SAUDER
School of Business

Real Estate Division

LESSON 3

Exploratory Data 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 3: Exploratory Data Analysis

2. Abromaitis, C. 2004. "Computers, Graphs And Statistical Thinking: More Than Pretty Pictures", *Valuation Insights and Perspectives*. Fourth-Quarter 2004. A light informal look at the use of graphing in the valuation process; a good overview of the "why" of graphical analysis.

Recommended Reading

1. Goddard, B.L. 1999. "The Role of Graphic Analysis in Appraisals". *Appraisal Journal*. October 1999.

2. Goddard, B.L. 2000. "The Power of Computer Graphics for Comparative Analysis". *Appraisal* Journal. April 2000.

3. Hartwig, F. & Dearing, B.E. 1979. *Exploratory Data Analysis*. Sage University Paper Series on Quantitative Research Methods, Vol. 16. Newbury Park, CA: Sage. (data examples are non-real estate, but serves well as a non-technical primer on graphical data analysis)

4. UBC Real Estate Division. 2009. *Advanced Computer-Assisted Mass Appraisal*. Vancouver: UBC Real Estate Division.
   Chapter 1: Introduction to Statistical Methods
   (sections on sampling theory, statistical inference, and confidence intervals)
   Chapter 3: Mass Appraisal Model Building
   (sections on exponential and logarithmic transformations)

5. UBC Real Estate Division. 2009. *Foundations of Real Estate Mathematics*. Vancouver: UBC Real Estate Division.
   Chapter 15: Introduction to Sampling Theory

6. Kane, M.S., Linne, M.R., and Johnson, J.A. 2004. *Practical Applications in Appraisal Valuation Modelling*. Chicago: Appraisal Institute.
   Chapter 2: Review of Statistical Analysis
   Chapter 3: Data Exploration.

Learning Objectives

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

1. explain the importance of exploratory data analysis in real estate applications;

2. describe the "Four Rs" and the function of each in exploratory data analysis;

3. differentiate between variable types and understand the possibilities and limitations for each in data analysis and model building;

4. use summary statistics to reduce the uncertainty in data;

5. use graphic analysis, including scatterplots, boxplots, and histograms, to reveal data characteristics and trends;

6. use transformations to re-express data into formats that better facilitate analysis;

7. recognize non-linear data relationships and understand how logarithmic transformations can be used to account for this; and

8. analyze residuals in order to review and assess model quality.

Instructor's Comments

Lessons 1 and 2 both explored statistical foundations, with Lesson 1 focusing on theoretical underpinnings and Lesson 2 exploring how computer applications can be used to do the necessary calculations and graphics. This lesson will focus on data analysis, the third and final "building block" in our statistical analysis foundations. You can think of our approach as building a three-legged stool, with statistics, computer software, and data each representing a leg. With the three legs in place, we can then make the "step up" to practical applications of statistics.

Data is a critical element in statistical analysis because it serves as the basis for all applications. If you do not have a solid understanding of your data, how it is compiled and arrayed, or the relationships within it, you are bound to end up with problems in the inputs and outputs of your analyses. Exploring data is necessary in all forms of real estate analysis, whether supporting a size adjustment for a comparable in a house appraisal, determining a market capitalization rate, or using regression modelling to estimate the values of condominiums throughout a large city. All of these analyses start with data exploration – and without sound data exploration they all inevitably end poorly!

Exploratory data analysis is based on the "Four Rs":

1. Reduction
2. Revelation
3. Re-Expression
4. Residuals (from models)

We will examine each of these in turn to better understand what comprises data analysis. By the end of the lesson, you will realize that we have already covered most of the essential tools in Lessons 1 and 2, just perhaps not yet focused on our exploratory goal. Throughout this lesson we will use real estate examples to showcase the Four Rs. Our goal is to illustrate how to structure problems towards finding productive solutions. We will illustrate how statistics and graphics can be used to properly analyze data, in order to find accurate and helpful answers. Along the way we will also examine common pitfalls in data analysis, hopefully heading off some problems before they occur.

Before we begin our foray into data analysis, a note on software: Lesson 2 focused on two software packages, Microsoft Excel and SPSS. This lesson and the remainder of the course will continue this multi-application approach, focusing on spreadsheet capabilities where feasible and then on statistical applications once we have reached the spreadsheet's limits. In this lesson, most of the output was produced in either Excel or SPSS. You will note we either abbreviate or do not provide calculation steps where they were already provided in Lesson 2.

An SPSS cross-reference guide is supplied as an Appendix at the end of this Course Workbook. This can serve to help you quickly find the full set of instructions for various SPSS procedures.

NOTE FROM THE TUTOR

Lesson 3 is our third and final "foundations" lesson. Following this lesson we will be focusing on applications, first for single property assignments, then for mass appraisal. Students often find it more enjoyable and interesting to review the more practical uses of statistics, so you have that to look forward to!

Introduction to Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an approach to learning from data. A by-product of both the computer revolution and the growth of the Internet has been an exponential growth in the amount and complexity of available data. EDA is a way to help make sense of the vast data facing today's business analyst. The analyst must adopt an active role, an exploratory point of view, to seek out the data patterns that lead to productive hypotheses. The EDA process is about evaluating, synthesizing, and leveraging information.

EDA employs mostly visual techniques, stressing a penetrating look at data to reveal suggestive patterns and its underlying structure. We will explore simple but effective graphical methods of exploratory data analysis in a real estate context, with the goal of finding and interpreting patterns in property markets. For example, by plotting real estate data and analyzing it graphically, we can obtain deeper insights into why prices vary.

Statistical graphics are widely used to communicate information, such as enlivening a presentation or focusing attention on specific important points. With today's powerful computer hardware, (reasonably) easy-to-use software, and increasingly available data, comprehensive data analysis is becoming a client expectation. Studies that were once either too expensive, too tedious, or too difficult to complete can now be done with the click of a mouse. The graph, initially a static object that had to be carefully constructed, is now a dynamic partner in the investigation of empirical phenomenon. Ease of use has transformed a somewhat obscure field into a business necessity.

Why Graphs?

Appraisers often use graphs or charts for presenting information, such as presenting demographic data in a market area description. However, they are not typically employed in the valuation section of a report. Bar and pie charts are probably the most commonly used charts. In general, the use of graphics as an analytical tool tends to be passive, a part of presentation, versus an active analytical tool that helps derive meaning from the data, facilitate understanding, and infer conclusions. In other words, graphs are sometimes (or often) used as "pretty pictures" and not much else.

Graphs help us process large amounts of data, taking what otherwise would be incomprehensible and allowing our limited brain processing power the ability to draw meaningful conclusions. Simply displaying data can sometimes help with more effective analysis. But in a graph, the analyst can incorporate a very large amount of quantitative information in a very efficient manner. Instead of reviewing a vast array of numbers in a grid, the analyst can instead see a curve and know in seconds the underlying relationship in the data. Without the graphic representation, understanding might not be possible. In this case, the graph helps by downplaying specific values and instead emphasizing important features like shape of the data distribution. The essence of visualizing data is to find a trend line or pattern within a large grouping of observations.

We will explore ways to, quite literally, look at your data. Our focus will be on easily producible and interpretable displays which provide maximum insight. For example, we will use histograms and boxplots to visualize the distribution and shape of single continuous variables. We will also use scatterplots to examine the relationship between two continuous variables.

Later in the course, we will look at geographic information systems (GIS). Graphic analysis is a form of spatial analysis, looking for patterns in two-dimensional space. GIS ties this spatial analysis to physical location, in effect adding a third dimension to our graphs. In this case, the graphs we look at are geographic maps, with helpful information coded into them.

The rest of this lesson focuses on dynamic tools for examining data. However, before exploring these tools, we will first discuss some underlying concepts in data management. You cannot explore data until you understand it!

Data Entry and Management

First, let's discuss what we mean by "data". A set of data involves a number of "variables" and "observations" or "cases". The "cases" are the observations accumulated into the dataset, such as 550 property sales, 120 leases, or 14,562 automobile purchasers. The "variables" are the characteristics of the cases, such as the number of bedrooms, square footage, base rents, or car colour preferences. The dataset will typically be constructed as a grid, with variables and cases arrayed along the columns and rows.

In order to explore data with the computer, we need it in digital form. A good paper data filing system may work for many uses, but it is close to useless for efficient data analysis. Data entry by typing is a time-consuming task. Whenever possible, you will want data that can be easily imported into the software packages you are using. For example, most statistical software can automatically convert the data from Excel files. Import techniques vary with the source of data and software package.

Helpful Hint!

The typical appraisal grid format places sales (cases) across the top in columns and the property attributes appearing as rows. However, statistical software packages are reversed, with cases in rows and variables in columns. In order to standardize datasets for ease of use and export to other software packages, you may want to organize your datasets using this format, where the rows are the individual sales or cases and the columns are the variables. You also want to avoid any empty rows and columns in your spreadsheets if you plan to use them for data entry into statistical software. Finally, you should avoid formatting the data in your spreadsheet too much, as this can interfere with the transfer of data into other software packages. Leave the formatting for later, when you are producing your final report.

Variable Types

In order to use statistics productively, the analyst first needs to understand the data being analyzed. There is a variety of data types, and each has its own peculiar characteristics affecting how it is measured and how to interpret its results. Confusing data types in any kind of analysis can lead to very bad results. Where data is not in a form that is useful, often you need to transform it to a more ideal type.

Ordinal variables are likely the easiest to understand – first, second, third are the fundamental examples of ordinal variables. Someone wins first prize in a chili cooking contest, someone wins second prize, and someone else third. What is the relationship between the three rankings (and others who finished out of the prizes)? There is none other than they all entered the same competition – you cannot say how much better the second place chili is when compared to the third place chili. As another example, how much "better" is a first place marathon runner over a second place marathon runner if you don't know the difference in their times?

Consider a real estate example: you have received data on ten single family detached residential sales, including all sorts of inventory variables such as lot size, main floor area, year built, and number of bathrooms. However,

for quality of construction all you have is a ranking, from first down to tenth. It would be difficult to do any meaningful analysis when this quality variable is a subjective ranking; perhaps it could be more useful if coded using an objective rank, such as a Marshall and Swift rating class.

Nominal variables contain less information than ordinal variables – there is absolutely no relationship between any two, other than the fact that they are different. The values for the variables are those of convenience rather than for analysis, e.g., type of construction: wood frame, concrete, or steel. It is difficult to add these together or get a maximum or minimum.

Interval variables have a relationship between them, e.g., House A built in 2011 is five years newer than House B built in 2006. However, the relative "distance" between the two does not have a direct mathematical relationship or meaning. For example, is House A 2011 ÷ 2006 = 1.00249 times better than House B based on year built? Probably not.

Ratio variables on the other hand can tell us a great deal. For example, a 5,000 square metre warehouse is twice as large as a 2,500 square metre one. A house with 50 metres of waterfront has 25% more than a house with 40 metres.

Variables can have three other pairs of characteristics as well:

1. *qualitative* (describing a quality) or *quantitative* (something that can be counted or measured).
   Example:
   - qualitative: excellent landscaping
   - quantitative: 3 bathrooms

2. *subjective* (based on opinion) or *objective* (based on fact).
   Example:
   - subjective: average view, on a five point scale – excellent, good, average, fair, poor; not every person would agree
   - objective: brick exterior finish or 4 bedrooms

3. *continuous* (given any two observations a valid observation could be found between the two) and *discrete* (two observations may not have a valid observation between them).
   Example:
   - Continuous: main floor area in square feet, with one house at 1,300, another 1,301, and yet another 1,308.5, and so on)
   - Discrete: number of bedrooms, where no valid value occurs between three and four bedrooms; quality of construction, rated as either excellent, good, average, fair, or poor, and with no valid value between good and average.

Binary variables (or *dummy variables*) are a special case of a discrete variable used for non-numeric variables, such as location, building features, and views. A binary variable, as the name implies, has only two possible values; the classic example is *on* and *off*. These are most often used in data analysis to indicate the presence or absence of a particular characteristic. For example:

- office mezzanine is present in a warehouse: 1 yes, 0 no
- a house has a swimming pool: 1 yes, 0 no

Binary variables serve a very important purpose in data analysis as will be examined in detail throughout the course. For example, consider a variable that describes properties as being in one of three distinct neighborhoods, A, B, or C. You could create three variables, NBHDA, NBHDB, and NBHDC, and each variable would have its own column. A property located in A is coded 1 for NBHDA, and 0 for NBHDB and NBHDC.

All of your variables must be carefully inspected to determine how best to use them in analysis – failure to do so could result in very weird and unsupportable findings. As well, different variable types are connected with different optimal analyses: for example, a continuous variable such as sale price is best viewed in a scatterplot, while a discrete variable such as bedrooms is best viewed in a boxplot or crosstabulation. This is explained further in the next section.

Exploratory Data Analysis Techniques: The "Four Rs"

In this section, we will explore the "four Rs" of EDA: reduction, revelation, re-expression, and residuals (from models).

Reduction: Simplifying the Data

> *Reduce: simplify…by classification or analysis*.
> Concise Oxford Dictionary

Reduction means simplifying the information, focusing it to a small enough "package" that it becomes comprehensible. As an analogy, consider how the term "reduction" is used in cooking: boiling down a soup until the excess liquid is evaporated, with the broth becoming increasingly concentrated. The same process is used in data analysis: "boil it down" until the essential elements become clear.

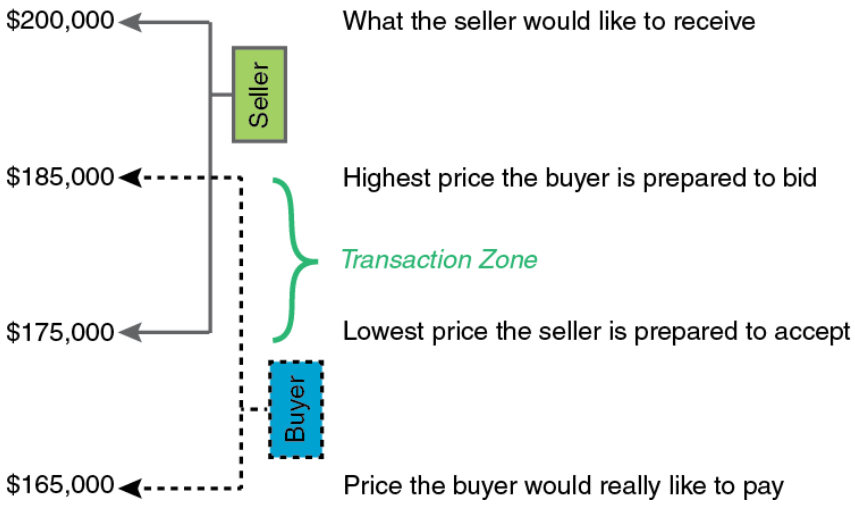
Let's start with a vast array of numbers all jumbled together and with no organization. You are appraising a property and you have a pile of sales reports, listings, inspections, all mixed together on your desk. In attempting to estimate the subject's market value, this disorganized data does you little good. So, our first step is to summarize the data in a way that makes sense. As discussed earlier, we will organize the data into a table in a computer program, with variables in the columns and the cases in rows.

At this point, we still have not found any patterns in the data, but we now have something we can look at and begin to make sense of. We have started our process of reducing the data to a manageable level.

The first thing we probably notice about our data is that it is not all the same. If the data was all identical, it probably would not tell us much nor would it require much analysis to identify whatever it could tell us. Variation is everywhere and the analyst's primary task is to explain this variation and use it to draw conclusions of value to a client. Think of it this way: without variation, the client probably wouldn't need an analyst (as some might say, "if it was easy, then why would they hire you?").

