplots can be very useful for examining sales ratios by neighbourhoods and other subgroups for models that estimate market value. Ideally, boxes and their medians would align, indicating equal ratios – in other words, showing that the estimated values are equivalent for properties across neighbourhoods and with/without a given attribute (e.g., those with pools versus those without).

Boxplots can be supplemented by statistical tests which allow one to conclude whether median assessment levels are equal. For example, the "Mann-Whitney" test, which is used for discrete variables that can take on only two values (binary variables), and the "Kruskal-Wallis" test, which can be used when discrete variables can take on more than two values. The Kruskal-Wallis test uses the "chi-square" statistic to determine whether there is a significant difference in the mean values of the groups of properties (chi-square is pronounced like "kye-square" and can be written as chi^2 or χ^2, which is the Greek letter chi; the test is described in Chapter 2 in the *Advanced Computer-Assisted Mass Appraisal* book). Both the Mann-Whitney and Kruskal-Wallis tests are described in more detail later in the course.

Crosstabulations

Crosstabulation tables are useful for examining the relationship between two discrete variables such as neighbourhood, construction quality, construction type, and number of bedrooms. They can also be used to examine continuous variables such as sale price and lot size, but only if meaningful groupings of these variables are created. If the variable is continuous (e.g., price, finished area, lot size), you will need to supply "breaks" to divide the data into categories as was done with YRCLASS earlier. Crosstabulations can also use the chi-square statistic to test for independence between two variables. Crosstabulations and chi-square statistics are discussed at considerable length in Chapter 2 of the Advanced *Computer-Assisted Mass Appraisal* book.

Visual analysis and crosstabulation tables show the analyst much the same information. While later lessons will emphasize visual analyses, crosstabulations will be introduced here for those students who prefer this type of analysis.

To facilitate our analysis, we will break S_PRICE into three groups: (1) $200,000 or less, (2) $200,001 to $300,000, and (3) greater than $300,000.

In SPSS, this is accomplished by the Recode procedure:

- Transform → Recode Into Different Variables.
- Reset, clears previous values.
- Enter S_PRICE into the Input Variable to be recoded.
- Type SP_RANGE into the Output Variable box.
- Click Old and New Values.
- Click the Range: Lowest through value button and enter 200000.
- Enter 1 into the New Value box and click Add.
- Click the Range: button and enter 200001 and 300000.
- Enter 2 into the New Value box and click Add.
- Click the Range: value through Highest button and enter 300001.

- Enter 3 into the New Value box and click Add.
- Continue → Change → OK, creates the new variable, SP_RANGE.

In Excel, you would use the IF function.

Notice that SP_RANGE has three values: 1 (low through 200,000), 2 (200,001 through 300,000) and 3 (greater than 300,000). In SPSS, a crosstabulation of SP_RANGE with QUALITY is produced as follows:

- Analyze → Descriptive Statistics → Crosstabs.
- Enter QUALITY as the row variable.
- Enter SP_RANGE as the column variable.
- Click Statistics, opens statistics selection window.
- Select Chi-square and Correlations.
- Continue → OK, produces the following report:

QUALITY * SP_RANGE Crosstabulation **Count**

		SP_RANGE			Total
		1.00	2.00	3.00	
QUALITY	.55	1			1
	.67	4	1		5
	.81	19	1		20
	.94	17	43	2	62
	1.00	3	20	6	29
	1.20		2	15	17
Total		44	67	23	134

Chi-square Tests[a]

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-square	121.822	10	.000
Likelihood Ratio	110.110	10	.000
Linear-by-Linear Association	69.299	1	.000
N of Valid Cases	134		

a 9 cells (50.0%) have expected count less than 5. The minimum expected count is .17

Symmetric Measures

		Value	Asymp. Std. Error[A]	Approx. T[B]	Approx. Sig.
Interval by Interval	Pearson's R	0.722	0.044	11.983	.000[C]
Ordinal by Ordinal	Spearman Correlation	0.715	0.048	11.756	.000[C]
N of Valid Cases		134			

a Not assuming the null hypothesis.
b Using the asymptotic standard error assuming the null hypothesis.
c Based on normal approximation.

