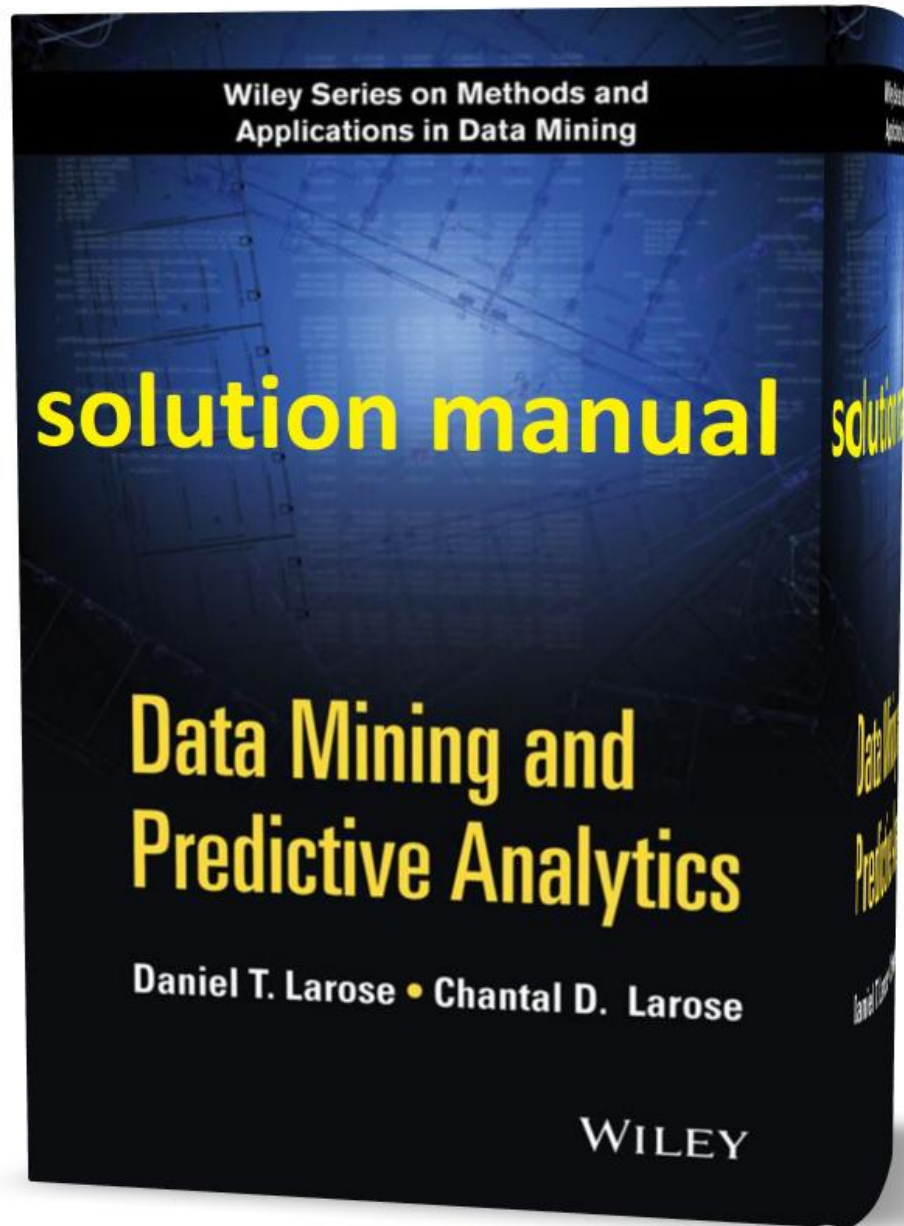


for download complete version of solution (all chapter 1 to 28 ) click here.



<https://gioumeh.com/product/data-mining-and-predictive-analytics-solutions/>

**Solutions to Chapter 1**  
**AN INTRODUCTION TO DATA MINING AND PREDICTIVE**  
**ANALYTICS**

Prepared by James Cunningham, Graduate Assistant

1. For each of the following, identify the relevant data mining task(s):

- a. **The Boston Celtics would like to approximate how many points their next opponent will score against them.**

Estimation

- b. **A military intelligence officer is interested in learning about the respective proportions of Sunnis and Shias in a particular strategic region.**

Classification

Clustering

Description

- c. **A NORAD defense computer must decide immediately whether a blip on the radar is a flock of geese or an incoming nuclear missile.**

Classification

- d. **A political strategist is seeking the best groups to canvass for donations in a particular county.**

Clustering

Classification

Description

- e. **A Homeland Security official would like to determine whether a certain sequence of financial and residence moves implies a tendency to terrorist acts.**

Prediction

- f. **A Wall Street analyst has been asked to find out the expected change in stock price for a set of companies with similar price/earnings ratios.**

Estimation

Data Mining and Predictive Analytics

for download complete version of solution (all chapter 1 to 28 ) [click here.](#)

2. For each of the following meetings, explain which phase in the CRISP-DM process is represented:

- a. **Managers want to know by next week whether deployment will take place. Therefore, analysts meet to discuss how useful and accurate their model is.**

Model Evaluation Phase

- b. **The data mining project manager meets with the data warehousing manager to discuss how the data will be collected.**

Data Understanding Phase

- c. **The data mining consultant meets with the Vice President for Marketing, who says that he would like to move forward with customer relationship management.**

Business Understanding Phase

- d. **The data mining project manager meets with the production line supervisor, to discuss implementation of changes and improvements.**

Model Deployment Phase

- e. **The analysts meet to discuss whether the neural network or decision tree models should be applied.**

Modeling Phase

**3. Discuss the need for human direction of data mining. Describe the possible consequences of relying on completely automatic data analysis tools.**

Data mining requires human direction in order to be both effective and appropriate as problem-solving is a human process requiring human critical thinking every step of the way. As stated in the text, data mining without proper human direction is something that is *very easy to do badly*. It is very easy to derive results that are damaging to business processes by (1) failing to understand the business problem at hand, (2) failing to understand the data sets at hand (and their interrelationships), (3) failing to select appropriate modeling techniques, and (4) failing to evaluate model results correctly.