Before we explore methods to reduce data, consider the concepts of unexplained and random variation. Not all variation can be explained; some variation is random. For example, human behaviour in the real estate marketplace is not always rational and sometimes cannot be explained by objective methods. Consider Figure 3.1 below, which illustrates how sale price for a property may be negotiated between parties. An appraiser can at best attempt to objectively estimate a value in the $175,000-$185,000 "transaction zone", although the exact value in here is difficult to pinpoint. However, the extreme edges of the seller's "ceiling price" and buyer's "floor price" are completely subjective and impossible to even guess at. If a seller is demanding $10 million for a modest house and happens to find the one person in the world willing to pay that much, it is difficult for an outsider to this transaction to explain this rationally. In statistical terms, this would be called an "outlier" and probably removed from the analysis. This will be discussed later in this lesson.

Figure 3.1: Sale Price and the Range of Values



Our goal is to explain as much variation as possible, keeping in mind that at some point there will always be uncertainty left-over. Statisticians call this "error", which to some gives a false impression of a mistake being made. We will instead call this a "residual" or what's left-over after we explain all rational variations possible.

Our first tool in data reduction is to view summary descriptive statistics. We will use the "Burnaby" dataset from Lesson 2 and examine the S_PRICE variable:

N	Valid	134
	Missing	0
Mean		243,586.78
Median		230,000.00
Std. Deviation		72,465.610
Range		328,000
Minimum		135,000
Maximum		463,000

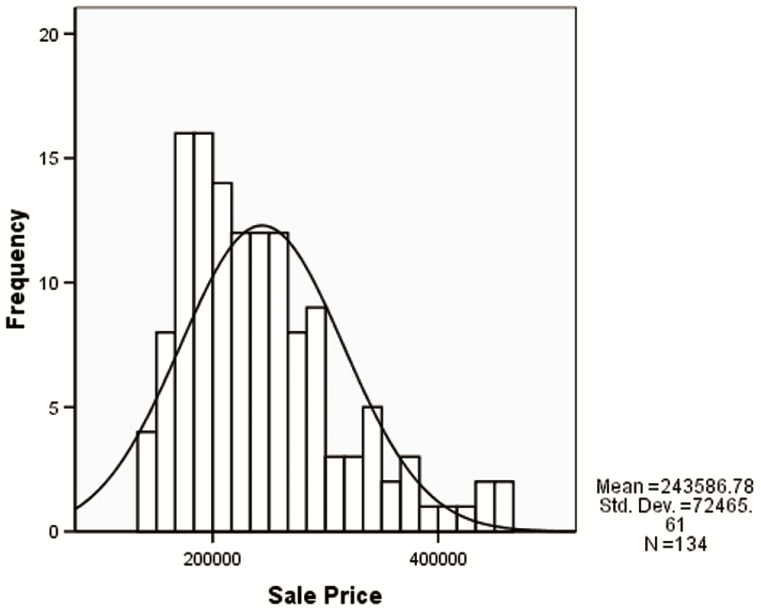
This output shows a few simple measures of central tendency and dispersion for the sale price variable. Very quickly we know more than we could from just scrolling through the dataset. We know the middle value is $230,000 and the mean is $243,600, showing the data is skewed to the right.

The measures of central tendency have little meaning without accompaniment by dispersion measures. Knowing the centre of the distribution is only half the story, we also need to know how spread out the data is around this centre. The range is $328,000 and the standard deviation is $72,465, indicating the data is fairly widely dispersed.

These summary statistics give us a sense of the data. For example, say the standard deviation was instead $10,000 or $100,000, what would that tell us about the spread of the data? What if the median was actually larger than the mean? Sometimes these statistics in themselves might not mean all that much, as their chief value may come in comparing to other similar data distributions. For example, how does S_PRICE relate to TESTVAL?

Next we'll view a histogram of the data. A histogram is analogous to a bar chart in Excel. It shows the data distribution in an array, e.g., how many observations are found in each $50,000 interval of price.

- Graphs → Legacy Dialogs → Histogram…
- Enter S_PRICE in the Variable field, select Display normal curve, then click OK.



The S_PRICE histogram shows more observations bunched up on the left and then tailing off to the right. As we inferred from our descriptive statistics, the variable appears to be "right-skewed". The normal curve is superimposed on the graph simply as a reference.

When analyzing histograms, you need to consider the group's centre point as well as the variability or spread on either side of the centre. For example, refer back to Lesson 1 where we showed graphs of different distributions that had similar medians but very different distributions because of their variability.

So, very quickly we have reduced some uncertainty about the data. We started with a hodgepodge of disorganized data. We organized the observations into a table. We produced summary statistics and a histogram to get a better sense of what the data is and how it is distributed. Now we'll proceed to the second "R", revelation.

Revelation: Separating the Signal from the Noise

> *As we chip away the unnecessary information, "the noise", we attempt, too, to find or impose patterns on what remains: Does it gestalt? If it does not, we keep nudging here and there, because we know that the information has a shape, and all we have to do is find it.*
>
> Gene Dilmore (Technology of Information Processing and Data Basing)
>
> [Note: *gestalt* means a configuration, form, or pattern that, as a unified whole or functional unit, has properties which cannot be derived by summation of the separate parts. *Webster's Dictionary*]

Our first step in exploratory data analysis was reducing the data to more manageable statistical or graphical summaries. Once the data is in a comprehensible form, then we can look for patterns to reveal themselves. Knowing the type of data under examination, we look for expected patterns and for striking deviations from those patterns. We then seek explanations within the problem's context.

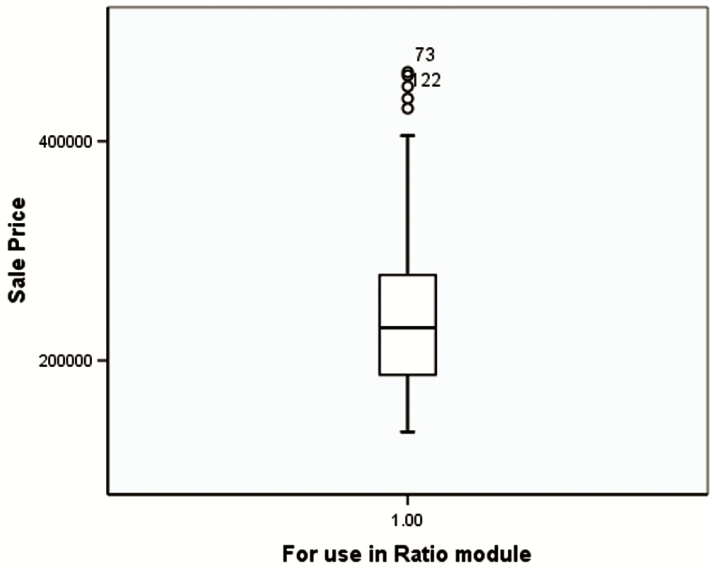
Before looking at techniques for pattern discovery, first a few words on data structure. We talked earlier in this lesson about the concept of explainable and unexplained variation in a variable – for example, sale prices for real estate, where there is some rational variation we can investigate and a non-rational element that is pretty much random. In data exploration, this can be called separating the "signal" from the "noise". Think of scientists monitoring radio signals from space, scanning millions of hours of "white noise" hoping to find some kind of pattern. A similar, but hopefully less daunting, process goes on in all data analysis – when faced with a large array of numbers, the analyst must try to separate out the random elements and find the patterns that contribute to solving the problem of interest.

In technical terms, real estate price data would typically be considered *stochastic* or *non-deterministic*. This means there is a random element in it which precludes fully explaining all variation with 100% accuracy. For example, you can use past sales transactions to partially forecast future prices, but never with complete accuracy. In contrast, *deterministic* events are those completely explained by existing or past causes and with no uncertainty whatsoever – e.g., when a tree releases an apple, gravity causes it to fall to the ground.

Another related distinction in data is its *smooth* and *rough* components. The smooth properties in data are those explained by measures of central tendency, dispersion, and distribution. The rough elements are those outside these smooth considerations, such as the residual between a given data point and the line of best fit or, at the extreme, outliers that are outside the range of reasonableness. Smooth elements are analyzed using the descriptive statistics discussed in the "Reduction" section; rough elements are analyzed using features that showcase outliers, such as boxplots, scatterplots, and histograms.

Let's examine the boxplot as our first revelation technique. Lesson 2 explained the basic features of a standard boxplot graphic. Consider the following boxplot of S_PRICE:[1]

- Graphs → Legacy Dialogs → Boxplox… → Simple, Define
- Enter S_PRICE in the Variable field and ONE in the Category Axis field, then click OK.



The "box" in the boxplot encloses 50% of the data, between the first and third quartiles. Each quartile represents 25% of the distribution, with the second quartile being the median value. The median, or 50% percentile, is shown by the horizontal line in the box. Extending from the box are "whiskers" showing the limits of the data's distribution, except for outliers. Outliers may be shown as circles, dots, asterisks, or numbers showing which

[1] SPSS requires a "category variable" to produce a boxplot – in other words, sale price by another variable. We have specified the category variable as "One", a variable that is equal to one for all observations.

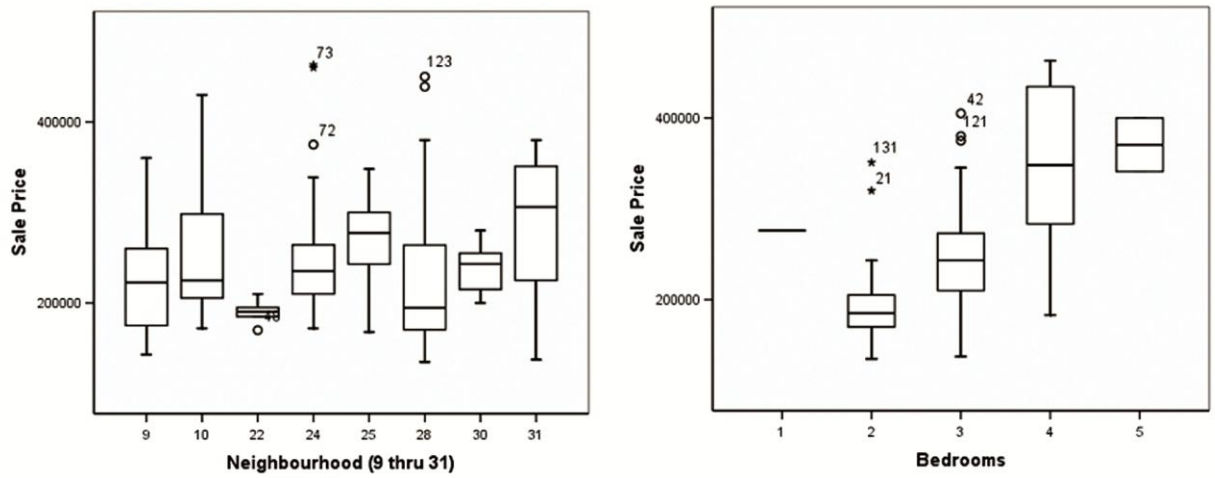
case is an extreme value. Boxplots can also indicate the skewness of the data, shown by the relative orientation of the median relative to the quartiles.

So far in this lesson we have only considered single variables, or *univariate* measures. We will expand our analysis to include *bivariate* (two variables) and/or *multivariate* relationships.

Boxplots are most useful for a continuous variable such as Price against a discrete variable where there are only a few possible values: e.g., bedrooms (number), presence of pool (yes/no), neighbourhood (number or letter code). The boxplot will very quickly show you whether the groups are equal or whether there are significant variations.

Consider the boxplot of S_PRICE by NBHD and by BEDROOMS:

Use Dialog Recall for the previous boxplot. Change the ONE to NBHD, and click OK. Open the boxplot module once again and change NBHD to BEDROOMS and click OK.

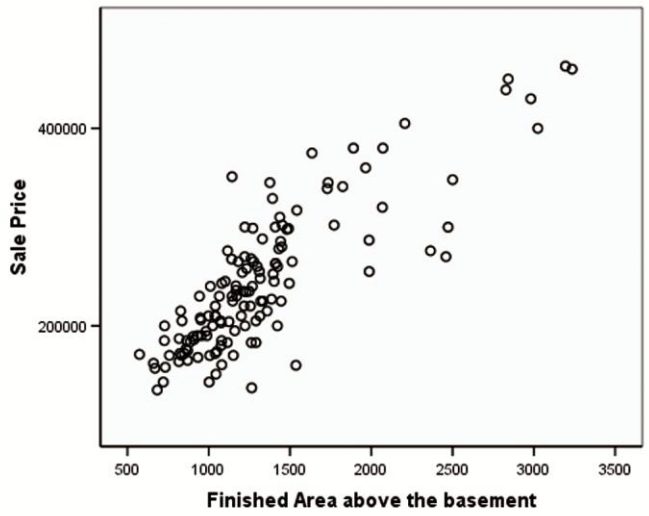


The left boxplot shows how neighbourhoods vary in prices. It appears that NBHDs 9 and 31 are quite different. It also appears that NBHDs 24 and 28 have some high priced outliers. These should be investigated further, as they could indicate a data error or perhaps sales are problematic in some way and might need to be removed from the analysis.

The right boxplot shows how prices vary by number of bedrooms in the home. As expected, the more bedrooms, the higher the price – the boxes barely overlap at all, indicating a significant variation. However, note how much overlap there is between the whiskers – amidst this general trend, there is still a lot of variation between individual properties. Hence the power of viewing this graphically!

Boxplots are not effective for analyzing two continuous variables, e.g., S_PRICE against FINAREA, as there would be too many boxes to analyze with one for every square footage in the database. For this, a scatterplot is more effective. As discussed in Lesson 2, a scatterplot measures the graphic relationship between two variables (X, Y), by locating each X-Y coordinate on a two-dimensional graph. Consider the scatterplot of S_PRICE by FINAREA:

- Graphs → Legacy Dialogs → Scatter/Dot… → Simple Scatter, Define
- Enter S_PRICE in the Y Axis field and FINAREA into the X Axis field
- Click OK.

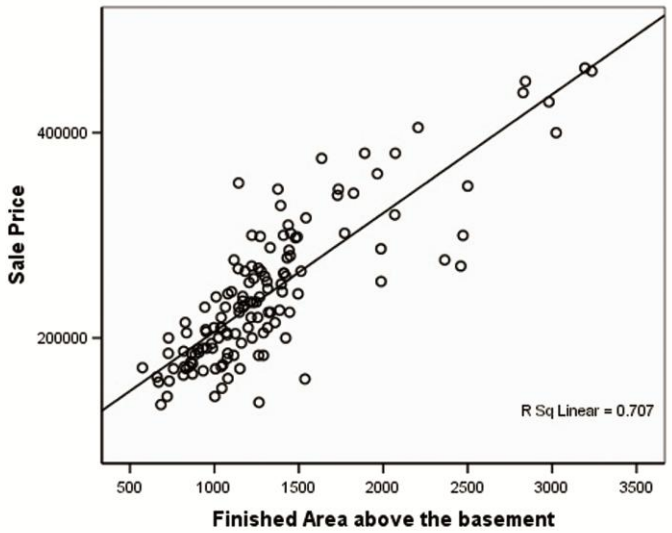


This shows data spread from approximately 600 square feet and $150,000 up to 3,300 square feet and $450,000. As expected the data appears to be upward sloping, indicating that larger houses tend to sell for more. But, again, note the wide variation for this trend. If you were interested in one value of the X variable, FINAREA, you could take a "slice" upward from, say, 2,000 square feet.