The chi-square statistic determines if two variables are significantly related. In the above analysis, the calculated chi-square statistic is 121.822 and the critical value of chi-square at 10 degrees of freedom[12] is 18.31 at a 95% confidence level (from "Critical Values of Chi-Square Table" found on the course website under "Online Readings"). As indicated by the corresponding "Significance" [Asymp. Sig. (2-sided)] value of .000, the probability of getting a value this high if the two variables, QUALITY and SP_RANGE, were not related is 0. Hence, the chi-square analysis strongly suggests that the two variables are related (statistically speaking, we

[12] Degrees of freedom indicates the number of statistics that can be accurately calculated from the limited data in a sample. The degrees of freedom are equal to the sample size minus the number of statistics calculated. For example, when you calculate the mean of a sample, you have used one degree of freedom. When you then use this mean in further calculations, you lose a second degree of freedom. You can think of it as advancing on the limbs of a tree: the trunk is solid (the data), the first branch is solid (mean), the second branch may still be solid (further statistic based on mean), but then the branches become increasingly weak. Continuing the analogy, the bigger the tree (bigger database), the farther out onto the branches you can safely go and still have reliable results. Degrees of freedom are used in many tests, such as chi-square tests, to indicate the statistical reliability of the results.

reject the null hypothesis that quality and price are not related; the other two statistics produced in the above report indicate the same conclusion and can be ignored).

This bears out our visual inspection of the relationship between the two variables. However, the accurate interpretation of these test results requires that the large majority of the cells in the crosstabs table contain at least five observations. The line below the "Chi-Square Tests" table indicates that nine of the 18 cells do not have the required five observations. Hence, we should be reluctant to draw conclusions about the relationships between the variables.[13] However, both the Pearson and Spearman correlation tests on the next table show a very high correlation (.72), which implies that the result of the chi-square test here is accurate.

Helpful Hint!

On the Crosstabs menu, you can click on Cells, under Percentages select Row, Column, Total to see the percentage of cases in each cell.

It may be noted that having the significance level (.000 in this case) makes it unnecessary to look up the critical chi-square value in a statistical table. At a 95% confidence level, the probability level must be less than 0.05 for the calculated chi-square value to be significant. If the probability level is less than 0.05, you can conclude the variables are not independent and that a statistically significant relationship exists between them.

As discussed above, the chi-square test needs at least five observations in each cell to produce accurate results. To account for this need while keeping a measure of quality in the analysis, the QUALITY variable will be grouped into two groups.

- Transform → Recode Into Different Variables.
- Reset.
- Enter QUALITY into the Input Variable to be recoded.
- Type Q_CLASS into the Output Variable box.
- Click Old and New Values.
- Click the Range: Lowest through value button and enter 1.
- Enter 1 into the New Value box and click Add.
- Click the Range: value through Highest button and enter 1.01.
- Enter 2 into the New Value box and click Add.
- Continue → Change → OK, creates the new variable, Q_CLASS.

Notice that Q_CLASS has two values: 1 (average and below quality) and 2 (above average quality).

We can now re-run the above analysis, replacing QUALITY with Q_CLASS.

- Dialog Recall → Crosstabs, to recall previous Crosstabs procedure.
- Replace QUALITY with Q_CLASS → OK, enter Q_CLASS as Row variable and run report.

Q_CLASS * SP_RANGE Crosstabulation

		SP_RANGE			Total
		1.00	2.00	3.00	
Q_CLASS	1.00	44	65	8	117
	2.00		2	15	17
Total		44	67	23	134

[13] There are special tables for interpreting the chi-square statistics when there are fewer than five observations in each cell. However, the SPSS manual and our course will not cover these tables. Therefore, if there are fewer than five observations in any cell, the crosstabulation results cannot be relied upon.

Chi-Square Tests[a]

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	69.383[a]	2	0
Likelihood Ratio	54.237	2	0
Linear-by-Linear Association	43.884	1	0
N of Valid Cases	134		

a 1 cells (16.7%) have expected count less than 5. The minimum expected count is 2.92.

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Interval by Interval	Pearson's R	0.574	0.059	8.062	.000[c]
Ordinal by Ordinal	Spearman Correlation	0.542	0.06	7.401	.000[c]
N of Valid Cases		134			

a Not assuming the null hypothesis.
b Using the asymptotic standard error assuming the null hypothesis.
c Based on normal approximation.