One very popular fallacy is that data mining can be completely autonomous and thus requires little to no human direction. Applying data mining software features at random is bound to produce the wrong answer to the wrong question with the wrong data. In fact business decisions made based on inappropriate analyses are much more damaging and costly than those made based on no analysis at all. Also, once a model is deployed, it must be monitored for its efficacy and will most often need to be tuned over time.

**4. CRISP-DM is not the only standard process for data mining. Research an alternative methodology (Hint: SEMMA, from the SAS Institute). Discuss the similarities and differences with CRISP-DM.**

SEMMA is a process developed by the SAS Institute for conducting a data mining project. Each letter in the acronym SEMMA identifies a separate stage of the data mining process as follows:

**Sample** – The first stage in SEMMA entails extracting a representative sample of a much larger data set. Please note that this stage is optional and thus used at the discretion of the analyst.

**Explore** – The second stage in SEMMA entails searching for unanticipated trends, patterns, and anomalies in order to gain an understanding of the data and develop ideas.

**Modify** – The third stage in SEMMA entails modifying the data set through a combination of selecting original variables and more importantly transforming variables and deriving new ones that would be most conducive to a data modeling exercise.

**Model** – The fourth stage in SEMMA entails allowing the software to determine the best combination of variables that predict a desired outcome.

**Assess** – The fifth and final stage in SEMMA entails evaluating model efficacy and estimating how well it will perform if deployed.

The CRISP-DM process was developed by a consortium pioneered by DaimlerChrysler, SPSS, and NCR and consists of six stages or *phases* as follows:

**Business Understanding** – The first phase entails gaining an understanding of the business problem at hand and translating this into a data mining problem to be solved and an initial solution approach. In direct contrast with SEMMA, we observe that CRISP-DM prescribes business-requirements development as an explicit activity and the specific data mining problem and solution approach as explicit deliverables whereas SEMMA does not. SEMMA prescribes delving right into the data set, which can lead to significant time wasted (that will most likely be proportional to the dimensionality of the data set being explored).

**Data Understanding** – The second phase entails determining how data will be collected and exploratory analysis. This phase is similar in nature to SEMMA's Explore stage, but in contrast with SEMMA, the exploratory analysis activities of the CRISP-DM Data Understanding phase are conducted from the perspective of solving a particular data mining problem. Therefore, while exploration conducted in SEMMA's Explore stage seems to be by pure brute-force, exploration conducted in CRISP-DM's Data Understanding phase is done from the perspective of a specific data mining problem to be solved. In other words, the exploratory analysis in CRISP-DM's Data Understanding phase is expected to be more effective and more efficient focusing on exploring correlations between predictors and interactions between predictors and a specific target variable.

**Data Preparation** - The third phase entails all of the actions (e.g. selections, transformations, derivations, etc.) needed to develop a data set that is most conducive to a data modeling exercise. This phase is similar to SEMMA's Modify stage, but contrast with SEMMA, the preparation activities conducted in the CRISP-DM Data Preparation phase are done so with a specific data mining problem and target modeling approach in mind. This is a critical distinction between the two processes. As an example, if we have data that is highly inter-correlated or *multicollinear*, we can leverage a dimensional transformation such one produced via Principal Components Analysis (PCA) to eliminate the multicollinearity, but only for certain types of modeling approaches. Therefore, since the CRISP-DM Data Preparation phase has a target modeling approach in mind when preparing data, it can leverage advanced transformational techniques like PCA appropriately and is thus superior to the SEMMA Modify stage.

**Modeling** - The fourth phase entails the human-directed application of multiple modeling techniques in order to (1) optimize the balance between model bias and model variance and (2) maximize the ability for these models to operate effectively on new observations. While this is similar to SEMMA's Model stage, the CRISP-DM Modeling phase is human-directed whereas SEMMA's Model stage appears to be autonomous with little or no human direction. As stated in the text, autonomous data mining is a dangerous practice.

**Evaluation** – The fifth phase entails thorough evaluation of both the (1) constructed models for their efficacy and performance as well as the (2) approach used to construct the models to ensure that the constructed models actually solve the business problem at hand. While this phase is similar to SEMMA's Assess stage, the CRISP-DM Evaluation phase verifies that the

for download complete version of solution (all chapter 1 to 28 ) [click here.](#)

models constructed actually solve the business problem at hand. Since SEMMA does not prescribe formal definition of the business problem to be solved, the SEMMA Assess stage may actually result in a model that performs well but operates on the wrong target variable and corresponding predictors and thus has little or no business value.

**Deployment** – The sixth and final phase entails preparing the model results so that it can be leveraged by the business sponsor. For simpler data mining projects, this may entail generating a report that the sponsor may use to base business decisions off of. For more complex projects, this may entail implementation of the final model in a commercial rules-engine software package. In direct contrast with SEMMA, there is no corresponding stage in the SEMMA process prescribing model deployment.

## Solutions to Chapter 2 DATA PREPROCESSING

Prepared by James Cunningham, Graduate Assistant

- 1. Describe the possible negative effects of proceeding directly to mine data that has not been preprocessed.**

Neglecting to preprocess the data adequately before data modeling begins will likely produce data models that are unreliable and whose results should be considered dubious at best. Performing data cleaning and data transformation during the data preparation phase is absolutely necessary for successful data mining efforts.