As explained in Lesson 2, the relationship in scatterplots can be made even clearer by fitting a line or a curve to the data. This is called *smoothing* the data, or in other words explaining the "smooth" element – again, the "rough" element being how the data points vary from this line.

The easiest and most common of these is the least squares regression line. This will indicate the correlation, or strength and direction of linear relationship between two variables. Revisiting Lessons 1 and 2, correlation is measured on a standardized scale between -1 and +1; these indicate a perfect relationship. 0 indicates no linear relationship.

See the S_PRICE by FINAREA graph below. Double-click on the chart, click "Add Fit Line at Total" in the Chart Editor window, close the Chart Editor.



This indicates a strong linear relationship.

Resolution problems in graphs may mask patterns in data. For example, in some situations, outliers may need to be removed in order to see a relationship otherwise difficult to decipher. If there was one very large house in this analysis, say a 20,000 square foot mansion that sold for $14 million, this might distort the analysis and make the trend line difficult to decipher. In this case, you could re-run the scatterplot with a filter (Data → Select Cases… → If condition is satisfied → If) set to "FINAREA < 5000". This would cause SPSS to filter out, or ignore, the mansion (filters will be discussed in more detail later in this lesson).

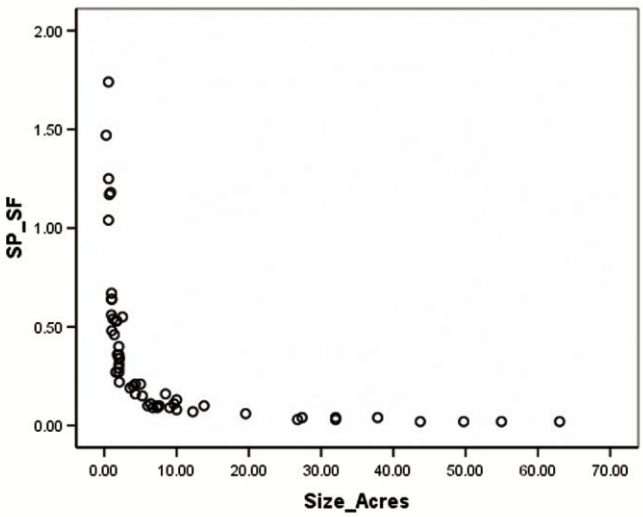
Alternatively, you may have a problem with a non-linear relationship. The examples above have all focused on linear relationships between variables, but not all variable relationships are linear. For example, land price per square foot often shows "decreasing returns to scale", meaning that as parcel size increases, the value per square foot drops. This is a non-linear function and this expression may require transforming the data. This leads to our next section on re-expression of data.

Re-Expression: Transforming the Data

The section above discussed scatterplots as a method for revealing patterns in data. However, in some cases the data is not in an optimal format for this form of expression and requires transformation. In particular, where there are non-linear relationships, the data may need to be re-expressed to facilitate analysis.

Continuous data, including real estate data, often displays a non-linear function. In order to fit this to a linear model, the data may need to be transformed. When you view a scatterplot of the data, if the graph(s) suggest the link between X and Y is not linear, either X or Y or both can be transformed (re-expressed) by their logarithmic functions (e.g., natural log or log base10),[2] different powers, their square root, or their inverse. By transforming the variable, the resulting graph may show a linear relationship. This is best explained through an illustration: we will show how graphical analysis can be used to substantiate a size adjustment.

The relationship between price and size in real estate market data can take many different forms, and these patterns can be discovered and highlighted through the use of graphics. If the value-size relationship shows a linear pattern, where price increases proportionately with size, then no adjustment is necessary – you may simply use a variable such as value per square foot or per front foot. However, the prevailing pattern found in most land markets is that total price increases with size, but at a diminishing rate. Economists call this a concave function. For example, see the scatterplot below, showing land sales on sale price per square foot versus size in acres:



[2] If you have trouble understanding natural logarithms intuitively, the following article is easy to understand and even fun! "Demystifying the Natural Logarithm (ln)" available *at http://betterexplained.com/articles/demystifying-the-natural-logarithm-ln/*

As you can see, the relationship of land price per square foot and parcel size is clearly curved. This is an example of how a graph can convey, in an understandable manner, the non-linear functional form of the typical size-price effect. Common forms of non-linear functions are illustrated in Chapter 3 of the *Advanced Computer-Assisted Mass Appraisal* book. This can be viewed in "Online Readings" on the course webpage.

Re-expressing the size-unit price variables by their natural logs linearizes the relationship. This will be illustrated briefly in a case study later in this lesson and in more detail in Lesson 5.

Residuals: We Missed by How Much?

"All models are wrong, but some are useful"
George E. P. Box

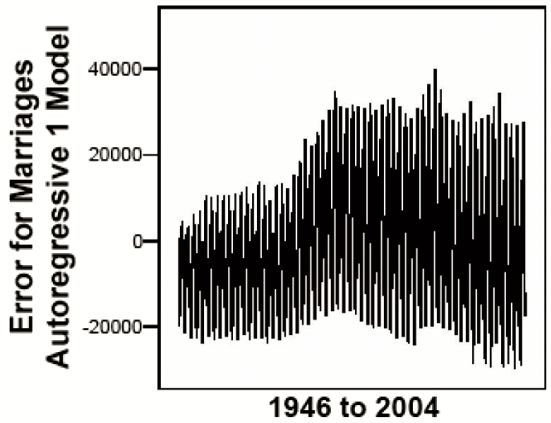
Real estate appraisal is a form of model building, where past sales data are used in an attempt to estimate the market value of a property that has not sold (the subject). We employ models to mimic reality, so that we can use past or current data to estimate or forecast unknown values. Because they are an approximation, models will never be 100% accurate (and in fact, if they were 100% accurate, that in itself would probably indicate a mistake!). An important aspect of model building, then, is to search our results for patterns in our "mistakes".

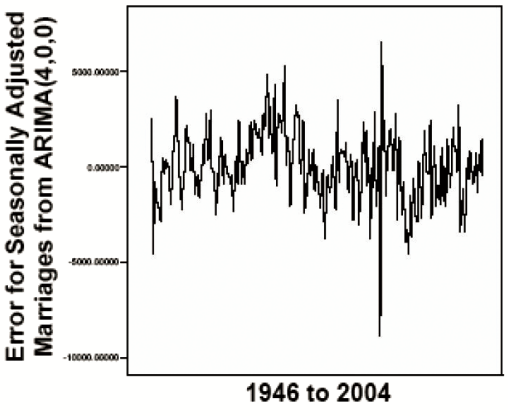
Estimated or predicted values from a model will always differ from observed values. In real estate valuation, for example, the market value estimate will rarely be the same as the actual selling price; in a mass appraisal for assessment purposes, the assessed values on the whole will vary above and below their observed selling prices for properties that sold near the valuation. These differences are called *residuals*.

As discussed earlier, data has two parts, systematic and random variation (or signal plus noise). Our goal as modellers is to minimize systematic variation. The leftover "noise" can be from measurement error, omission of important variables, or purely random unexplained variation.

Before a valuation model is used to address questions about the relationship between price and predictors, the fit of the model to the data should be assessed. This is the diagnostic part of fitting models. When analyzing "fit", this can be as simple as examining the resulting values against intuition (i.e., a common sense review) or can involve detailed statistical analysis.

One important diagnostic tool for the regression modeller is the residual plot. This graphs the "errors" (residuals) between the model predictions and the actual observed values. Consider the following forecasting example, showing residual plots from two different models attempting to predict the number of marriages in Canada.





The errors/residuals in the graph on the left increase from the beginning of the data to the end, with the variance getting wider with each successive year. This is a problem called *heteroskedasticity*. If the residual graph shows heteroskedasticity, then you should change the way you measure your variable. Usually the problem is that the errors get larger over time, partly because the size of the variable gets larger over time – for example, real estate prices tend to increase over time, so a forecast of future real estate prices will have errors that become more and more magnified. The way to fix this is to transform the variable so that it does not grow over time. Mathematically, a function called the natural logarithm is the ideal solution for any numbers that grow exponentially over time – you will need to re-express your data in a format that works.

The graph on the right shows errors with a constant size as we move from left to right in the graph, but with one spike. In this example, the marriage data required a seasonal adjustment to account for more weddings in summer than winter.

Another related diagnostic tool is the partial regression plot (also referred to as added variable plots, adjusted variable plots, and individual coefficient plots). When performing a linear regression with a single independent variable, a scatterplot of the response variable against the independent variable provides a good indication of the nature of the relationship. However, if there is more than one independent variable, things become more complicated. It is useful to generate scatterplots of the response variable against each of the independent variables, but this does not take into account the effect of the other independent variables in the model. Partial regression plots attempt to show the effect of adding an additional variable to the model given that one or more independent variables are already in the model.

Finally, outlying data points merit special attention. Outliers can mask important trends by reducing the resolution of a plot. They can also significantly influence the results of an analysis. We must consider why any outlier value is extraordinary before deciding to keep it or discard. In other words, we must weigh the benefits against the costs: on the one hand, by removing an outlier we may be able to better identify the relationships in the data, but we must be careful not to "over-manage" our data. At what point does our refinement become manipulating the data to meet our pre-conceived conclusions? Removing outliers is a very common necessity in modelling, but it must be done critically and with care.

The Four Rs: A Final Word

Each case of exploratory data analysis is different; every set of data is different. In fact, two different sets of sales data drawn from the same city during the same time frame may tell completely different "stories". At the outset, you do not know what you will find. Rest assured that once you have developed some experience in data analysis the four Rs will run together into a smooth process.

One word of warning which will be touched on in the next five lessons: watch for outliers; data which simply does not fit with the rest of the information you have…a living area for a condominium of 100,000 square feet or a lot size of 0.25 square feet (the first likely has too many zeros and the second may be the result of entering the number of acres instead of square feet). You can discover these data problems at any stage of your analysis. The earlier you find them the better, as they can easily taint any analysis if discovered too late, and restarting may be the only option.

We will now continue this lesson with six case studies which demonstrate the Four Rs.

Exploratory Data Analysis Case Studies

Case Study 1: Data Reduction with Summary Statistics and Histograms

Helpful Hint!

We will be using several databases in this lesson and throughout the course. You can download these from "Online Readings" on the course website. If you wish, you may download each file separately as you progress through the course. Alternatively, you can save time by instead downloading the entire course's data all at once using the "zip" files.

This case uses the dataset titled "Industrial".[3] This dataset provides the indicated price per square foot for 20 sales of industrial properties in four different markets (Market 1, Market 2, Market 3, and Market 4). The data is summarized in the following table:

Obs No.	Market 1	Market 2	Market 3	Market 4
1	44.30	45.20	36.70	41.70
2	40.00	46.20	41.40	42.20
3	43.40	39.00	40.50	40.70
4	41.50	45.00	37.60	39.30
5	52.00	47.80	39.60	43.30
6	32.00	46.60	44.00	41.50
7	38.00	36.20	40.20	38.30
8	40.60	46.90	52.90	41.90
9	39.70	37.10	49.50	41.80
10	39.00	49.30	38.30	42.10
11	29.50	34.70	45.90	49.90
12	43.00	37.40	48.70	39.60
13	44.00	37.80	37.20	42.30
14	42.50	50.20	35.80	34.10
15	46.00	38.50	38.10	40.10
16	54.50	50.40	40.20	38.50
17	47.00	38.80	43.80	42.50
18	41.00	33.60	51.70	39.40
19	37.00	45.50	31.10	63.00
20	45.00	33.80	46.80	37.80

Viewing the raw data does not help us much in understanding the patterns in the data or implications for their application. To get a sense of the range of values in each market, we could re-sort the data in each column in ascending order. The following illustrates the range in sale prices in Market 1:

[3] This is an artificially constructed dataset, based on data from *Statistical Graphics for Univariate and Bivariate Data* by William G. Jacoby, Sage Publications, 1997. While the data are not real estate focused in reality, they are consistent with the expected values for industrial properties and will be used for this purpose.

Obs No.	Market 1
1	29.50
2	32.00
3	37.00
4	38.00
5	39.00
6	39.70
7	40.00
8	40.60
9	41.00
10	41.50
11	42.50
12	43.00
13	43.40
14	44.00
15	44.30
16	45.00
17	46.00
18	47.00
19	52.00
20	54.50

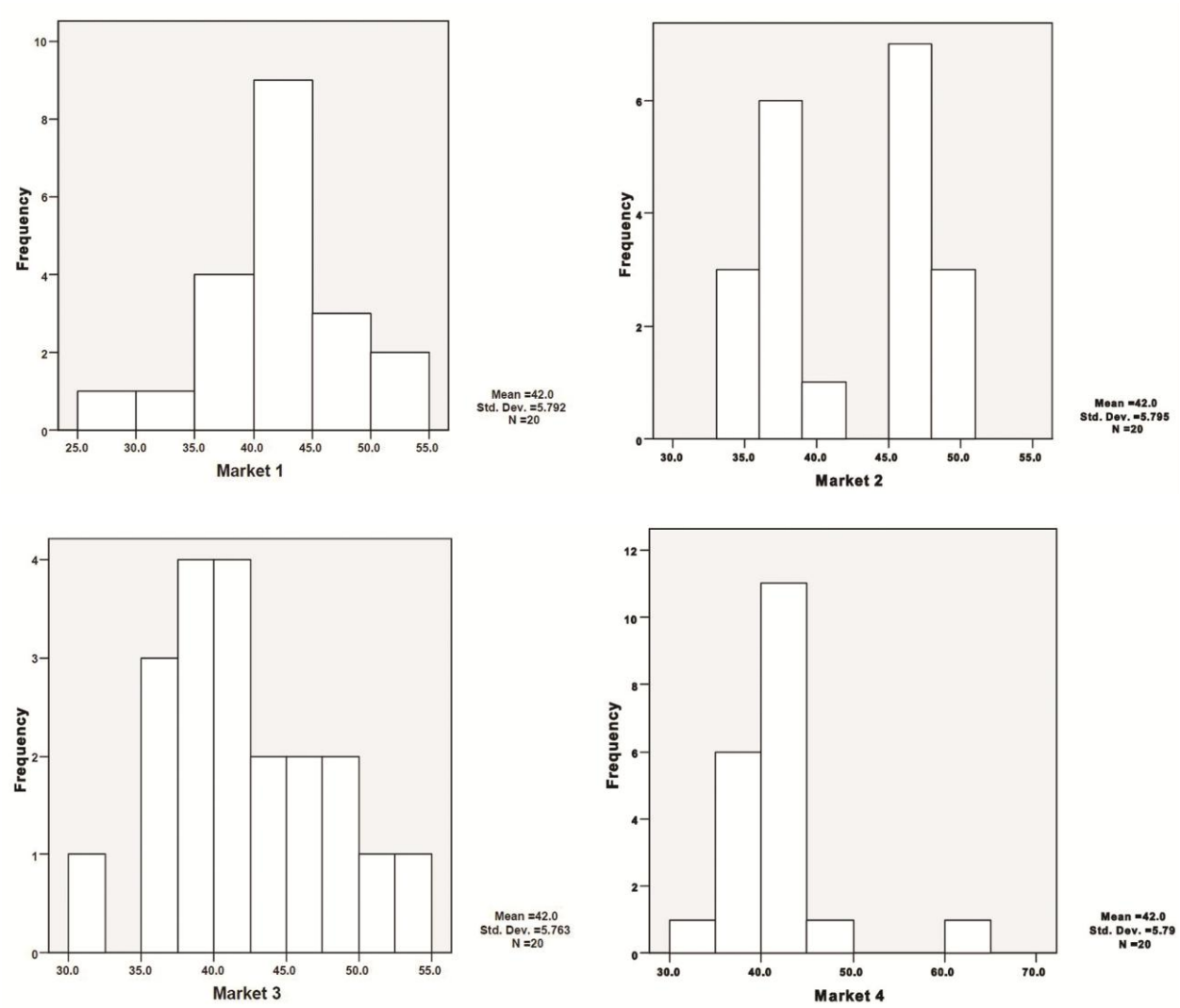
A quick scan of the sorted sale price per sq. ft. values for industrial buildings in market 1 shows a range of prices between $29.50 and $54.50, with the bulk of values falling in the $30s and $40s. A visual inspection of the numeric values, however, probably does not lead to any further useful insights into this or the other three markets represented by variables Market 1 to Market 4.