Even with this recode, there is still one cell with less than 5 observations, which means that the chi-square statistics remain inaccurate.

We will now produce a crosstabulation of YRCLASS and Q_CLASS.

- Dialog Recall → Crosstabs.
- Replace SP_RANGE with YRCLASS as the column variable.
- OK, produces the following report:

QCLASS * YRCLASS Crosstabulation

		YRCLASS					Total
		1.00	2.00	3.00	4.00	5.00	
QCLASS	1.00	5	11	77	16	8	117
	2.00	0	0	0	5	12	17
Total		5	11	77	21	20	134

Chi-Square Tests[a]

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	56.276[a]	4	.000
Likelihood Ratio	51.970	4	.000
Linear-by-Linear Association	42.721	1	.000
N of Valid Cases	134		

a 5 cells (50.0%) have expected count less than 5. The minimum expected count is .63.

Symmetric Measures

		Value	Asymp. Std. Error [a]	Approx. T[b]	Approx. Sig.
Interval by Interval	Pearson's R	.567	.062	7.903	.000 [c]
Ordinal by Ordinal	Spearman Correlation	.561	.058	7.780	.000 [c]
N of Valid Cases		134			

a Not assuming the null hypothesis.
b Using the asymptotic standard error assuming the null hypothesis.
c Based on normal approximation.

Notice that all 17 properties with above-average construction quality fall into the two newest age groups (1981 to 2000, and 2001 and above).

Try other crosstabulation tables to become familiar with the operation of this module and the various options available. In applying crosstabulation, you will need to recode all continuous variables. These include S_PRICE, EFFYRBLT, TOTAREA, FINAREA, and LOT_SQFT. Discrete variables, which do not necessarily require breaks, include NBHD, QUALITY, BEDROOMS, BATHS, FAMROOM#, and

YRCLASS. In particular, check for relationships between SP_RANGE and the discrete variables mentioned above. In most cases, you will find a relationship. However, you should also test the relationship between SP_RANGE and LOT_SQFT.

Keep in mind that you could also review the relationships of these variables using scatterplots and boxplots.

Closing, Saving, and Backing Up Data

This is the end of our work with the "Burnaby" database. You can now close the database. You may wish to save the revisions made during this lesson, either by selecting File → Save and overwriting the original database or File → Save As… and specifying a new name. In SPSS, you may also save any output files or syntax files created, using Save As and specifying a filename.

Helpful Hint!

It is good practice to get into the habit of backing up data throughout this course (and in your work and personal lives). Backing up your database periodically is a good idea in case you make a mistake or somehow lose your data – as an example, having to start over again on your major project after losing your data would obviously not be ideal!

Recommended saving and backing up practices include:

- First download all original files to a directory on your computer and then make a working copy elsewhere – save the databases in original form in case you need to refer back to them.
- As you work with data, especially if you are running intensive modelling repeatedly on the same data, get in the habit of periodically re-saving the data with a new name, e.g., "Burnaby Oct1 initials", so that if you come into data problems you can easily revert back to a version before errors were introduced.
- Be VERY careful about "Save" versus "Save As" – if you use "Save", it permanently overwrites whatever you had with what you have now, so you need to be careful not to accidentally overwrite data.
- Periodically save backup copies with a new name and to a new directory on your computer, so that you avoid overwriting your data accidentally.
- To be really safe, save a backup copy onto a CD and store it in a different location than your computer (in case of fire or theft).

Summary

This lesson built on the introduction to statistics in Lesson 1 by showing how these statistics can be quickly and easily computed using computer applications. We have illustrated common statistical procedures using a statistical program, SPSS, as well as in a standard business spreadsheet, Microsoft Excel, wherever possible. In this lesson, we covered descriptive statistics, transformations, graphic analysis, and crosstabulations.

Now that we have introduced basic statistics and their computation, we can now proceed to demonstrate their practical application.

- Lessons 4-5 will illustrate practical applications of statistical procedures in everyday valuation work. These will largely focus on Microsoft Excel.
- Lessons 6-8 will illustrate valuation modelling, using databases, statistical procedures, and software to create models that estimate the value of real property. These will focus on SPSS.

The course will provide a variety of case studies showing real-life examples of how these statistical procedures have been applied in valuation work (for fun and profit).