For example, suppose you are analyzing a data set that includes a person's Age and Date\_of\_Birth attributes, and you want to calculate the average Age. Now, if 5% of the records contain a value of 0 for Age, the mean value would be very misleading and inaccurate. One solution to this problem would be to derive Age for the zero-based records based on information contained in the Date\_of\_Birth variable. Now, the mean value for Age is more representative of those persons in the data set.

- 2. Refer to the income attribute of the five customers in Table 2.1, before preprocessing.**

- a. Find the mean income before preprocessing.**

The mean value for Income before preprocessing is 38,999.80 and is derived by the possible inclusion of Income values -40,000 (erroneous) and 100,000 (possible outlier).

- b. What does this number actually mean?**

In this case the mean value has little meaning because we are combining real data values with erroneous values.

- c. Now, calculate the mean income for the three values left after preprocessing. Does this value have a meaning?**

However, the mean value for Income produced by values 75,000, 50,000, and 10,000 (9,999 rounded to nearest 5,000) is 45,000. The latter value is certainly more representative of the true mean for Income, now that the records containing questionable values have been excluded.

**3. Explain why zip codes should be considered text variables rather than numeric.**

Zip codes should be considered text variables because they cannot be quantified on any numeric scale. Even their order has no numerical significance.

**4. What is an outlier? Why do we need to treat outliers carefully?**

Consider a set of numerical observations and the center of this observation set. An outlier is an observation that lies much farther away from the center than the majority of the other observations in the set.

We must treat outliers carefully because they can cause us to misrepresent the true center of an observation set incorrectly if they lie significantly farther away from the other observations in the set.

**5. Explain why a birthdate variable would be preferred to an age variable in a database.**

A birthdate variable is preferable to an age variable in a database because (1) one can always derive age from birthdate by taking the difference from the current date, and (2) age is relative to the current date only and would need to be updated continuously over time in order to remain accurate.

**6. True or false: All things being equal, more information is almost always better.**

The answer is true. In general, more information is almost always better. The more information we have to work with, the more insight into the underlying relationships of a particular domain of discourse we can glean from it.

**7. Explain why it is not recommended, as a strategy for dealing with missing data, to simply omit the records or fields with missing values from the analysis.**

It is not recommended to omit records or fields from an analysis simply because they have missing values. The rationale for this recommendation is that omitting these fields and records may cause us to lose valuable insight into the underlying relationships that we may have gleaned from the partial information that we do have.



**8. Which of the four methods for handling missing data would tend to lead to an underestimate of the spread (e.g., standard deviation) of the variable? What are some benefits to this method?**

Replacing a missing value by the attribute value's mean artificially reduces the measure of spread for that particular attribute. Although the mean value is not necessarily a typical value, for some data sets this form of substitution may work well. Specifically, the effectiveness of this technique depends on the size of the variation of the underlying population. In other words, the technique works well for populations having small variations, and works less effectively for populations having larger variations.

Several benefits to leveraging this method include (1) ease of implementation (i.e. only one value to impute), (2) preservation of the standard error (i.e. no additional residual error is introduced).

**9. What are some of the benefits and drawbacks for the method for handling missing data that chooses values at random from the variable distribution?**

By using the data values randomly generated from the variable distribution, the measures of center and spread are most likely to remain similar to the original; however, there is a chance that the resulting records may not make intuitive sense.

**10. Of the four methods for handling missing data, which method is preferred?**

Having the analyst choose a constant to replace missing values based on specific domain knowledge is overall, probably the most conservative choice. If missing values are replaced with a flag such as "missing" or "unknown", in many situations those records would ultimately be excluded from the modeling process; that is, all remaining valid, potentially important, values contained in those records would not be included in the data model.

11. Make up a classification scheme which is inherently flawed, and would lead to misclassification, as we find in Table 2.2. For example, classes of items bought in a grocery store.

Breakfast	Count
Cold Cereals	72
Sugar Smacks	1
Cheerios	2
Hot Cereals	28
Cream of Wheat	3

Using the table above, the “Breakfast” categorical attribute contains 5 apparent classes. However, upon further inspection the classes are discovered to be inconsistent. For example, both “Sugar Smacks” and “Cheerios” are cold cereals, and “Cream of Wheat” is a hot cereal. Below, the cereals are now classified according to one of two classes, “Cold Cereals” or “Hot Cereals.”

Breakfast	Count
Cold Cereals	75
Hot Cereals	31

12. Make up a data set, consisting of the heights and weights of six children, in which one of the children is an outlier with respect to one of the variables, but not the other. Then alter this data set so that the child is an outlier with respect to both variables.

In the table below, Child #1 is an outlier with respect to Weight only. All children in the table are close in Height differing at most by 9 inches. However, all children except for Child # 1 are close in Weight differing at most by 7 pounds. Child #1 is an outlier as the Weight differs by 18 pounds from the second-heaviest child (Child #6), making this right-tailed difference in Weight greater than the entire Weight range for the other five children.

Child	Height (in)	Weight (lbs)
1	49	100
2	50	75
3	52	77
4	55	79
5	57	80
6	58	82

In the table below, Child #1 is an outlier with respect to both Height and Weight. All children except for Child #1 in the table are close in Height differing at most by 8 inches and are close in Weight differing at most by 7 pounds. Child #1 is an outlier for both Height and Weight as the Height differs by 14 inches from the second-shortest child (Child#2) (which is greater than the entire Height range of the other five children), and

for download complete version of solution (all chapter 1 to 28 ) click here.

the Weight differs by 18 pounds from the second-heaviest child (Child #6) (which is greater than the entire Weight range of the other five children).