Data reduction can help overcome the shortcomings of a visual scan, for example by calculating summary statistics. Summary statistics for the four market variables are in the table below (calculated in SPSS using "Descriptives"):

		Market1	Market2	Market3	Market4
N	Valid	20	20	20	20
	Missing	0	0	0	0
Mean		42.000	42.000	42.000	42.000
Median		42.000	42.000	40.350	41.600
Std. Deviation		5.7925	5.7950	5.7629	5.7901
Range		25.0	16.8	21.8	28.9
Minimum		29.5	33.6	31.1	34.1
Maximum		54.5	50.4	52.9	63.0
Percentiles	25	39.175	37.175	37.725	39.325
	50	42.000	42.000	40.350	41.600
	75	44.825	46.825	46.575	42.275

You will notice that the ranges of the price distributions for each market variable are different, but the mean and standard deviation statistics are virtually identical. By relying on these two statistics of central tendency spread, it is tempting to conclude that the four industrial property markets are the same.

To confirm this, let's take a visual look at these four markets by plotting a histogram of each variable and viewing the shape of the distribution for the four industrial property markets.



The distribution for Market 1 is symmetrical and could possibly approximate a normal bell-shaped distribution. The mean and standard deviation statistics, therefore, are probably good estimators of the centre and spread of this market.

Market 2's distribution shows a distinct bimodal shape. The mean and standard deviation statistics of $42 and $5.80 per sq. ft., respectively, are not good estimators of the centre and dispersion of this market. This market segment may actually be comprised of two different markets.

The prices in Market 3 show a positive (to the right) skew. The median, at about $40 per sq. ft., is a better representation of the centre than the mean.

The sale price distribution of market segment 4 also show a positive (right) skew, but possibly due to a single outlying observation. If this outlier is removed (by setting a filter, e.g., "If Market4 < 60"), then the distribution looks more normal.

Helpful Hint!

To set this filter in SPSS, follow these steps:

- Data → Select Cases... → If condition is satisfied, click If...
- In the text box, type (or use the mouse to point and click) Market4 < 60, click Continue
- Click OK

In the Data View, you should now see one record number (on the far left hand side) with a slash through it – this is the record that is now "filtered out" because its Market 4 value is greater than or equal to 60. This record will be ignored in any analysis or chart completed from now on.

To turn off a filter (it is very important to remember to do this when you do not need the filter any longer), follow these steps:

- Data → Select Cases... → All Cases → OK

All records will now be used.

Data reduction is typically accomplished by statistical summaries, but these are based on assumptions of the variable's distribution structure that may or may not hold, as shown in this case study. Although the mean and standard deviation statistics obtained for Market 1 are good representations of that market's price distribution, numerical summaries by themselves clearly misrepresent the other three markets. Graphical representations of these markets are easily generated and clearly showcase their differences.

Case Study 2: Reduction and Revelation Using Graphs and Correlation

This case uses the dataset "Vancouver",[4] providing hypothetical square footage and monthly rent for 18 apartments in downtown Vancouver. The data is summarized as follows:

Case No.	Building	SqFt	Rent
1	800 Burrard	530	915
2	800 Burrard	830	1115
3	800 Burrard	1230	1680
4	550 Robson	706	1316
5	550 Robson	775	1449
6	550 Robson	886	1623
7	550 Robson	967	1692
8	550 Robson	1053	1811
9	550 Robson	1130	2114
10	550 Robson	1346	2357
11	370 Howe	678	1219
12	370 Howe	730	1633
13	370 Howe	901	1963
14	370 Howe	1147	2223
15	370 Howe	1254	2349
16	370 Howe	1445	2654
17	370 Howe	1564	3281
18	370 Howe	1705	3426

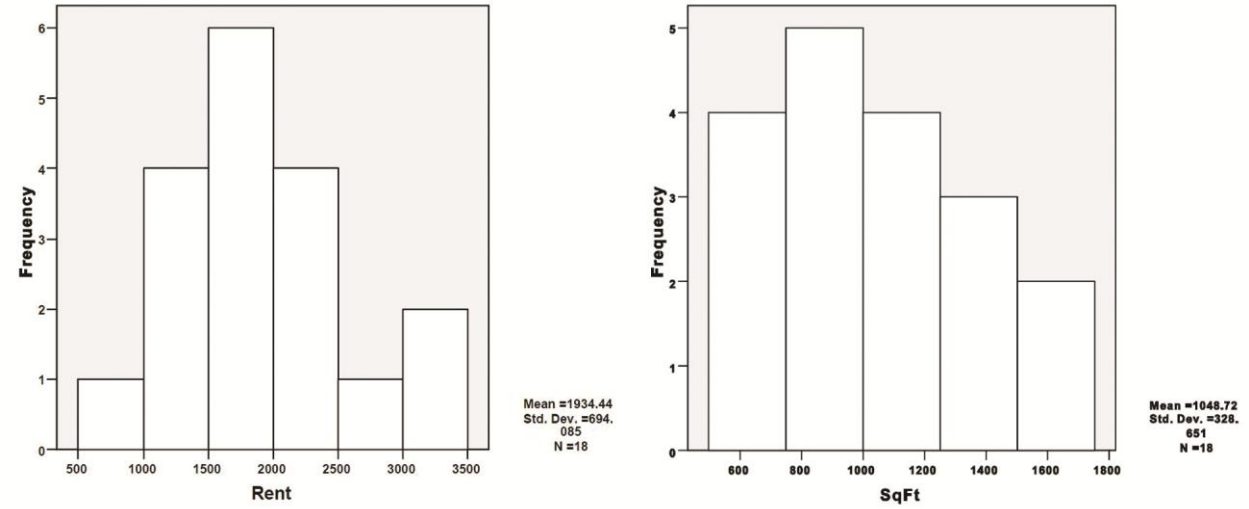
[4] This data is loosely based on a study of Chicago's North Shore luxury apartment market by student members of the real estate club at the Kellogg School of Management.

The first data reduction strategy is to obtain some basic statistics of the centre and dispersion of the variables to be analyzed:

		Rent	SqFt
N	Valid	18	18
	Missing	0	0
Mean		1934.44	1048.72
Median		1751.50	1010.00
Std. Deviation		694.085	328.651
Minimum		915	530
Maximum		3426	1705

The living area of the apartments have a low of 530 sq. ft. and a high of 1,705 ft. The rents have a wide range, between $915 and $3,426. As the median rent of $1,751 is significantly lower than the average rent of $1,934, we suspect that the distribution of rents may be skewed to the right.

It is useful to take a look at histograms to visually observe the shape of the distribution.



The histogram of the rent variable confirms that the distribution of rent is indeed skewed to the right – towards the higher rents. Living area is also somewhat skewed to the right, but more normally distributed than rent.

We have also produced a correlation matrix to illustrate the relationship between variables. Correlation analysis shows how specified variables are correlated with each other. Go to Analyze → Correlate → Bivariate, and select Rent and SqFt as the variables. Ensure that "Pearson" is the Correlation Coefficient selected and click OK. The following correlation matrix results:

		Rent	SqFt
Rent	Pearson Correlation	1	.940(**)
	Sig. (2-tailed)		.000
	N	18	18
SqFt	Pearson Correlation	.940(**)	1
	Sig. (2-tailed)	000	
	N	18	18

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

The correlation coefficient between rent and living area in square feet is expressed as 0.94, a very strong relationship. The table also shows, as expected, the correlation coefficient between the rent variable and itself is 1 or perfect as the values are identical.

Helpful Hint!

The correlation matrix produced in SPSS shows three lines for each variable, the correlation, the significance, and the N. When producing these, we only care about the correlation, in a matrix of 10 or more variables, these extra lines for significance and N can make the matrix very cluttered. As well, readers often confuse the significance number with the correlation, so be careful you are reading the correct number. You can remove the two extraneous rows from the correlation matrix if you wish. To do so double click on the correlation matrix and for the Sig. and N rows do the following:

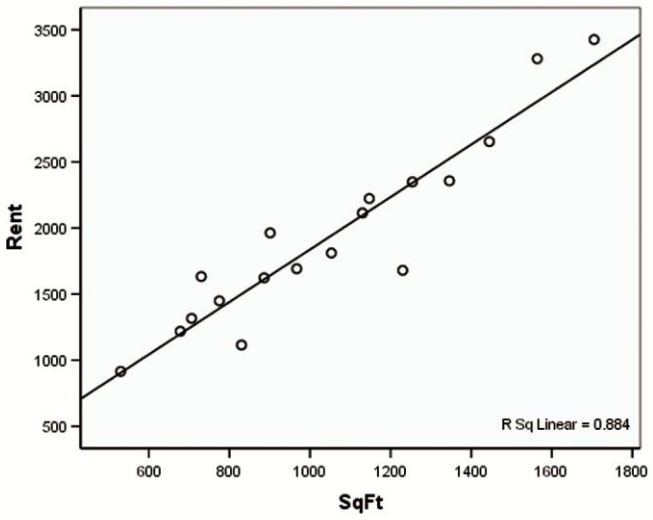
- select the desired row (click on either Sig. or N)
- from the menu, click Edit → Select → Data and Label Cells
- from the menu, click Edit → Delete (the selected row will be deleted for the entire table)

This simplified table is shown below:

		Rent	SqFt
Rent	Pearson Correlation	1	.940(**)
SqFt	Pearson Correlation	.940(**)	1

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

Although the correlation coefficient between two variables is a useful statistic for getting a fix on the strength of a relationship, its limitations for measuring linear relationships can be very misleading. It is always a good idea to also look at a scatterplot for visualizing the relationship between two variables. The scatterplot below, which plots the 18 observations of apartment rents against their respective living areas, more accurately displays the nature of the strong linear relationship between these two variables that was indicated by their correlation coefficient.



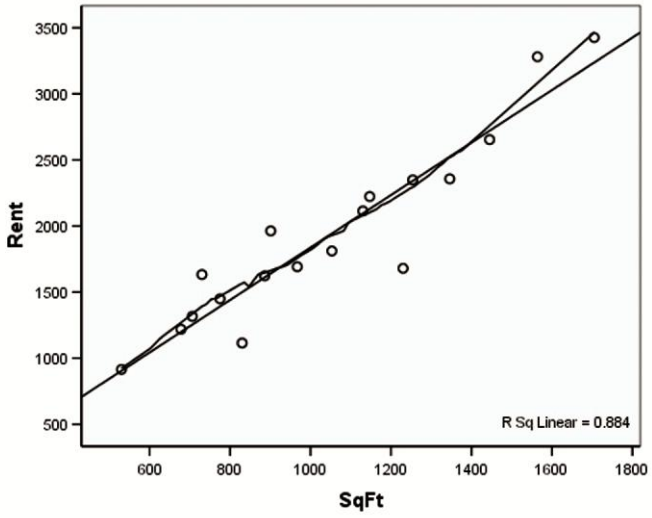
The scatterplot shows that rent has a strong positive relationship with living area. As the size of the living area increases, the rent also increases. This relationship also appears to be linear in form, which means that the rate of increase is constant.

As an aid to assessing linearity, we have imposed a best fit regression line, based on least squares. When we fit a straight line to the data, we are automatically making a judgment about the structural form of the data.

We can see there are several data points that lie below and to the right of the line that are somewhat inconsistent with the rest of the data points. This implies there is curvature in the data, that could possibly indicate a non-linear relationship.

In SPSS, we can fit a "Loess" fit line to see how closely the data reflects a linear relationship. A Loess fit line is comprised of a series of mini-regressions along the full range of the data values. Whereas the single regression line fits all of the data, the Loess line examines the pattern found in local areas within the data. The SPSS steps are as follows:

- Double-click on the chart to open the "Chart Editor".
- Select Elements → Fit Line at Total.
- Select "Loess". The "% of points to fit" box will determine the distance between the "mini-regressions", with 99% being very close to the linear fit. You may leave the % at the default 50.
- You may want to select File → Save Chart Template, in order to easily add "Loess" lines to other scatterplots later; e.g., save template as "loess50", for "Loess line set at 50%".
- Click Apply → Close → X in top right corner to obtain the graph below:



The Loess curve shows there is some curvature in the data, but that this curvature is not overly significant. Therefore, a straight line, indeed, appears to fit this data well.

We see there are some outliers that stand out from the rest of the data. In reviewing the data, we find that most of these cases are in the 370 Howe building. Depending on the purpose of our analysis, we might want to investigate the 370 Howe building to see what features it had that might cause these rents to deviate from those obtained in other buildings. If desired, we could set a filter to eliminate these cases from the analysis and see how they affected the analysis. The analyst can run "what if" simulations with the data to visually discover the sensitivity of the observed relationships to particular data points that do not seem to fit.

Helpful Hint!

If you wanted to quickly know which observations are the outliers in this graph, in SPSS you could double click on the graph to open the Chart Editor, click the "Data Label Mode" icon (looks like a square target or bullseye), and then click on the observations of interest. Their case number will appear. You can then click on the "Go to Case" icon (looks like the end of a ruler) or you click on the "Data View" tab and review the data window manually.

Our graphical investigation of the relationship between rents and area using scatterplots shows that a straight line relationship is appropriate for forecasting apartment rents in this market. However, it also highlighted several cases in one building that do not appear to be consistent with the other observations and perhaps should be investigated further. Our key result, though, is showcasing the scatterplot as a workhorse graphic for discovering the nature of relationships between two variables.

Case Study 3: Revelation Using Correlation and Scatterplots

This case uses the dataset "officerent",[5] providing office rent observations that occurred on various floors in four different office buildings A, B, C, and D. The data is summarized as follows:

Rent_A	Floor_A	Rent_B	Floor_B	Rent_C	Floor_C	Rent_D	Floor_D
8.04	10	9.14	10	7.46	10	6.58	8
6.95	8	8.14	8	6.77	8	5.76	8
7.58	13	8.74	13	12.74	13	7.71	8
8.81	9	8.77	9	7.11	9	8.84	8
8.33	11	9.26	11	7.81	11	8.47	8
9.96	14	8.1	14	8.84	14	7.04	8
7.24	6	6.13	6	6.08	6	5.25	8
4.26	4	3.1	4	5.39	4	12.5	19
10.84	12	9.13	12	8.15	12	5.56	8
4.82	7	7.26	7	6.42	7	7.91	8
5.68	5	4.74	5	5.73	5	6.89	8

We can visually inspect the numeric values, but scanning such a large table of numbers is not likely to lead to any useful insights into the relationships that exist between rent and floor level in each of four office buildings. However, one initial observation that jumps out is that the rents obtained in the dataset for Building D are all on the eighth floor except for one case that is on the 19th floor. For this building there is hardly any variation in floors, with 10 of the 11 observations on one floor. Conversely, the rents per square foot vary more widely. This suggests that rents are not meaningfully related to the floor level in this particular building. Therefore, we might come to the conclusion that dataset D is unlike datasets A to C.