Before proceeding to illustrate these applications, we have one further foundational topic to cover, data analysis. All applications of statistics in real estate work are based on effective use of data, analyzing it to find the relationships between variables and unearth the information hidden beneath the numbers. Lesson 3 will examine exploratory data analysis in detail, helping you to seek out the relationships forming the basis for our further work in this area.

NOTE FROM THE TUTOR

Following up the note at the start of this lesson: we have covered a lot of ground in Lessons 1 and 2 – you should not be surprised if you're feeling a bit overwhelmed. You may wish to consider re reading this lesson a second time, and referring back to Lesson 1 as well, because you will likely notice things the second time through that you missed the first time. Beyond this lesson we will begin to apply these concepts. If you think of this like carpentry, we have now shown you how to use a hammer and saw, and from here we will begin to demonstrate how these tools can be used to build things of use.

Review and Discussion Questions

1. Would you agree with the following statement: "Using a sophisticated statistical analysis tool, such as SPSS, will produce a more reliable forecast than the simple statistic tools available in Excel."

2. Assume you had completed a statistical analysis in Excel and produced descriptive statistics for a small sample of office property rents for Class B buildings in Kelowna. Is there any point in considering the *Count* function when you interpret the findings and report to your client?

3. The descriptive statistics summary provides a mean and median of the dataset. Can you think of any reason(s) why the median may be better indication of central tendency than the mean?

4. In the Burnaby database, determine the lower and upper bounds for a 95% confidence level for the mean for the *TOTAREA* variable. What happens to the confidence interval when the level is set to 90%? What changes if it is instead the confidence interval for median instead of the mean?

5. For the ratio of *TESTVAL* by *S_PRICE*, determine the confidence interval at a 90% confidence level. How did the outcome change from the example in Lesson 2? Why?

6. The median Assessment-Sale Ratio (ASR) and Coefficient of Dispersion (COD) are commonly used in property tax assessment. Assessment organizations normally set goals or targets for desired levels of the median ASR and Coefficients of Dispersion (COD) to ensure quality outcomes for assessments. Can you think of ways to manipulate the statistics in order to achieve target levels?

7. What are some statistical measures that you can use to test for the "normality" of your data sample? What are the strengths and weaknesses of the various approaches?

8. Test the *LOT_SQFT* variable for normality. What are your conclusions?

9. Assume you were working with a large dataset of office vacancy observations in a large market area. The data consists of observations of vacant space (sq ft) and gross leasable area (GLA) for each building sampled. Your interest is predicting vacancy for a specific range of office buildings. How would you accomplish this task with SPSS? What variables would be required?

10. City Planners are generally interested in trends for density of various housing types for specific neighbourhoods. Considering the Burnaby data, how would you transform the data to produce one or more measures of housing density? What do you conclude about changes in residential density over time?

11. For the Burnaby data, provide some examples of data fields you would recommend for "re-coding" to improve the statistical analysis. What benefits would re-coding provide in these scenarios?

12. For the Burnaby data, determine the Coefficient of Determination for sale price and quality.

    (a) What is the dependent variable? What can you conclude about the relationship between these two variables?

    (b) Consider how you could improve the R^2 value for this analysis, or in other words, reduce the sources of unknown variation. Re-run the analysis to determine whether the R^2 value has improved.

13. Create a boxplot comparing sales price per square foot of finished area and neighbourhood. In which neighbourhood(s) does the sales price per square foot appear to have the highest number of occurrences between the 25th and 75th percentiles? What would you conclude about neighbourhood 22?

14. Test the hypothesis that there is a positive relationship between the number of bedrooms in a house and the effective year built. In other words, you would like to determine if the number of bedrooms in a house has increased over time. Using the Burnaby data, develop a crosstabulation table and examine the chi-square statistics. What did you learn?

ASSIGNMENT 2

LESSON 2: Statistical Software Applications for Real Estate Analysis

Marks: 1 mark per question.

THE FIRST 18 QUESTIONS REFER TO THE BURNABY DATABASE USED IN THIS LESSON.

1. Determine the following descriptive statistics for the *LOT_SQFT* variable: mean, median, mode, standard deviation, standard error of the estimate, and range.

   Given these statistics, what can you conclude about the data?