Child	Height (in)	Weight (lbs)
1	36	100
2	50	75
3	52	77
4	55	79
5	57	80
6	58	82

Use the following stock price data (in dollars) for Exercises 13–18

---

10	7	20	12	75	15	9	18	4	12	8	14
----	---	----	----	----	----	---	----	---	----	---	----

---

**13. Calculate the mean, median, and mode stock price.**

The *mean* is calculated as the sum of the data points divided by the number of points as follows:

$$\text{Mean Stock Price} = (10+7+20+12+75+15+9+18+4+12+8+14) / 12 = 204 / 12 = \$17.$$

The *median* is calculated by placing the prices in order and (a) selecting the middle value if the number of points is odd, or (b) taking the average of the two middle values if the number of points is even. Since we have twelve points, median is calculated as follows:

$$\text{Median Stock Price} = \text{mean of center values } \{4,7,8,9,10,12,12,14,15,18,20,75\} = 24/2 = \$12.$$

The *mode* is calculated as the value that occurs the most often in the set and is calculated as follows:

$$\text{Mode Stock Price} = \text{highest frequency of } \{4,7,8,9,10,12,12,14,15,18,20,75\} = \$12.$$

**14. Compute the standard deviation of the stock price. Interpret what this number means.**

The *standard deviation* represents the expected distance of a point chosen at random from a data set to the center of that set and is calculated by taking the square root of the *variance*. The variance is the average of the sum of squared distances of each point from the data-set mean. Given that the mean is \$17 (see Exercise #13) for this set, the variance for the set of stock prices is calculated as follows:

Stock Price Variance (Var) =

$$(4-17)^2+(7-17)^2+(8-17)^2+(9-17)^2+(10-17)^2+(12-17)^2+(12-17)^2+(14-17)^2+(15-17)^2+(18-17)^2+(20-17)^2+(75-17)^2 =$$
$$(-13)^2 + (-10)^2 + (-9)^2 + (-8)^2 + (-7)^2 + (-5)^2 + (-5)^2 + (-3)^2 + (-2)^2 + (1)^2 + (3)^2 + (58)^2 =$$
$$169 + 100 + 81 + 64 + 49 + 25 + 25 + 9 + 4 + 1 + 9 + 3364 = 3900 / 12 = \mathbf{325 \$^2}.$$

Taking the square root of the Variance, the Standard Deviation (SD) is calculated as follows:

$$\text{Stock Price Standard Deviation (SD) of Stock Price} = \sqrt{(325)} = \pm\mathbf{\$18.03}.$$

Since the mean is \$17 and the standard deviation is plus/minus \$18.03, the expected price of a stock drawn at random from the set of twelve stocks is expected to lie mathematically between  $(\$17-\$18.03) = \mathbf{-\$1.03}$  (i.e. \$0.01 since we assume that a stock price can never be less than one penny USD) and  $(\$17+\$18.03) = \mathbf{\$35.03}$ .

As we can see, each stock with the exception of the one priced at \$75 is priced within this range.

**15. Find the min-max normalized stock price for the stock worth \$20.**

Min-Max normalization scales an observation relative to the data-set's range resulting in a value between 0 and 1 (this value has no units) and is formulated as follows:

$$\text{MinMax}X_i = [X_i - \text{Min}(X)] / [\text{Max}(X) - \text{Min}(X)]$$

Therefore, the min-max normalized stock price of \$20 is calculated as follows:

$$\text{MinMax}(\$20) = (\$20 - \$4) / (\$75 - \$4) = (\$16) / (\$71) = \mathbf{0.2254}.$$

**16. Calculate the midrange stock price.**

The midrange stock price is the central price for the entire price range and is formulated as follows:

$$\text{MidRangeX} = [\text{Max}(X) + \text{Min}(X)] / 2$$

For the problem at hand we have as follows:

$$\text{MidRangeX} = (\$75 + \$4) / 2 = (\$79) / 2 = \mathbf{\$39.5}$$

**17. Compute the Z-score standardized stock price for the stock worth \$20.**

Z-Score standardization scales an observation where the mean value is zero, the SD is 1 and most values lie between -4 and 4 (this value has no units) and is formulated as follows:

$$\text{Z-Score}(X) = [X_i - \text{Mean}(X)] / |\text{SD}(X)|$$

Given the mean of \$17 (see Exercise #13) and |SD| of 18.03 (see Exercise #14), The Z-Score for the stock price of \$20 is calculated as follows:

$$\text{Z-Score}(\$20) = (\$20 - \$17) / \$18.03 = (\$3) / \$18.03 = \mathbf{0.1664}.$$

Please note that this value makes sense as it is slightly greater than zero just as \$20 is slightly greater than \$18.03.

**18. Find the decimal scaling stock price for the stock worth \$20.**

Decimal standardization scales an observation to a value between -1 and 1 (this value has no units) and is formulated as follows:

$$\text{Decimal}(X_i) = X_i / 10^d$$

where d is the number of digits in the observation in the data set having the largest absolute value. Since the largest stock price is \$75, d = 2 as there are two digits in this price. The decimal standardization is then calculated as follows:

$$\text{Decimal}(\$75) = \$75 / \$10^2 = \$75 / \$100 = \mathbf{0.75}$$

**19. Calculate the skewness of the stock price data.**

Skewness is the lack of normalization of a Z-Score-standardized distribution and is measured using the following formula:

$$\text{Skewness} = 3 [\text{Mean}(X) - \text{Median}(X)] / \text{SD}(X)$$

Given the mean of \$17 and median of \$12 (see Exercise #13), and an SD of \$18.03 (see Exercise #14), the skewness for the stock price distribution is calculated as follows:

$$\text{Skewness} = 3 [\$17 - \$12] / \$18.03 = 3[\$5] / \$18.03 = \$15 / \$18.03 = \mathbf{0.8319}.$$

We observe that this distribution is right-skewed since a right-skewed distribution has a mean that is greater than its median yielding a positive skewness value. In contrast, a left-skewed distribution will have a mean that is less than its median and thus a negative skewness value.