[5] This is an artificially constructed collection of four datasets of 11 cases in each dataset. F.J. Anscombe invented these datasets to demonstrate the importance of graphing the data before finding the correlation and best fit regression line between the variables. They were first published in "Graphs in Statistical Analysis", *American Statistician*, 27 (February 1973), pp. 17-21.

We will employ our first reduction strategy, by more concisely describing the data using summary statistics. Below are summary statistics and correlations for all the variables:

	N		Mean	Std. Deviation	Minimum	Maximum
	Valid	Missing				
Floor_A	11	0	9.00	3.317	4	14
Floor_B	11	0	9.00	3.317	4	14
Floor_C	11	0	9.00	3.317	4	14
Floor_D	11	0	9.00	3.317	8	19
Rent_A	11	0	7.5009	2.03157	4.26	10.84
Rent_B	11	0	7.5009	2.03166	3.10	9.26
Rent_C	11	0	7.5000	2.03042	5.39	12.74
Rent_D	11	0	7.5009	2.03058	5.25	12.50

Correlations (Note: the N and Sig lines have been removed for clarity)

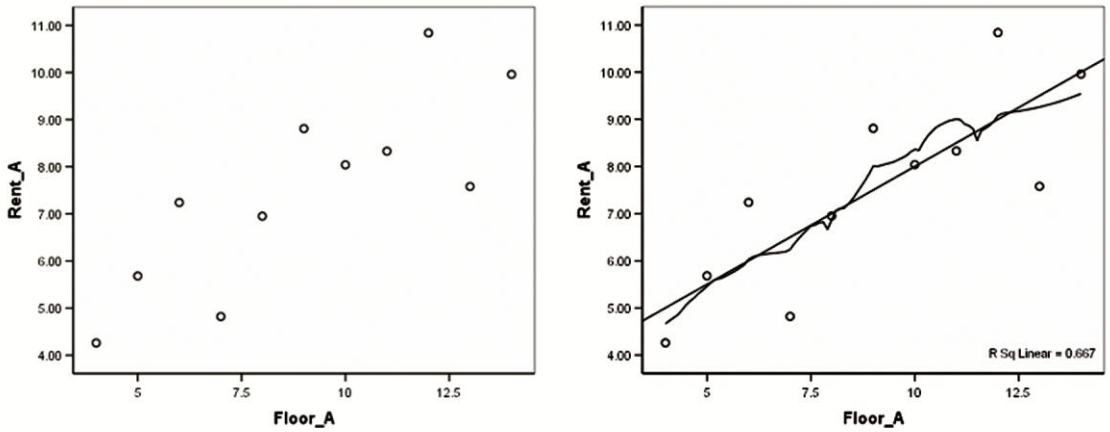
	Floor_A	Floor_B	Floor_C	Floor_D	Rent_A	Rent_B	Rent_C	Rent_D
Floor_A	1	1.000(**)	1.000(**)	-.500	**.816(**)**	.816(**)	.816(**)	-.314
Floor_B	1.000(**)	1	1.000(**)	-.500	.816(**)	**.816(**)**	.816(**)	-.314
Floor_C	1.000(**)	1.000(**)	1	-.500	.816(**)	.816(**)	**.816(**)**	-.314
Floor_D	-.500	-.500	-.500	1	-.529	-.718(*)	-.345	**.817(**)**
Rent_A	**.816**(**)	.816(**)	.816(**)	-.529	1	.750(**)	.469	-.489
Rent_B	.816(**)	**.816**(**)	.816(**)	-.718(*)	.750(**)	1	.588	-.478
Rent_C	.816(**)	.816(**)	**.816**(**)	-.345	.469	.588	1	-.155
Rent_D	-.314	-.314	-.314	**.817**(**)	-.489	-.478	-.155	1

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

Just as we discovered in Case Study 1, the ranges of the price and floor distributions for each dataset are different, but the mean and standard deviation are identical. The average floor level of the observations in each dataset is 9 and the average rent per square foot is $7.50. Standard deviation is 3.3 floors and $2.03 per square foot for the floor and rent variables, respectively.

The correlation coefficient between rent per square foot and floor is 0.8 (rounded) for each of the four datasets in the table above. These values are bolded to make them easier to find within the matrix. The linear relationship between rent per square foot and floor level in each dataset is very strong, but it is peculiar that they are identical for each dataset. Our initial examination of the data values in the table suggested that dataset D was unlike the other three datasets. Yet, the correlation statistic suggests the very opposite. We will use graphical tools to investigate this mystery.

We will first plot the rent variable against the floor variable of dataset A. The floor level variable will be placed on the horizontal axis (X-axis) and the rent per square foot will be placed on the vertical axis (Y-axis). We first show the observations alone then with a linear line and Loess line imposed.



On the left, we see quite clearly that rent per square foot increases with floor level. This relationship also appears to have a linear pattern, with rent per square foot increasing steadily towards upper floors. We confirm this with the graph on the right, adding the simple regression line between the two variables. We can also examine if this data fits a linear relationship. The Loess line fit line detects some curvature in the function between the two variables in the upper floor range, but overall the linear fit seems appropriate.

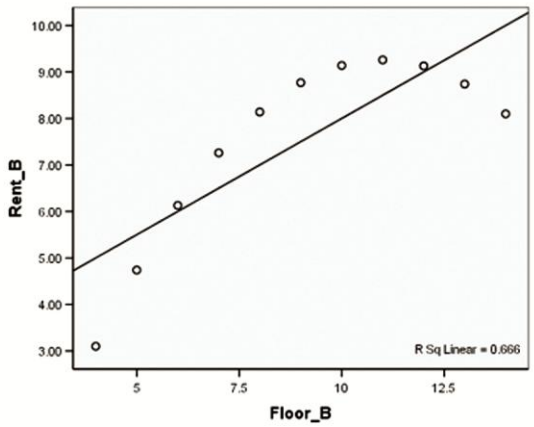
The mathematical expression of this regression line[6] is Rent_A = 3.0 + 0.50 × Floor_A. This means rent per square foot increases by $0.50 by going up each additional floor. By inserting different values for the floor, we can use this model to predict for rent per square foot.

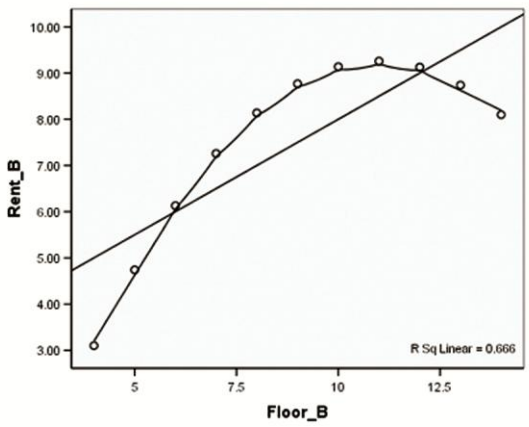
SPSS also provided the R-square statistic, explaining how much of the variation in rent is explained by floor number. Our simple model with one explanatory variable, floor level, explains 67% of the variation in rent per square foot.

Remember our correlation coefficient of 0.816? The R-square statistic from the regression line of Rent_A and Floor_A is simply the square of the correlation coefficient (0.816 × 0.816 = 0.666). You may also recall that the correlation coefficient matrix we obtained for all four datasets indicated they all had an identical coefficient of approximately 0.82. This means that if we relied solely on the R-square measure to tell us how good our model was, we would have to conclude that regressions fit to all four datasets would have an identical goodness of fit.

Let's examine how reality fits with this conclusion. Our next graph is a scatterplot of the second dataset, Rent_B against Floor_B with the simple regression line applied and then a Loess line to show any curve. We anticipate the fit we obtain will be similar to the first dataset because we know they have identical correlation coefficients.

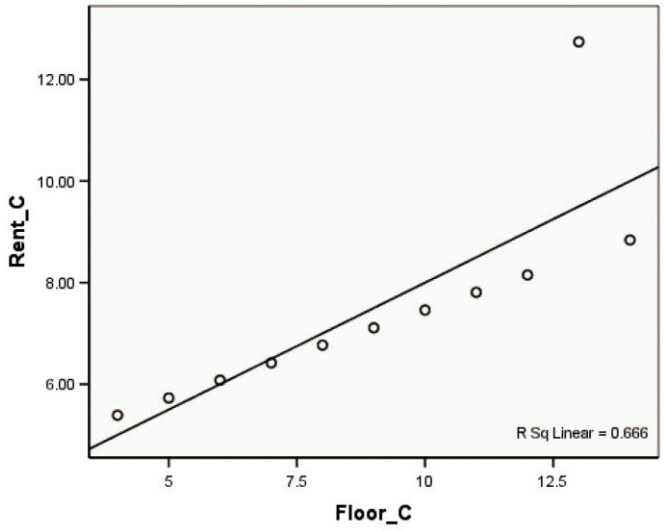
[6] In SPSS, you can find this by selecting Analyze →Regression →Linear, with RENT_A as Dependent Variable and FLOOR_A as Independent Variable.



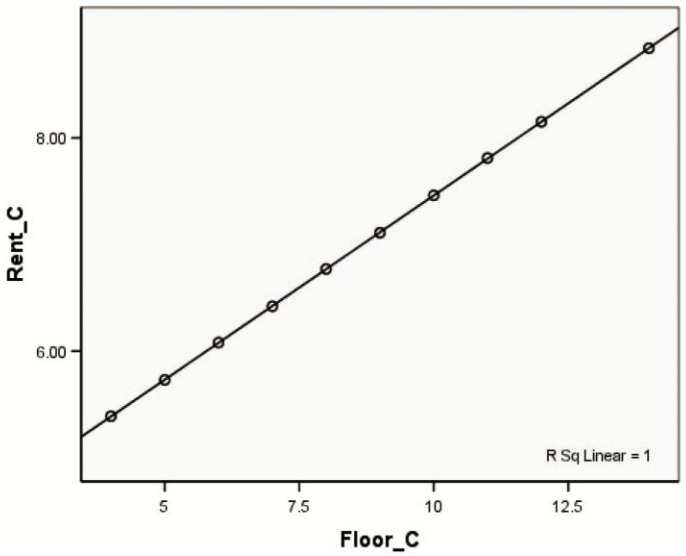


The results are surprising: we can plainly see that rents per square foot initially rise with floor level, then stabilize, and then start to decrease on the highest floors. This is a definite pattern, but certainly not linear – a straight line does not represent this data well at all. Yet, if we rely on the R-square measure only, we may be convinced that this model is as good a fit to the data as it was for the first dataset. Only a visual examination employing a graph has shown us how wrong our assumption of linearity would have been. This tells us that one of the most important things we look for in the scatterplot is to see if the trend of any pattern is linear or non-linear.

Let's now take a graphical look at the relationship between the two variables in dataset C:

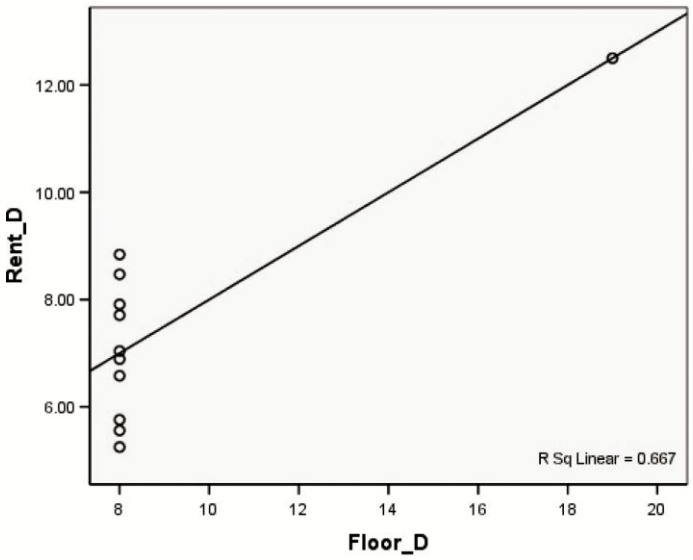


In dataset C, we can see one rent observation that clearly falls outside the perfect linear pattern of the other rents. This "outlier" pulls the regression line up and thus exerts more influence on the fit line than it should. It would be a mistake to rely on a model that includes this outlier because it would not reflect the pattern shown by most of the data points. It is also of interest to note that the R-square measure for this simple regression is identical to those of the first two datasets. If we re-ran this graph with a filter set to eliminate this outlier (Data → Select Cases → If condition is satisfied → If… → Rent_C < 10 → Continue → OK), the following graph results:

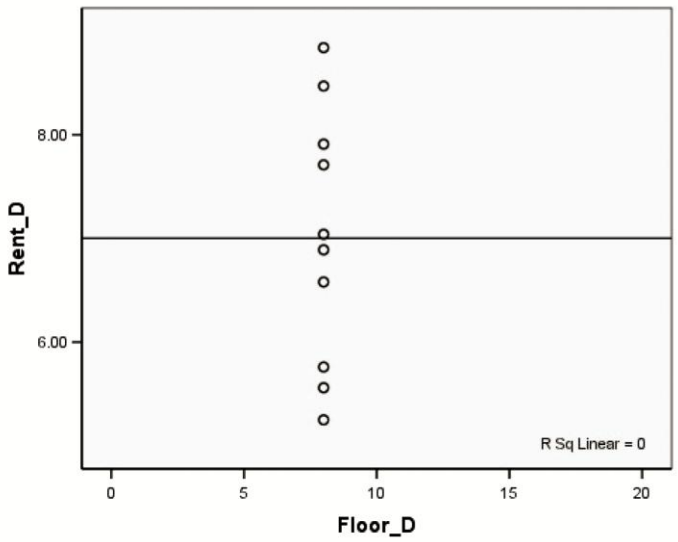


The straight line now fits the remaining data points perfectly. Of course, you would never see this exact type of pattern in the real world. The point of this exhibit is to stress the importance of locating any unusual observations and dealing with them before proceeding with an analysis.

Now let's look at a scatterplot of the dataset D (remember to remove the filter first):



In this graph, we not only see a single observation exerting undue influence on the placement of a regression line, it completely controls its placement. This particular regression has the same R-square as the regressions performed on the other three datasets. We can filter out this last case in order to focus in on the remaining data (If… → Floor_D < 10). The scatterplot below reflects this influential case temporarily removed. As we first noticed when we scanned the table that contained all the data, all but one of the rent observations were from one floor. As can be seen, the regression model fit to all of the data certainly did not reflect most of the observations.

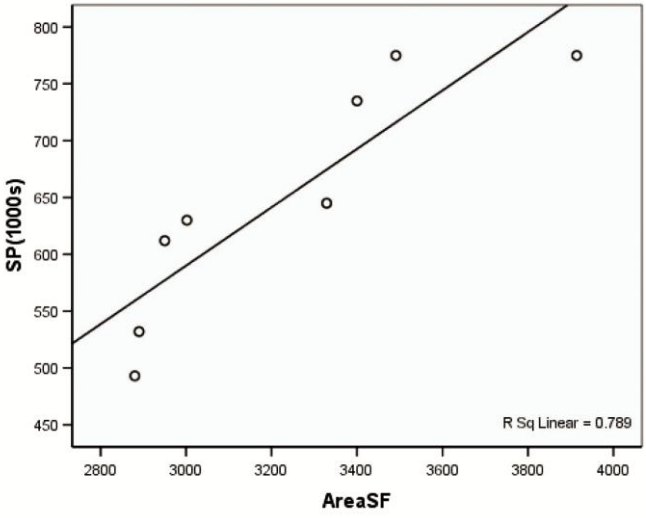


The lesson learned is that if you are examining the relationship between any two variables, you must always examine a picture of your data before proceeding with any more formal analysis. This short case study very strikingly shows why graphic analysis is necessary in order to see the structure or patterns in data.