   (1) Since the mean and median are very similar for the *LOT_SQFT* variable, either measure would appear to be a useful indication of central tendency.
   (2) The similarity of the median and mean suggests that the data may not follow a normal distribution (e.g., bell-shaped curve).
   (3) The standard deviation is a significant amount in relation to the mean so we expect the data occurrences to be tightly clustered about the mean.
   (4) None of the above.

2. Calculate the following statistics:

   - Mean, median, and COV (Coefficient of Variation) of *TESTVAL*
   - Mean, median, and COV (Coefficient of Variation) of *FINAREA*
   - R^2 (R-square or Coefficient of Determination) of *TESTVAL* versus *FINAREA*

   Which of the following statements is TRUE?

   (1) The mean of *TESTVAL* is less than the median.
   (2) The COV of *FINAREA* is less than the COV of *TESTVAL*.
   (3) The R^2 of *TESTVAL* versus *FINAREA* is 0.875.
   (4) All of the above.

3. Determine the confidence intervals for *QUALITY* (requires SPSS or NCSS). What can you conclude from the statistics?

   (1) Given a sample size of 134 we can be 90% confident that the mean quality lies between 0.9354 and 0.9713.
   (2) Given a sample size of 134 we can be 95% confident that the mean quality lies between 0.9501 and 0.9672.
   (3) The confidence interval for the mean quality decreases as the confidence level is increased from 90% to 95%.
   (4) None of the above.

Assignment 2 continues on next page

4. Create the histogram for *TOT_AREA*. From this, you can conclude this variable is:

   (1) normally distributed.
   (2) skewed to left.
   (3) skewed to right.
   (4) flat.

5. The variable *BEDROOMS* is best analyzed against *TOT_AREA* by which of the following?

   (1) Scatterplot
   (2) Boxplot
   (3) Crosstabulation
   (4) Normality test

6. Which of the following best describes the boxplot for sale price by bedrooms?

   (1) Four boxes, with the median of each increasing from left to right.
   (2) Approximately 11 boxes, each higher as you move right.
   (3) Four boxes, generally increasing, with approximately equal dispersions.
   (4) Seven boxes, with no clear pattern of high and low medians.

7. Which of the following best describes the scatterplot of *TESTVAL* against *EFFYRBLT?*

   (1) A strong positive relationship, closely dispersed along the best fit line.
   (2) A moderate positive relationship, but with a fair amount of dispersion at the high end.
   (3) A moderate negative relationship, with more dispersion at the low end.
   (4) A weak relationship, with no clear positive or negative relationship.

8. For the *SP_SQFT* variable transformed in Lesson 2, determine the coefficient of variation (mean centred).

   (1) 22.1%
   (2) 44.0%
   (3) 28.5%
   (4) 25.7%

9. Transform a new variable for sale price per square foot of lot size, by dividing sale price by lot square footage. The mean of this new variable is:

   (1) 39.88
   (2) 35.47
   (3) 0.025
   (4) 16.45

Assignment 2 continues on next page

10. Create a crosstabulation table of *YRCLASS* and *QUALITY*. Which one of the following statements is correct?

    (1) There are 8 observations in Class 1 of *YRCLASS*.
    (2) There are 11 observations in Class 2 of *YRCLASS*.
    (3) There are 57 observations for *YRCLASS* = 5, *QUALITY* = .94.
    (4) There are 14 observations for *YRCLASS* = 3, *QUALITY* = .81.

11. What can you conclude from examining the normal probability plot for *TESTVAL*?

    (1) *TESTVAL* appears to be normally distributed.
    (2) *TESTVAL* deviates from a normal distribution at the high and low end.
    (3) *TESTVAL* deviates from a normal distribution at the high end, indicating it is skewed to the right.
    (4) *TESTVAL* deviates from a normal distribution at the low end, indicating it is skewed to the left.

12. Analyze *TESTVAL* using *BEDROOM* value as a break variable. What is the standard deviation for the 2 bedroom sub-group?