**20. Explain why data analysts need to normalize their numeric variables.**

Data analysts need to normalize their numeric variables as it places all variables on the same scale. Normalizing all variables to the same scale is critical when performing operations that are sensitive to data variation or *spread* so that variables having larger variations do not adversely overpower variables having smaller variations. Most (if not all) analytic operations involving linearization (e.g. Regression, PCA, MANOVA, etc.) are sensitive to data spread.

**21. Describe three characteristics of the standard normal distribution.**

The three main characteristics of the Standard Normal Distribution are as follows:

- The mean is zero
- The SD is 1
- It is symmetric (equal and opposite in shape and size) about the mean and normal (the mean has the highest frequency, and frequency decreases symmetrically as distance from the mean increases).

**22. If a distribution is symmetric, does it follow that it is normal? Give a counterexample.**

If a distribution is symmetric, it is not guaranteed to be normal. In order for a distribution to be normal it has to have a single expected value (i.e. the value with the highest frequency).

A classic counterexample is the Uniform Distribution, which is symmetric about the center of its interval, yet since it all values on the interval occur with equal frequency, it has an infinite number of expected values making it non-normal.

**23. What do we look for in a normal probability plot to indicate non-normality?**

A normal probability plot is simply a plot of the quantiles of a given distribution to the quantiles of the Standard Normal Distribution. If the quantiles are approximately equal, then the plot will approximate a straight line indicating that the given distribution is normal.

In contrast, if the quantiles of the distribution are not equal to the Standard Normal Distribution, then the plot will not approximate a straight line indicating non-normality.

**Use the stock price data for Exercises 24–26.**

**24. Do the following:**

**a. Identify the outlier.**

The outlier is the stock price of \$75. The difference from the next-closest stock price (\$20) is \$55, which is nearly 3.5X larger than the entire range of the other eleven stocks (i.e. \$16).

**b. Verify that this value is an outlier, using the Z-score method.**

We can also verify that \$75 is in fact an outlier using the Z-score method. The Z-score for this stock is calculated using our mean of \$17 (see Exercise #13) and our SD of \$18.03 (see Exercise #14) as follows:

$$\text{Z-Score}(\$75) = (\$75 - \$17) / \$18.03 = (\$58) / \$18.03 = \mathbf{3.2169}.$$

Since a Z-score that is less than -3 or greater than 3 is considered an outlier, we conclude that stock price \$75 is an outlier as its Z-score is 3.2169 which is greater than 3.

**c. Verify that this value is an outlier, using the IQR method.**

We can also verify that \$75 is in fact an outlier using the Inter-Quartile Range or IQR method. The quartiles are determined by placing the stock prices in ascending order and dividing them onto four parts as follows:

The ordered stock prices are: {4,7,8,9,10,12,12,14,15,18,20,75}, and since there are an even number of values, we partition as {4,7,8,9,10,12} and {12,14,15,18,20,75}

The quartiles are then determined as follows:

$$Q1 = \{4,7,\mathbf{8},9,10,12\} = \$8$$

$$Q3 = \{12,14,\mathbf{15},18,20,75\} = \$15$$

We then calculate  $IQR = Q3 - Q1$  as follows:

$$\text{IQR} = \$15 - \$8 = \$7$$

If an observation is an outlier, then it will have a value that is less than  $Q1 - 1.5\text{IQR}$  or a value greater than  $Q3 + 1.5\text{IQR}$ . We then calculate the upper and lower boundary values for the stock price set as follows:

$$\text{LowerBound} = Q1 - 1.5\text{IQR} = 8 - 1.5(7) = 8 - 10.5 = \mathbf{-\$2.50}$$

$$\text{UpperBound} = Q3 + 1.5\text{IQR} = 15 + 1.5(7) = 15 + 10.5 = \mathbf{\$25.50}$$

Since \$75 is greater than \$25.5, we conclude that \$75 is an outlier.

## 25. Identify all possible stock prices that would be outliers, using:

### a. The Z-score method.

The ordered stock prices are: {4,7,8,9,10,12,12,14,**15,18**,20,75} where the mean is \$17 lying between the \$15 and \$18 stock indicated in bold text, and the SD is \$18.03. Working from the left, we have as follows:

$$\text{Z-Score}(\$4) = (\$4 - \$17) / \$18.03 = (-\$13) / \$18.03 = \mathbf{-0.7210}.$$

$$\text{Z-Score}(\$7) = (\$7 - \$17) / \$18.03 = (-\$10) / \$18.03 = \mathbf{-0.5546}.$$

$$\text{Z-Score}(\$8) = (\$8 - \$17) / \$18.03 = (-\$9) / \$18.03 = \mathbf{-0.4992}.$$

$$\text{Z-Score}(\$9) = (\$9 - \$17) / \$18.03 = (-\$8) / \$18.03 = \mathbf{-0.4437}.$$

$$\text{Z-Score}(\$10) = (\$10 - \$17) / \$18.03 = (-\$7) / \$18.03 = \mathbf{-0.3882}.$$

$$\text{Z-Score}(\$12) = (\$12 - \$17) / \$18.03 = (-\$5) / \$18.03 = \mathbf{-0.2773}.$$

$$\text{Z-Score}(\$14) = (\$14 - \$17) / \$18.03 = (-\$3) / \$18.03 = \mathbf{-0.1664}.$$

$$\text{Z-Score}(\$15) = (\$15 - \$17) / \$18.03 = (-\$2) / \$18.03 = \mathbf{-0.1109}.$$

$$\text{Z-Score}(\$18) = (\$18 - \$17) / \$18.03 = (\$1) / \$18.03 = \mathbf{0.0555}.$$

$$\text{Z-Score}(\$20) = (\$20 - \$17) / \$18.03 = (\$3) / \$18.03 = \mathbf{0.1664}.$$

We already know that \$75 is an outlier having a Z-score of **3.2169** (see Exercise #24). However, no other outliers were identified using Z-score standardization.