Case Study 4: Revelation Using Boxplots and Compare Means

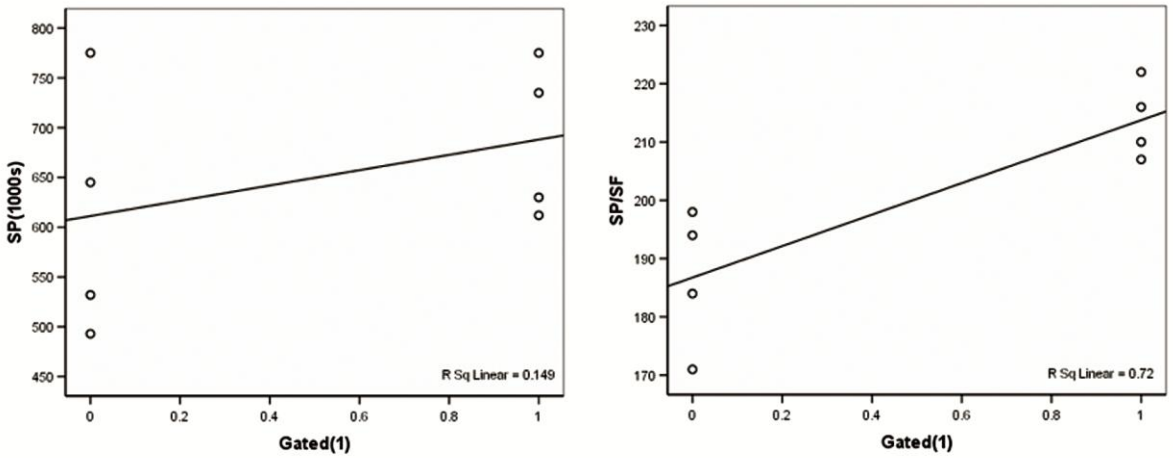
This real estate problem came about from an appraiser in the South-Western US trying to determine an adjustment for location, specifically if the community was "gated" or not. The appraiser had eight single-family home sales to work with, four in gated communities, four not. The sales data provided the house price, living area, and whether or not it was in a gated community. The question: do similar homes in gated communities sell for more than those in non-gated communities?

We created a dataset called "gated" for this sales data. We created a new variable (SPSF), sale price per square foot, in order to account for the influence of size differences on price, and thus allow consistent comparison of sale prices. This makes the sample more or less similar in characteristics, isolating gated location as the variable of interest. To confirm this was appropriate, we viewed a scatterplot of sale price against area and found a strong linear relationship between these variables.

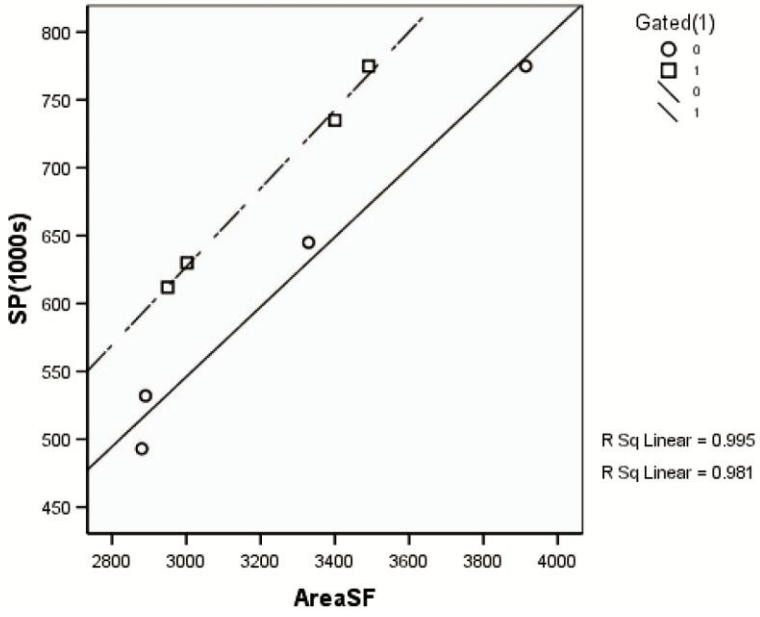


We then created a dummy/binary variable (Gated1) to indicate whether each property was in a gated community or not. The variable was coded 1 = gated and 0 = not gated. This facilitates comparison of these as two distinct groups.

First, we examine a scatterplot of sale price and sale price per unit by the gated variable.



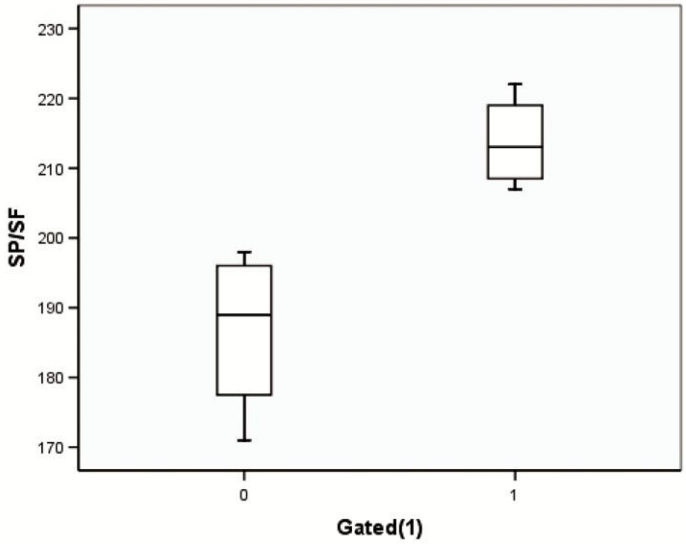
The graphs show a relatively weak relationship of sale price to gated location, but a stronger relationship when using sale price per unit, presumably because the effect of size has been accounted for. However, we find the scatterplot is not particularly suitable for analyzing discrete variables and the correlations indicated are not likely reliable.



Instead, we will view a scatterplot with a separate line for sales in gated communities and those not. This was explained in Lesson 2 in the "Scatterplot with Multiple Trend Lines" section. We will view sale price against area, with Gated1 in the "Set Markers By" box (ensure the "Use chart specifications from:" box is de-selected to not use the RSQ1.sgt template). After running the graph, double click on chart to open the edit chart feature and select Elements → Fit Line at Subgroups. This displays the regression line and R-square value for gated and non-gated sales.

The scatterplot results indicate that gated sales (marked by squares) tend to sell, on average, at a higher price than non-gated homes (marked by circles). The best fit regression lines are useful to visually indicate the spread and linear function for both groups.

A boxplot is probably superior for visualizing differences between distinct groups. We will view sale price per unit against gated, as shown below:



The first box shows non-gated locations, the second shows gated. A quick overview of boxplots: the actual box encloses the middle 50% of the data and the line through the box marks the median. As can be seen from this plot, there is a distinct difference between the centres of the two distributions of prices. Given that the boxes do not overlap, we can be quite confident these are statistically different.

We now use Analyze → Compare Means → Means… to see a numerical summary of the centre and spread of the two distributions of prices. Set SPSF as the Dependent variable and put Gated1 in the Independent List field. We get the following results:

SPSF

Gated1	Mean	N	Std. Deviation
0	186.75	4	12.038
1	213.75	4	6.652
Total	200.25	8	17.011

Based on four sales in each group the results show that homes in the gated community sample sold for $213 per SF while those in the non-gated areas sold for less at $187 per SF. The difference as a ratio is 213.75/186.75 = 1.145 or the gated homes in the sample sold for 14.5% more. This is our best estimate of the adjustment needed for location, when sale comparables are in gated communities and the subject is not. In completing an appraisal, we would then seek further evidence to test and support this estimate.

Case Study 5: Re-Expression – Identifying Unit of Comparison

This case uses the dataset "madison",[7] providing sales of commercial buildings in the downtown of Madison, Wisconsin, U.S.A. between 1973 and 1976. There are 13 cases with data on eight variables, seven continuous and one descriptive. The variables are as follows:

- Total Sale Price (SPRICE);
- Gross Building Area in square feet (GBAREA);
- Gross ground floor area in SF (GGFAREA);
- Front feet of exposure on primary road (FFPRIMARY);
- Front feet of exposure on all roads (FFTOTAL);
- Gross rentable area in SF (GRENTABLE);
- Lot area in SF (LOTSIZE); and
- the property's location (ADDRESS).

There are an additional six variables in the database, which are the sale price divided by each of these units, e.g., SPGBAREA is the sale price per square foot of buildable area. These will be explained in more detail below.

In many appraisals it is standard procedure to divide sale price by a unit measure and then use sale price per unit as the means of comparison: e.g., sale price per square foot or sale price per front foot. This procedure is performed so often that the appraiser sometimes may lose sight of what the objective of the exercise is. When this happens, we may routinely resort to ad hoc choices for units of comparison that may not be the most appropriate. For example, one rule-of-thumb is to pick a unit of comparison we believe is the one most frequently employed by typical market participants. However, in exploring this decision we may find that what we initially believe is not actually true.

The units of comparison may be physical or economic units. In choosing the best unit of comparison, one optimal goal may be to choose the one that explains as much of the difference in prices between the sales as possible. We do this by finding the characteristics of the properties that affect the price – in effect, finding a signal amongst the noise.

Therefore, we will seek a unit of comparison whose re-expression of price results in the largest reduction in initial variation in the total prices paid for properties. The remaining variance is left to be explained by other attributes, but the less remaining variation requiring explanation, the better off we will be.

In this case study, we have five different measurement variables that are candidates for a comparative unit measure. We have in our data variables that reflect the sale price per unit for each of the unit measurement variables in the "madison" dataset. We will now employ two techniques to see which unit of comparison explains the most variation in selling price: (1) the coefficient of variation (COV) statistic, and (2) the coefficient of determination (R-square) obtained through simple linear regression.

First, we will compare the coefficient of variation (COV) statistic for each possible unit of comparison. We used the SPSS "Descriptives" module to compute the mean and standard deviation for each variable. We then manually calculated the COV by dividing the standard deviation by the mean (this could also be calculated using the "Ratio" module in SPSS):

[7] These sales come from several appraisal demonstration reports produced by students of one of the late James Graaskamp's appraisal classes at the University of Wisconsin.

	N				
	Valid	Missing	Mean	Std. Deviation	COV
SPFFPRIMARY	13	0	3354.56	2039.83	2039.83 ÷ 3354.56 = 60.8%
SPFFTOTAL	13	0	2479.40	1955.68	1955.68 ÷ 2479.40 = 78.9%
SPGBAREA	13	0	14.67	5.57	5.57 ÷ 14.67 = 38.0%
SPGGFAREA	13	0	40.09	14.84	14.84 ÷ 40.09 = 37.0%
SPGRENTABLE	13	0	16.71	5.34	5.34 ÷ 16.71 = 32.0%
SPRICE	13	0	184653.85	118184.07	118184.07 ÷ 184653.85 = 64.0%
SPSFLOT	13	0	31.95	13.53	13.53 ÷ 31.95 = 42.3%

Helpful Hint!

Tables in SPSS occasionally show ********* for some results. By double-clicking on the table you will open the SPSS Pivot Table. 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 may wish to change the number of decimal places (select the cell or cells, click Format → Cell Properties...and under the Format Value Tab, change the number of decimal places). Close the SPSS Pivot Table when you are finished.

You can also adjust the number of decimal places that a variable has by using the Variable View tab when looking at the data – this is often a good idea when looking at data for the very first time. Simply click the Variable View tab and quickly scan down the Decimals column. Three decimal places (as a maximum) for most numeric variables is a good rule of thumb; many, such as a count of fireplaces, would require zero.

With a COV of 32%, the sale price per square foot of gross rentable area (SPGRENTABLE) shows the smallest average variation about the mean. Compare this with the relative variation for overall sale price, which had a 64% COV. By using the selling price per square foot of gross rentable area in our direct comparison analysis instead of overall sale price, we have reduced the amount of unexplained variation in prices by one-half.

Based on a COV analysis, sale price per square foot of gross rentable area appears to be the best unit of comparison here, followed by sale price per square foot of gross ground floor area (SPGGFAREA) with a COV of 37%.

For a second analysis, we will use simple linear regression to explore the correlation between the sale price and the corresponding unit measurement. This will focus on goodness of fit as measured by the coefficient of determination (R-square). This statistic is produced by a regression of the dependent variable sale price against the various unit of comparison predictor variables.

We first examine the correlation matrix for the unit variables (Analyze → Correlate → Bivariate → check Pearson under Correlation Coefficients → click OK):

Correlations

	FFPRIMARY	FFTOTAL	GBAREA	GGFAREA	GRENTABLE	LOTSIZE	SPRICE
FFPRIMARY	1	.896(**)	.651(*)	.750(**)	.629(*)	.831(**)	.569(*)
FFTOTAL	.896(**)	1	.739(**)	.725(**)	.724(**)	.855(**)	.663(*)
GBAREA	.651(*)	.739(**)	1	.895(**)	.965(**)	.931(**)	.934(**)
GGFAREA	.750(**)	.725(**)	.895(**)	1	.883(**)	.858(**)	.802(**)
GRENTABLE	.629(*)	.724(**)	.965(**)	.883(**)	1	.897(**)	.948(**)
LOTSIZE	.831(**)	.855(**)	.931(**)	.858(**)	.897(**)	1	.875(**)
SPRICE	.569(*)	.663(*)	.934(**)	.802(**)	.948(**)	.875(**)	1

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

Gross rentable floor area (GRENTABLE) shows the highest linear correlation with sale price among the various units of measurement. This agrees with our COV conclusion. Note that squaring the correlation coefficients (R) gives us the R-square or coefficient of determination statistic. We could also find these values through six simple linear regression runs: Analyze → Regression → Linear… → SPRICE as the Dependent variable, and each of the six variables as Independent variables (run one at a time) → OK. This will run a regression for each variable against sale price. For example, part of the output for the regression of sale price against gross rentable floor area is below. The R-square statistic is bolded for emphasis.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.948(a)	**.898**	.889	39456.018

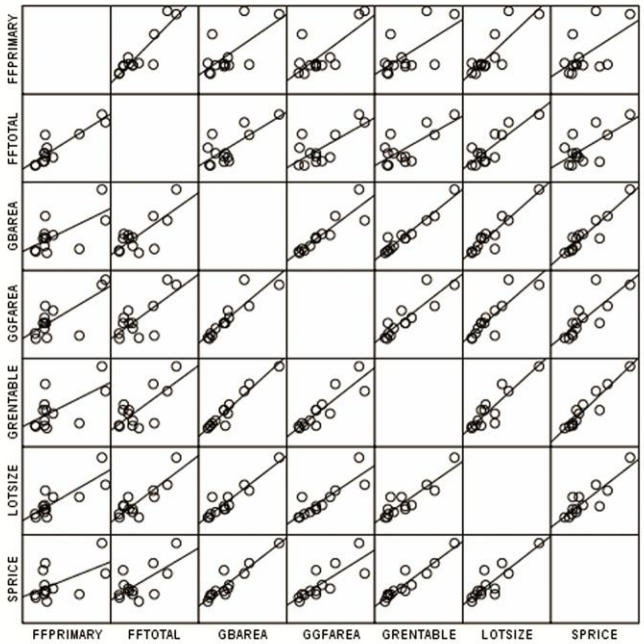
a Predictors: (Constant), GRENTABLE

The R-square statistics for all the variables are summarized below:

	SPRICE
FFPRIMARY	0.323
FFTOTAL	0.439
GBAREA	0.871
GGFAREA	0.643
GRENTABLE	0.898
LOTSIZE	0.766
SPRICE	1

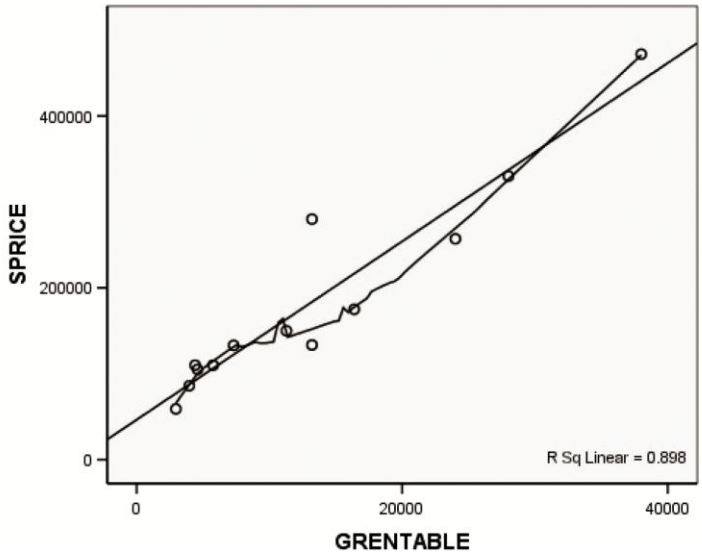
It is a good idea to visually examine the fit between sale price and the variables to ensure that a linear function is appropriate for this data. To do this quickly, we can specify a scatterplot matrix, which provides a comprehensive view of the relationships between the variables specified. It is similar to the correlation matrix we examined earlier, only that the relationship between pairs of variables is visualized graphically as opposed to interpreting the correlation coefficients. SPSS instructions are as follows:

- Select Graphs → Legacy Dialogs → Scatter/Dot… → Matrix Scatter → Define.
- Specify the seven variables above as "Matrix Variables".
- Click on "Use chart specifications from" and select "RSq1.sgt", to display regression lines.