    (1) 78,185
    (2) 39,782
    (3) 39,500
    (4) 34,349

13. Analyze the normality of sale price by bedroom. What can you conclude about the normality of the 3-bedroom group? (Hint: Explore in SPSS, Descriptive Statistics in NCSS, Normal Probability Plot in Excel SSC-Stat using only records with 3 bedrooms)

    (1) The relationship between sale price and 3-bedroom units is moderately normal.
    (2) The relationship between sale price and 3-bedroom units is very non-normal.
    (3) The relationship between sale price and 3-bedroom units is perfectly normal.
    (4) The relationship between sale price and 3-bedroom units is insignificant.

14. Use a scatterplot to compare *S_PRICE* to *YRCLASS*. What can you conclude about the relationship between these two variables?

    (1) The relationship is not significant.
    (2) The relationship between *S_PRICE* and *YRCLASS* is moderately significant with an R^2 value of 0.56.
    (3) The relationship between *S_PRICE* and *YRCLASS* is highly significant with an R^2 value of 0.77.
    (4) None of the above.

Assignment 2 continues on next page

15. Would you rely on a boxplot to examine the relationship between *TEST_VAL* and *TOTAREA*?

(1) Yes, since both are discrete variables.
(2) Yes, since *TEST_VAL* is continuous variable and *TOTAREA* is a discrete variable.
(3) No, since both are continuous variables.
(4) No, since there is insufficient alignment of the medians.

16. Use crosstabulation (and chi-square statistics), a boxplot, and a scatterplot to investigate the relationship between *YRCLASS* and *FAMROOM#*. Which of the following statements is correct?

(1) The variables are strongly related, with a high chi-square and clear relationships in both boxplots and scatterplots.
(2) The variables are unrelated, with a low chi-square and no relationship shown in the scatterplot.
(3) The chi-square statistic appears to show a significant result, but the scatterplot shows a relatively weak relationship.
(4) All of the above are correct.

17. A common theory is that the average size of a house has grown over the years. Test this hypothesis by examining the relationship between total area and effective year built. Choose the correct response below.

(1) There is no clear relationship between the size of a house and year built beyond the year 2000.
(2) There is a moderate relationship between the size of a house and year built up to year 2000.
(3) The R^2 value for the relationship of house size to year built is .534.
(4) All of the above are correct.

18. Create a *TESTASR* variable by dividing *TESTVAL* by *S_PRICE*. If you viewed a scatterplot of *TESTASR* by *TOTAREA*, and then added *BEDROOMS* as third factor/marker variable, what would learn?

(1) *TESTASR* and *TOTAREA* have a strong positive relationship: larger units have larger ASRs.
(2) Most of the two-bedroom units are above a *TESTASR* of 1, meaning they are over-assessed.
(3) Most of the three-bedroom units are below a *TESTASR* of 1, meaning they are under-assessed.
(4) None of the above.

Assignment 2 continues on next page

THE FOLLOWING TWO QUESTIONS RELATE TO THE BUSI 344 PROJECT 1 DATABASE CALLED "CONDOSALES". PLEASE DOWNLOAD THE PROJECT 1 FILES FROM THE ONLINE READINGS ON THE COURSE RESOURCES WEBPAGE. YOU MAY ANALYZE THE FILES IN MS-EXCEL OR SPSS.

19. Create a cross-tabulation table between the Competitive Set Rank and the Location Quality Rank. Given that the Competitive Set Rank for University Gate is ABOVE AVERAGE (value 4) and the Location Quality Rank for University Gate is AVERAGE (value 3). Which of the following statements is FALSE?

    (1) The 23 sales of ABOVE AVERAGE competitive set rank well-bracket University Gate across the location quality rank
    (2) There is a strong relationship between the location quality and competitive set rankings
    (3) There are only two sales that match the location quality and competitive set rankings of University Gate
    (4) The majority of the sales with AVERAGE location quality rank have a BELOW AVERAGE competitive set rank.

20. Run a scatter plot of the Sale Price versus the Size of the condominiums and add a fit line. What does this show you?

    (1) There is a weak relationship between the Sale Price and the Size of the condominiums in the "condosales" database
    (2) The relationship between the Sale Price and the Size of the condominiums is logarithmic
    (3) A size adjustment is needed
    (4) Both (1) and (3)

20 Total Marks

SPSS CROSS-REFERENCE GUIDE

In an appendix at the end of this workbook, you will find an "SPSS Cross-Reference Guide". This reference document will help you quickly locate instructions for various SPSS procedures used in BUSI 344.

End of Assignment 2