**b. The IQR method.**

The ordered stock prices are: {4,7,8,9,10,12,12,14,**15,18**,20,75} where we have an IQR of \$7, a Lower Bound of \$2.5, and an Upper Bound of \$25.50.

Therefore, stock prices \$75 is once again the only outlier as it is greater than the upper bound of \$25.50.

**26. Investigate how the outlier affects the mean and median by doing the following:**

**a. Find the mean score and the median score, with and without the outlier.**

The mean for the entire set of stock prices is \$17 (see Exercise #13), and the mean without the \$75 outlier is calculated as follows:

$$\text{Mean}_{\text{No\_Outlier}} = (10+7+20+12+15+9+18+4+12+8+14) / 11 = 129 / 11 = \$11.73.$$

The *median* is calculated by placing the prices in order and (a) selecting the middle value if the number of points is odd, or (b) taking the average of the two middle values if the number of points is even. Since we have twelve points, median is calculated as follows:

$$\text{Median Stock Price} = \text{mean of center values } \{4,7,8,9,10,\mathbf{12,12},14,15,18,20,75\} = 24/2 = \$12.$$

$$\text{Median}_{\text{No\_Outlier}} = \text{mean of center values } \{4,7,8,9,10,\mathbf{12,12},14,15,18,20\} = \$12.$$

**b. State which measure, the mean or the median, the presence of the outlier affects more, and why.**

It is obvious that the presence of the outlier affects the mean more than the median. It increases the mean by \$5, and has no effect on the median.

For this particular data set, the outlier affects the mean more than the median because the mean determines the numerical center of the data set through interpolation and this data is right-skewed having a large right-tailed outlier. In contrast, the median determines the distributive center of the dataset through physical partitioning and the largest value of the lower half of the data is equal to the smallest value of the upper half of this data set.

**27. What are the four common methods for binning numerical predictors? Which of these are preferred?**

The four common methods for binning numerical predictors are as follows:

1. **Equal width binning** – this method divides into k categories of equal width chosen by the client or analyst.
2. **Equal frequency binning** - this method divides the numerical predictor into k categories, each having k/n records, where n is the total number of records.
3. **Binning by clustering** – this method uses a clustering algorithm, such as k-means clustering.
4. **Binning based on predictive value** - this method partitions the numerical predictor based on the effect each partition has on the value of the target variable.

The preferred methods are Binning by Clustering (method #3) and Binning based on predictive value (method #4). Both methods determine the partitions by the nature of the data and its underlying relationships.

In contrast, the Equal Width Binning (method #1) and Equal Frequency Binning (method #2) determine the partitions simply by their individual numeric values. In general, Equal Width Binning should not be used for anything more than rough exploration as the use of equal width bins is very susceptible to outliers. The method of Equal Frequency Binning is inherently flawed in that it can produce bins having overlapping values.

**Use the following data set for Exercises 28–30: 1 1 1 3 3 7**

**28. Bin the data into three bins of equal width (width = 3).**

Using equal-width binning with width=3, the bin boundaries are calculated as follows:

- Bin1:  $0 \leq X < 3$  containing {1,1,1}  
Bin2:  $3 \leq X < 6$  containing {3,3}  
Bin3:  $6 \leq X < 9$  containing {7}

**29. Bin the data into three bins of two records each.**

Binning this set into three bins of two records each is an application of equal-frequency binning with k=3, and since n=6, the bin size is  $k/n \Rightarrow 3/6 = 2$ . The bins are as follows:

- {1,1}, {1,3}, {3,7}

**30. Clarify why each of the binning solutions above are not optimal.**

Although this toy data set is relatively small and the clarification may not be obvious, the equal-width binning from Exercise #28 is suboptimal since Bin3 contains the outlier value 7, giving the illusion that the discretized class Bin3 is just as close to Bin2 as Bin1 when in fact it is much farther away.

The equal-frequency binning from Exercise #29 is also suboptimal since the three bins contain overlapping values. For example, the value 1 lies in both the first and second bins, and the value 3 lies in both the second and third bins. Therefore, models constructed from this new set of discretized classes are bound to produce unpredictable results.

**31. Explain why we might not want to remove a variable that had 90% or more missing values.**

In general, analysts should not remove variables that have a large number of missing values. The small number of values that are present may in fact be representative of the underlying population, and it would therefore be worthwhile to attempt to impute the missing values. However, if the small of values are not representative of the underlying population, the fact that the variable has missing values may actually be able to be correlated to other variables and produce predictive power that would have been lost if this data were discarded.

**32. Explain why we might not want to remove a variable just because it is highly correlated with another variable.**

In general, analysts should not remove variables, even when they are highly correlated. An analyst may be tempted to remove one of a pair of highly correlated variables in order to avoid over-emphasizing a particular informational characteristic. However, although removing a highly correlated variable avoids potential “double-counting”, doing so may also cause a loss of valuable predictive relationships to other variables that the target variable is not highly correlated with.

As an alternative to removing highly correlated variables, it is recommended that the analyst employ Principle Component Analysis to translate the highly correlated variables into a set of uncorrelated principal components.

for download complete version of solution (all chapter 1 to 28 ) click here.

## HANDS-ON ANALYSIS

Use the churn data set on the book series website for the following exercises.