The bottom row of the matrix provides the plots we are interested in. These six scatterplots graph sale price against the six unit measurement variables. For example, the left-most scatterplot on the bottom row graphs sale price on the vertical axis against front foot primary on the horizontal axis. All six scatterplots appear to show a positive relationship between sale price and the paired unit of comparison. However, the strength of correlation varies.

Gross rentable area had the highest R-square, so let's examine its scatterplot more closely.



A linear function appears appropriate for the relationship between sale price and gross rentable area. Therefore, we confirm gross rentable area as the best unit of comparison according to both the coefficient of variation (COV) and the coefficient of determination (R-square).

A summary of results is provided below (please note: this table has been manually compiled from the previous results, with rankings added). It is interesting to note gross building area ranks second according to the R-square technique, but third if we use the COV to make the decision.

	Mean	Std. Deviation	COV	Rank	R-square	Rank
FFPRIMARY	3354.56	2039.83	60.8%	5	32.4%	6
FFTOTAL	2479.40	1955.68	78.9%	6	44.0%	5
GBAREA	14.67	5.57	37.9%	3	87.2%	2
GGFAREA	40.09	14.84	37.0%	2	64.3%	4
GRENTABLE	16.71	5.34	32.0%	1	89.8%	1
LOTSIZE	31.95	13.53	42.3%	4	76.6%	3

Our goal in this case study was to reduce unsubstantiated guesswork in the application of the direct sales comparison approach. Through exploratory data analysis, we better understand the relationship between price and unit measures and can proceed knowing we are on the right track.

Case Study 6: Residuals – Identifying the Need for Further Adjustments

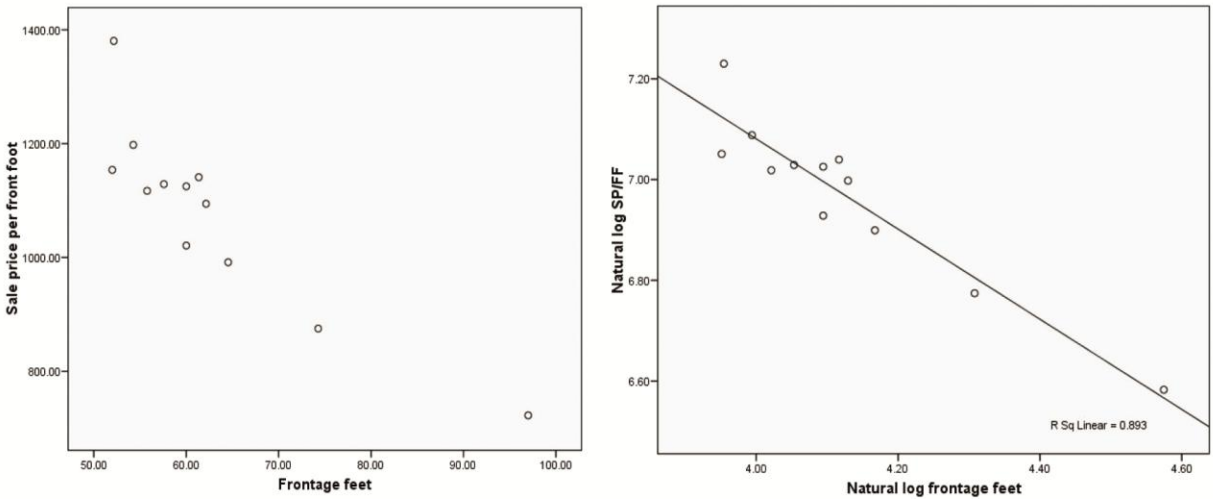
In this case study, we are appraising the market value of a vacant single-family residential lot in Moulton, Ontario. The subject lot has 83 feet of frontage, 153 feet of depth, and is considered a deep lot in this market. We have gathered data for 12 comparable sales, provided in the "Moulton" dataset and listed in the table below. As you will see, we chose frontage as the unit of comparison and adjusted the comparable sales based on a logarithm-based regression line in order to account for decreasing returns to scale. Then, after checking the residuals, we made a further adjustment for lot depth. The table results below show the completed analysis, illustrating the final result and how the results were improved along the way by successive adjustments. This case study explains the process we went through in analyzing the data and making these adjustments.

Sale #	Address	Sale Price	Front-age	LN Front-age	Depth	Deep Lots(1)	Area	SP/SF	Price /FF	LNPrice /FF	FF Adj Factor	FFAdj SP/FF	Adj Total SP(FF)	DepthAdj SP/FF	Full Adj Total SP
1	444 Lakeview	70,093	97.00	4.57	114.00	0.00	11147	6.29	723	6.58	1.1488	830.14	68,901	883	73,306
2	469 Lakeview	65,000	74.28	4.31	114.00	0.00	8622	7.54	875	6.77	0.9059	792.75	65,799	846	70,202
3	400 Lakview	70,000	61.35	4.12	200.00	1.00	10895	6.42	1141	7.04	0.7641	871.88	72,366	872	72,366
4	408 Lakeview	68,000	62.14	4.13	171.92	1.00	9881	6.88	1094	7.00	0.7729	845.78	70,200	846	70,200
5	702 Gladstone	64,000	64.54	4.17	104.11	0.00	6835	9.36	992	6.90	0.7994	792.72	65,796	846	70,200
6	696 Gladstone	61,250	60.00	4.09	106.92	0.00	6523	9.39	1021	6.93	0.7492	764.77	63,476	818	67,880
7	440 Lakeview	67,500	60.00	4.09	196.52	1.00	11793	5.72	1125	7.03	0.7492	842.81	69,953	843	69,953
8	Wellington	60,000	52.00	3.95	100.00	0.00	6027	9.96	1154	7.05	0.6596	761.05	63,167	814	67,571
9	Eastview	72,000	52.16	3.95	128.35	1.00	6361	11.32	1380	7.23	0.6614	912.95	75,775	913	75,775
10	476 Lakeview	65,000	54.27	3.99	120.24	1.00	5705	11.39	1198	7.09	0.6851	820.60	68,110	821	68,110
11	122 Kingfisher	62,300	55.77	4.02	101.71	0.00	6081	10.25	1117	7.02	0.7020	784.16	65,085	837	69,489
12	499 Parrott	65,000	57.58	4.05	105.62	0.00	9257	7.02	1129	7.03	0.7222	815.27	67,668	868	72,071
MEAN		65,845						8	1,079			820	68,025	850	70,594
MEDIAN		65,000						8	1,121			818	67,889	846	70,200
ST DEV		3741						2	165			45	3707	29	2418
COV		6%						24%	15%			5%	5%	3%	3%

You will note that the data includes a binary variable for deep lots and calculated variables for sale price per square foot, sale price per front foot, and the natural logs of both the frontage and the sale price per front foot (these variables will be used in the analysis).[8] The other variables will be described as we proceed through the case study.

During our data analysis, we discovered an inverse nonlinear relationship between front feet and sale price per front foot – see graph on left below. The data includes the natural logs of both variables (the "LN" notation in the data: LNFrontage and LNPriceFF). This would be considered a re-expression of these two variables by taking their natural logs. Continuing on, we graphed them and found that the natural logarithms of the variables had a linear relationship – see graph on right below. Because of the strength of this linear relationship, we can calculate the size adjustment factor directly from the graph's regression line.

[8] Natural logarithms will be covered in detail in Lesson 5.



For the graph on the right, we used the following procedure to determine the equation for the simple regression line:

- Analyze → Regression → Linear…
- Enter LNPriceFF as the Dependent variable and LNFrontage as the Independent variable
- Click OK.

Here are the results, garnered from the table of coefficients:

$$\text{LNPrice/FF} = 10.660724 - 0.8949763 \times \text{LN Frontage}$$

This indicates a slope of -0.8949763 (rounded to -0.89).[9]

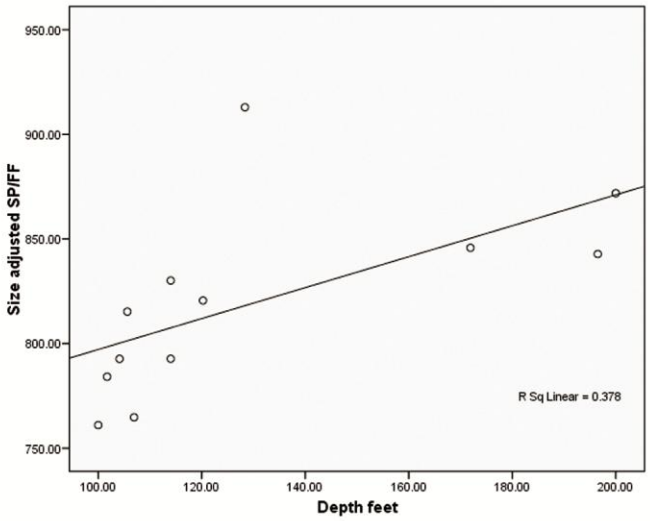
Next, we applied the value for the slope as an exponent in the formula for the size adjustment factor,[10] which allowed us to create a new variable, the size-adjusted sale price per front foot (FFAdjSP/FF). The COV for the size-adjusted sale price per front foot (which you can see is 5% in the bottom row of the data table presented earlier) is far better than the COV for the unadjusted sale price per front foot, which was 15%. This is a large improvement in explaining the variation in sale prices.

Before concluding our analysis, we should examine our size-adjusted price per front foot to ensure we have explained as much variation as possible. In effect, this is examining the residuals from the model. Can we improve results further?

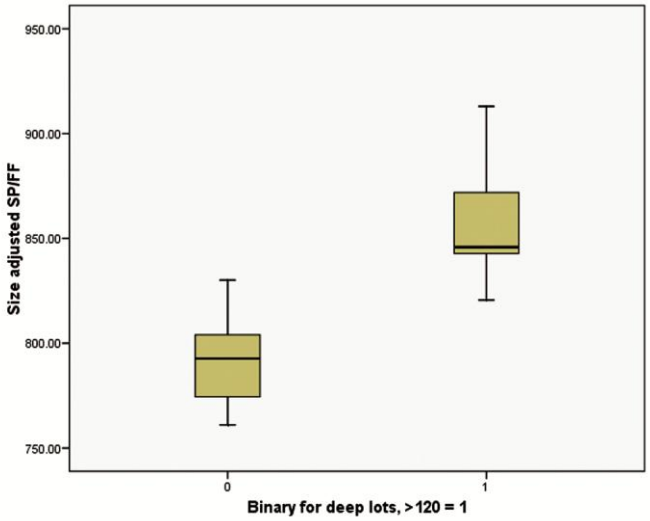
We have used front feet as our unit of comparison, but does depth have any effect on value? Looking at the data, there is a fairly wide spread in depth, from 100 to 200 feet. We will examine the relationship of size-adjusted price per front foot with depth. The scatterplot of "FF/Adj SP/FF" against Depth showed a grouping of lower priced lots with smaller depths and another group of higher prices lots with larger depths.

[9] To interpret this, think of a 10% increase in frontage resulting in an 8.9% drop in the sale price per front foot.

[10] The size adjustment factor is a somewhat complicated calculation. It is $(83 \div \text{front foot})^{-0.89}$ for each record, where 83 is the front footage value from the subject property. This will be more fully illustrated in Lesson 5.



We will use the binary variable DeepLots1 which is equal to 1 for lots deeper than 120 feet and is equal to 0 for those lots less than 120 feet deep. The resulting boxplot of FF/Adj SP/FF against Deep Lots (binary) shows this relationship clearly.



The seven sales with shallow lots (identified with a 0) have the lowest prices per front foot, while the five sales considered to be deep lots (identified with a 1) have the highest prices per front foot. This is summarized in the table below:

FFAdjSPFF

		Deep Lot = 0	Deep Lot = 1
N	Valid	7	5
	Missing	0	0
Median		792.72	845.78
Minimum		761.05	820.60
Maximum		830.14	912.95
Percentiles	25	764.77	831.70
	50	792.72	845.78
	75	815.27	892.42

Note: the preceding table was constructed manually from two tables created by first using:

- Data → Split File → Organize output by groups → enter DeepLots1 in the Groups Based On: field → OK;
- then Analyze → Descriptive Statistics → Frequencies… for FFAdjSPFF (and choosing the appropriate statistics from the Statistics… button).

The two groups have distinctly different selling prices per front foot. The difference between the medians of the two groups is $53.06 per front foot (845.78 – 792.72) and we will therefore adjust the prices per front foot of the shallow lots upward by $53.06 per front foot (since the subject lot would be considered a deep lot). The adjustments can be seen in the last two columns of the table at the start of this case study (FullAdjTotalSP). Adjusting for the depth of the lots further reduces the variation to a COV of 3%, indicating that both the size (front foot) and depth adjustments substantially improved the valuation.

The highlighted cases in the table, Sales 5, 6, and 8 will be used in the adjusted grid for the direct comparison approach in the appraisal. We will note the adjustments made in the grid for these three sales. This analysis could be a called a "paired group comparison". It is similar to paired sales, but technically more accurate because we consider variation in the distribution of prices and the paired sales technique does not.

This case study illustrated the need for examining a model's residuals. After preliminary analysis, you should look at your results, confirm they make sense, and then see if there are opportunities for refinement and improvement. In this case, we found a further adjustment was necessary, and improved our model. Our appraisal results improve and our client is happy!

Summary: Exploratory Data Analysis

Exploratory data analysis is a critical element in any statistical analysis. Without a solid understanding of your data, what it consists of and the patterns within it, you will not be able to properly analyze it or create models using this data. Therefore, exploratory data analysis is a key part of any practical application of statistics in real estate.

Exploratory data analysis is based on the "Four Rs":

- Reduction
- Revelation
- Re-expression
- Residuals (from models)

In this lesson, we provided six case studies to highlight these. To review:

- Case Study 1: *Reductio*n using summary statistics and histograms
- Case Study 2: *Reduction and Revelation* using graphs and correlation and scatter plots
- Case Study 3: *Revelation* using correlation and scatter plots
- Case Study 4: *Revelation* using box plots and compare means
- Case Study 5: *Re-expression* to find a unit of comparison via the Coefficient of Variation (COV)
- Case Study 6: *Residuals* (and *Re-expression* with natural logarithms) using scatter plots, box plots and compare means to identify the need for further adjustment

After working through these cases, it should be clear that the four Rs are inter-related, with each being important considerations in all analyses. They can occur in any order and may need to be repeated as the analysis proceeds (e.g., Reduction carried out after some Re-expression).

With this lesson now done, we have completed the introductory foundations section of this course. Lesson 1 covered basic statistical theory, Lesson 2 explored the use of computer applications for statistics, and this lesson highlighted the need for comprehensive data exploration and analysis. Continuing our metaphor from this lesson's introduction: we have now built all three legs for our statistical foundations stool and it is now time to step up to practical applications.

Lessons 4 and 5 will focus on single-property applications of statistical analysis. We will look at how statistical analysis can be used in everyday real estate practice, in particular in valuation work. The focus will be on practical uses for the typical real estate professional, keeping in mind the constraints of limited data and limited computing resources (hardware and software). In other words, our goal will be to present "job ready" skills and techniques you can immediately apply.

Lessons 6 to 8 will focus on the use of multiple regression for mass appraisal. In other words, given a database of property sales, how you can apply regression analysis to predict the market value of other properties that did not sell? This is particularly important for those working in property assessment for taxation purposes, but it is also important for anyone whose work may be affected by appraisal valuation models (AVMs).

With that said, we will now proceed on to Lesson 4 and begin our exploration of applications for statistical analysis.

Review and Discussion Questions

1. You hired a consultant to complete a statistical analysis predicting the need for seniors housing in Langley, BC. In reviewing the results, should you focus on the reliability of the forecast in relation to other benchmark data? Or do you need to examine the consultant's interpretation of the underlying data relationships?

2. How could you use visual presentation aids to help a client understand a statistical analysis?

3. Real estate terminology is very specialized. A variable describing office building class (e.g., Class A, Class B, Class C) would be what type of data variable? What possible problems might you experience in relying on this building class variable?

4. Do you agree with the following statement: "The goal of exploratory data analysis is to identify and account for every source of variation in data relationships"?

5. Assume you are looking at a histogram with a normal distribution. If some data was removed from the dataset, resulting in the median being lower than the mean, how do you think the new histogram would look?

6. Which would be a better tool for analyzing the relationship between two continuous variables: a boxplot, scatterplot, or histogram?

7. You want to "smooth" the data relationship in a scatterplot. How might you do this?

8. Consider three sample datasets drawn from the same population of high-rise condo sales in Vancouver. The datasets include a number of variables: sale price, unit size, floor height, view, and parking. The descriptive statistics for each dataset indicates similar mean and median values for sale price per front foot. What following steps should you take in comparing and analyzing the datasets? What should you be attempting to uncover?

9. If the correlation between two variables is strong (say, between rent per square foot and quality of retail space), is the relationship deemed to be linear?

10. If two data variables, say, price per square foot and finished floor space in new housing, had a linear relationship, what would be an easy way of determining the linear regression equation?

11. Ideally you want data to have a linear relationship and strong correlation, but this is often not present in real estate data. What would be the risk of relying on a regression equation where most of the data occurrences were concentrated at one end of the regression with few occurrences at the other end?

12. Assume you are comparing two datasets for apartment rents in Victoria, BC. One dataset reflects 3-storey "walk-up" apartment rents in the James Bay community and the other dataset includes similar property rents in another Victoria community, Fernwood. You want to determine the effect of location on rent. Assuming both datasets have a similar structure (e.g., same variables), what approach would you use to compare the datasets?

13. How can the coefficient of variation be used to help determine appropriate units of comparison for real estate data?

14. You have a database of recreational lot sales and are forecasting sale price per front foot for a certain size of waterfront lot. How can you account for non-linear data relationships in your forecast?

ASSIGNMENT 3

LESSON 3: Exploratory Data Analysis

Marks: 1 mark per question.

1. When exploring data for the first time, preliminary screening is important so you can:

   (1) seek patterns in the data.
   (2) understand relationships within the data.
   (3) eliminate data you do not need or identify data that seems odd or impossible.
   (4) All of the above.

2. Which of the following statements is TRUE regarding the "Moulton" case study?

   (1) Sale price per square foot of lot area was revealed to be the best unit of measure.
   (2) Re-expression of sale price per front foot was found to be unnecessary, given its linear relationship.
   (3) Identifying the residuals confirmed the size adjustment alone was insufficient; an adjustment for depth was also necessary.
   (4) The analysis of residuals confirmed no further adjustments were necessary after size (front feet) was accounted for.

3. You have created a model to estimate office vacancies over time. You notice that the error from your prediction does not have constant variance and you believe it shows heteroskedasticity. This analysis would be an example of which of the "Four Rs"?

   (1) Reduction
   (2) Revelation
   (3) Re-Expression
   (4) Residuals

THE FOLLOWING SIX (6) QUESTIONS REFER TO THE "BURNABY" DATASET FROM LESSON 2:

4. How would you describe the "bedrooms" variable type?

   (1) Nominal variable scale with continuous characteristics.
   (2) Interval variable scale with discrete characteristics.
   (3) Ordinal variable scale with subjective characteristics.
   (4) Ratio variable scale with continuous characteristics.

Assignment 3 continues on next page

5. Key goals of exploratory data analysis are to simplify the complexity of data available for analysis and understand the distribution and variance of the data. How would you describe the relationship between the *quality* and *sale price* variables?

   (1) The variables are stochastic in nature since quality is an excellent predictor of sale price.
   (2) The variables are deterministic in nature.
   (3) A scatterplot of the variables reveals the data has a smooth relationship.
   (4) The variables are stochastic in nature since quality does not explain all the variation in sale price.

6. Assume you have begun studying the relationship between *TESTVAL* and *EFFYRBLT*. However, you are only interested in carrying out further research if at least 70% of the variance in *TEST_VAL* can be explained by *EFFYRBLT*. Should you conduct further research?

   (1) Yes, since a boxplot graph of the relationship reveals a moderate amount of "outliers".
   (2) No, since a scatterplot graph of the relationship reveals a R-square value of 0.564.
   (3) No, since the significance of the ANOVA between the two groups is > 0.
   (4) Yes, since the confidence intervals include 1.

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. Analyze the relationship between *TOTAREA* and *FINAREA*. What is the level of correlation between the two variables?

   (1) Poor correlation with a Pearson Correlation of 0.2110.
   (2) According to high R^2 value of 0.756, the two variables are well correlated and the Loess line indicates very high linearity.
   (3) Although both variables are well correlated with a Pearson Correlation of 0.883, the Loess line may indicate some non-linearity.
   (4) Low, according to the R^2 value of 0.779.

9. You want to determine which has less variation: sale price per square foot of finished area (above the basement) or sale price per square foot of lot size. Which of these shows LESS variation?

   (1) Price per square foot of lot area since its COV is approximately double the COV for finished area.
   (2) Price per square foot of finished area since its COV is approximately half the COV for lot area.
   (3) Price per square foot of lot area since its standard deviation is less than the standard deviation for finished area.
   (4) Cannot be determined, not enough information.

Assignment 3 continues on next page

THE FOLLOWING TWO (2) QUESTIONS REFER TO THE "INDUSTRIAL" DATASET FROM LESSON 3:

10. You want to examine data for Market Area 1, but only for observations greater than $38. Which of the following is correct?

    (1) A histogram of the new dataset shows a perfect normal distribution.
    (2) A histogram of the new dataset shows the data is now skewed to the right.
    (3) A histogram of the new dataset shows the data is now skewed to the left.
    (4) The mean has increased to $45.

11. You want to examine data for Market Area 3, but only for observations less than $45. Which of the following is correct?

    (1) The mean of the filtered dataset is $38.89.
    (2) A histogram of the filtered dataset shows the data is now skewed to the right.
    (3) The range of filtered dataset is $31 to $44 and the standard deviation is $4.10.
    (4) The number of observations in the filtered dataset is 18.

12. An appraiser needs to determine the appropriate units of comparison for surface parking lot comparable sales. Choices include price per parking stall, price per square foot, and overall lot price. What key factor must the appraiser consider in selecting units of comparison?

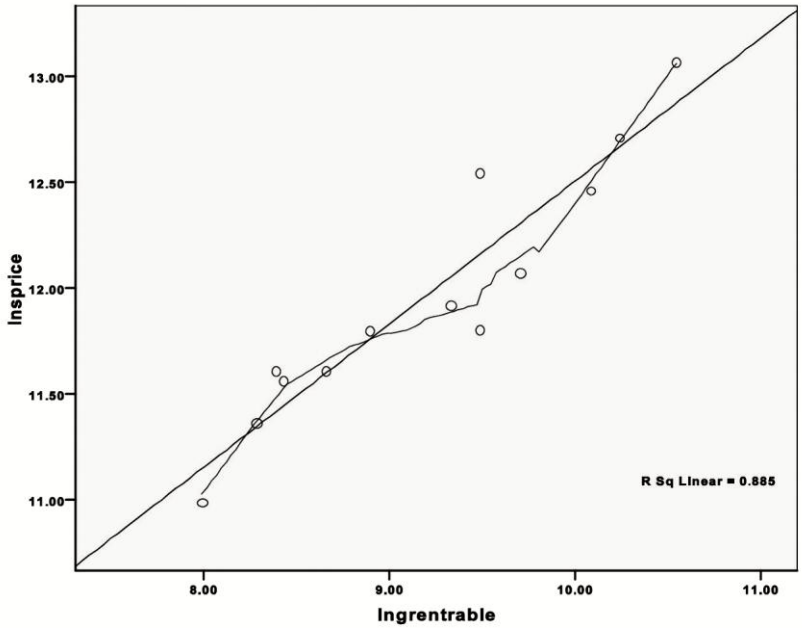
    (1) The unit of comparison (variable) must explain as much variation in value as possible.
    (2) The unit of comparison must always reflect the typical units adopted by other appraisal professionals for the same property type.
    (3) A low R^2 value for the relationship between sale price and the preferred unit of comparison.
    (4) A high COV indicates a superior result.

13. You are involved in an assessment appeal for a property that has a fabulous view, for which the owner feels she has been over-assessed. You have analyzed the sale of all view properties in this area over the past year and coded the quality of each as Excellent, Good, Fair, or Poor. However, you find the variable's current format means it cannot easily be used in further analysis in your statistical software program. In order to make this variable more useful, which of the "Four Rs" is necessary?

    (1) Reduction
    (2) Revelation
    (3) Re-Expression
    (4) Residuals

Assignment 3 continues on next page

14. In the "madison" dataset, we analyzed Sale Price versus Gross Rentable Area and determined a linear function was appropriate. To test this assumption, we also transformed both variables using natural logarithms and created the scatterplot below. Did using logarithmic analysis improve the analysis of the relationship between these two variables?



(1) Creating a logarithm for sale price versus gross rentable area greatly improved the R-Square.
(2) There is no significant difference; both outcomes indicate a strong relationship with low variance.
(3) The plot of sale price versus gross rentable area is a far superior outcome using a logarithmic analysis.
(4) There is no significant difference; both outcomes indicate a very weak relationship with high R-square values.

15. Your analysis of two area variables confirms that the COV and R-square indicate similar statistical outcomes, but you are concerned about possible non-linearity. How would you decide which variable to select for further analysis?

(1) Run scatterplots for each variable to learn if the relationships are linear or non-linear; non-linear relationships make a variable impossible to use.
(2) Run scatterplots for each variable and compare the "fit of the data" for each outcome; the highest slope in the regression line will determine which unit of comparison is preferred.
(3) Compare the standard deviation for each dataset; lowest will be best.
(4) Run scatterplots for each unit of comparison, evaluate the result, and conduct logarithmic testing to determine which unit of comparison best accounts for variation in the dependent variable.

Assignment 3 continues on next page

16. A colleague in your consulting firm has quit and you inherited his files. He was in the middle of large-scale market study for retail development sites and the data is a mess. You're not sure what's relevant and you definitely can't see any patterns or make any conclusions. Which of the following is a recommended exploratory data analysis technique?

    (1) Reduce the uncertainty by organizing the data into a database and eliminating unneeded cases and variables.
    (2) Run summary statistics to better understand the range in the data and the averages.
    (3) Create scatterplots and boxplots to get a sense of the relationships between variables.
    (4) All of the above.

17. Assume you completed a non-linear analysis of rents per front foot versus store frontage for retail properties on Robson Street, an exclusive shopping precinct in downtown Vancouver. The COV analysis indicates that a logarithmic regression accounts for most of the variation in rents. However, a significant variation remains – you suspect it is related to excess retail store depth (e.g., more than is required for storage). What could you do to verify your hypothesis?

    (1) Nothing. This problem cannot be solved.
    (2) Set a filter to eliminate newer properties from the analysis.
    (3) Re-express rent per front foot as rent per deep foot.
    (4) Re-code depth into a new variable which can be graphed in relation to size adjusted rents to see if a significant relationship exists.

18. What would be the best statistic to help answer the following question: If a person purchases a 1,000 square foot high-rise condo in downtown Toronto, how likely are they to purchase a flat-screen TV (assuming data on both variables was collected)?

    (1) Coefficient of variation since the statistic explains the variation of one variable in relation to another.
    (2) Standard deviation since it is a measure of dispersion for both variables.
    (3) Correlation coefficient since it measures the degree to which the values of two variables are proportional to each other.
    (4) Mean since it explains the central tendencies within the data.

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.

19. The New Construction variable is binary – either 'Y' for new construction or 'N' for older. Using the Year Built variable as a guide enter the correct values ('Y' or 'N') for the records that are missing their New Construction data. Then run two box plots (1) Sale price versus New Construction and (2) Sale Price per Square foot versus New Construction. Which of the following statements is TRUE?

    (1) New Construction has an impact on the Sale Price, but not the Sale Price per Square Foot
    (2) New Construction has an impact on the Sale Price per Square Foot, but not the Sale Price
    (3) New Construction has an impact on both the Sale Price and the Sale Price per Square Foot
    (4) New Construction has no impact on either the Sale Price or the Sale Price per Square Foot

Assignment 3 continues on next page

20. Run scatterplots of (1) the Sale Price versus the Sale Date, and (2) the Sale Price versus Year Built, add fit lines. Which of the following statements is TRUE?

(1) The Year Built has a greater impact on the Sale Price than the Sale Date
(2) The market has been flat for the past year
(3) The relationship between the Sale Price and the Year Built is logarithmic
(4) The R-squared for the Sale Price versus Sale Date fit line indicates that a time adjustment of 0.105% per month (compounded) is needed

20 Total Marks

Planning Ahead

You may wish to continue working ahead by investigating the data for Project 1. Using the techniques outlined in this lesson you can Reduce, Reveal, and Re-express the data. Eventually when you have created your "model" you can measure your Residuals. As you proceed thorough Lessons 4 and 5 (and perhaps even 6) you can refine your analysis somewhat.

End of Assignment 3