The Churn data set is located at the following location:

<http://www.dataminingconsultant.com/DKD2e.htm>

### 33. Explore whether there are missing values for any of the variables.

An analysis using SPSS Modeler v16 identifies that there are no missing values for any variables and is depicted below:

Field	Measurement Level	Null	Blank	Value	Invalid Values	Valid	% Valid	Valid Records	Null Value	Empty Value	Invalid Record	Blank Value
State	Categorical	---	---	Never	Never	130	100%	130	0	0	0	0
Area Code	Categorical	7	None	Never	None	130	100%	130	0	0	0	0
Area Code 2	Categorical	4	None	Never	None	130	100%	130	0	0	0	0
Phone	Categorical	---	---	Never	Never	130	100%	130	0	0	0	0
Area Code 3	Categorical	---	---	Never	Never	130	100%	130	0	0	0	0
Area Code 4	Categorical	---	---	Never	Never	130	100%	130	0	0	0	0
Area Code 5	Categorical	---	---	Never	Never	130	100%	130	0	0	0	0
Area Code 6	Categorical	---	---	Never	Never	130	100%	130	0	0	0	0
Area Code 7	Categorical	---	---	Never	Never	130	100%	130	0	0	0	0
Area Code 8	Categorical	---	---	Never	Never	130	100%	130	0	0	0	0
Area Code 9	Categorical	---	---	Never	Never	130	100%	130	0	0	0	0
Area Code 10	Categorical	---	---	Never	Never	130	100%	130	0	0	0	0
Area Code 11	Categorical	---	---	Never	Never	130	100%	130	0	0	0	0
Area Code 12	Categorical	---	---	Never	Never	130	100%	130	0	0	0	0
Area Code 13	Categorical	---	---	Never	Never	130	100%	130	0	0	0	0
Area Code 14	Categorical	---	---	Never	Never	130	100%	130	0	0	0	0
Area Code 15	Categorical	---	---	Never	Never	130	100%	130	0	0	0	0
Area Code 16	Categorical	---	---	Never	Never	130	100%	130	0	0	0	0
Area Code 17	Categorical	---	---	Never	Never	130	100%	130	0	0	0	0
Area Code 18	Categorical	---	---	Never	Never	130	100%	130	0	0	0	0
Area Code 19	Categorical	---	---	Never	Never	130	100%	130	0	0	0	0
Area Code 20	Categorical	---	---	Never	Never	130	100%	130	0	0	0	0
Area Code 21	Categorical	---	---	Never	Never	130	100%	130	0	0	0	0
Area Code 22	Categorical	---	---	Never	Never	130	100%	130	0	0	0	0
Area Code 23	Categorical	---	---	Never	Never	130	100%	130	0	0	0	0
Area Code 24	Categorical	---	---	Never	Never	130	100%	130	0	0	0	0
Area Code 25	Categorical	---	---	Never	Never	130	100%	130	0	0	0	0
Area Code 26	Categorical	---	---	Never	Never	130	100%	130	0	0	0	0
Area Code 27	Categorical	---	---	Never	Never	130	100%	130	0	0	0	0

### 34. Compare the area code and state fields. Discuss any apparent abnormalities.

Although each record has a value for Area Code and for State, there are only three distinct values for Area Code in the entire data set (408, 415, and 510), and each of the three values for Area Code is associated to each of the values for State. This presents an abnormality as each of these three Area Codes are for the state of California, so each record that does not have the State code CA has an invalid combination of State and Area Code values. The complete distribution of State and Area Code mappings are depicted below.

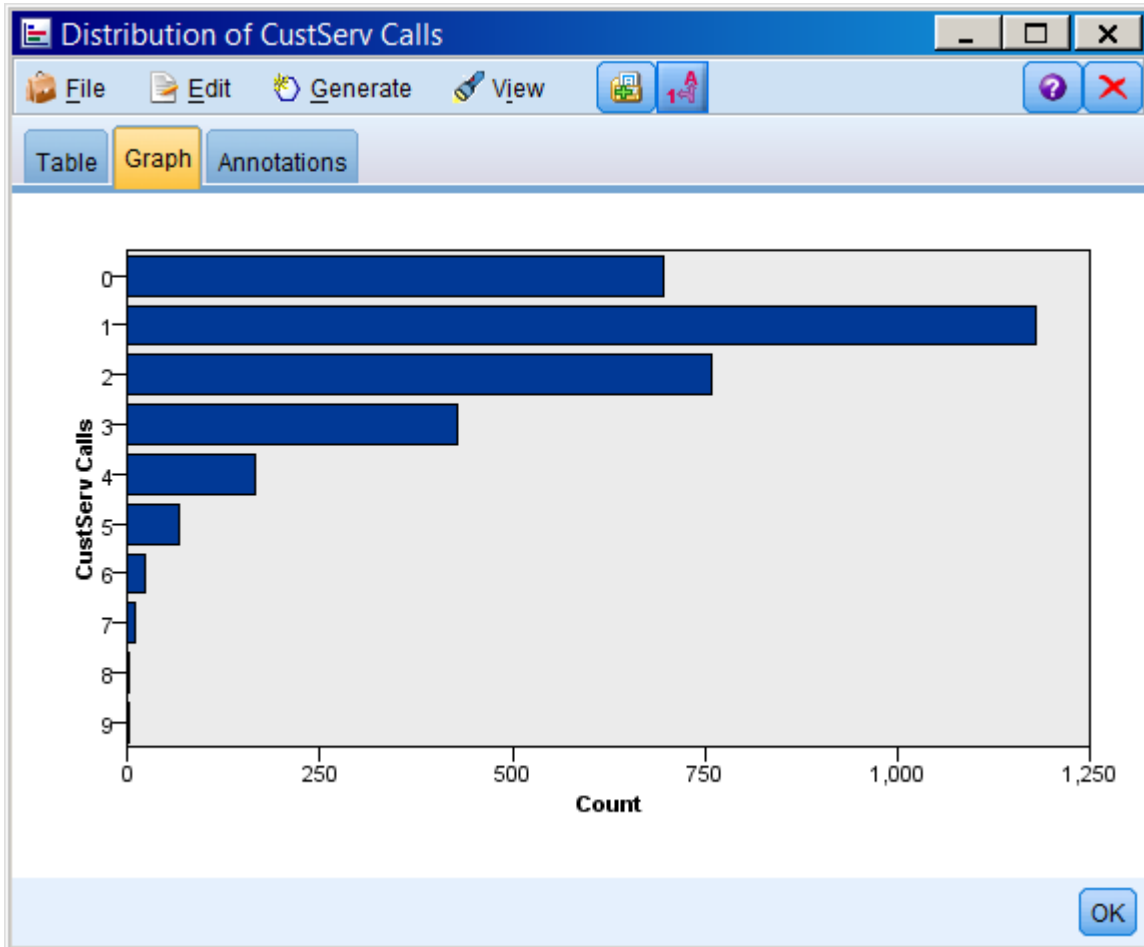
for download complete version of solution (all chapter 1 to 28 ) click here.

State	408	415	510
AK	14	24	14
AL	25	40	15
AR	13	27	15
AZ	15	36	13
CA	7	17	10
CO	25	29	12
CT	22	39	13
DC	14	27	13
DE	13	31	17
FL	12	31	20
GA	15	21	18
HI	15	30	8
IA	8	20	16
ID	12	41	20
IL	15	28	15
IN	18	33	20
KS	12	37	21
KY	15	32	12
LA	13	27	11
MA	24	29	12
MD	16	39	15
ME	15	25	22
MI	12	39	22
MN	20	40	24
MO	15	37	11
MS	15	31	19
MT	17	34	17
NC	25	28	15
ND	19	28	15
NE	13	34	14
NH	25	19	12
NJ	15	34	19
NM	16	35	11
NV	14	34	18
NY	19	47	17
OH	22	40	16
OK	17	27	17
OR	14	44	20
PA	14	19	12
RI	12	35	18
SC	13	30	17
SD	16	28	16
TN	11	30	12
TX	20	37	15
UT	12	37	23
VA	25	35	17
VT	17	36	20
WA	23	26	17
WI	22	35	21
WV	20	52	34
WY	17	41	19

for download complete version of solution (all chapter 1 to 28 ) click here.

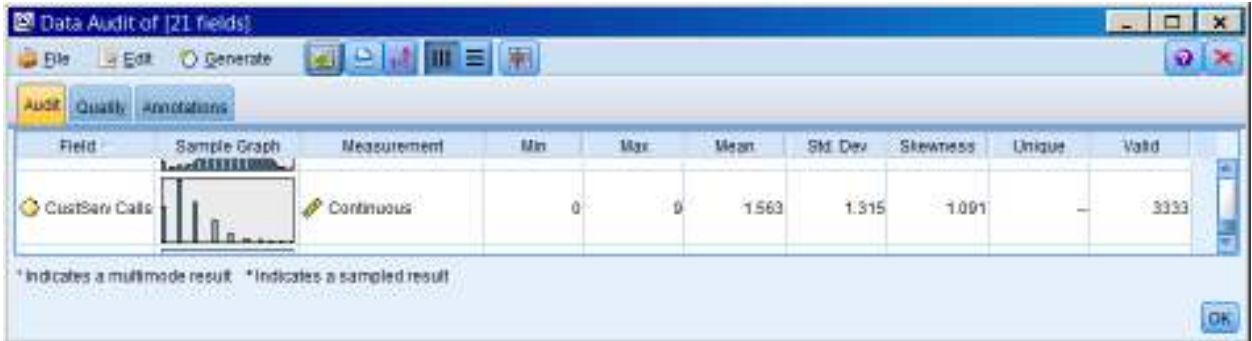
**35. Use a graph to visually determine whether there are any outliers among the number of calls to customer service.**

Using Modeler, we set the type the Customer Service Calls to Ordinal and produce a Distribution graph as depicted below in order to identify potential outliers. The center of the distribution appears to lie around 1 service call making potential outliers appear to be at 7, 8, and 9 service calls.



36. Identify the range of customer service calls that should be considered outliers, using:
- The Z-score method, and

We treat the Customer Service Calls as Continuous, which gives us a mean of **1.563**, and an SD of **1.315** as depicted in the Data Audit below.



The Z-scores for the number of service calls in descending order are as follows:

$$\begin{aligned} Z\text{-Score}(9) &= (9 - 1.563) / 1.315 = 7.437 / 1.315 = \mathbf{5.656} \Rightarrow \text{Outlier} \\ Z\text{-Score}(8) &= (8 - 1.563) / 1.315 = 6.437 / 1.315 = \mathbf{4.885} \Rightarrow \text{Outlier} \\ Z\text{-Score}(7) &= (7 - 1.563) / 1.315 = 5.437 / 1.315 = \mathbf{4.134} \Rightarrow \text{Outlier} \\ Z\text{-Score}(6) &= (6 - 1.563) / 1.315 = 4.437 / 1.315 = \mathbf{3.374} \Rightarrow \text{Outlier} \\ Z\text{-Score}(5) &= (5 - 1.563) / 1.315 = 3.437 / 1.315 = 2.614 \Rightarrow \text{Not an Outlier} \\ Z\text{-Score}(4) &= (4 - 1.563) / 1.315 = 2.437 / 1.315 = 1.853 \Rightarrow \text{Not an Outlier} \\ Z\text{-Score}(3) &= (3 - 1.563) / 1.315 = 1.437 / 1.315 = 1.093 \Rightarrow \text{Not an Outlier} \\ Z\text{-Score}(2) &= (2 - 1.563) / 1.315 = 0.685 / 1.315 = 0.521 \Rightarrow \text{Not an Outlier} \\ Z\text{-Score}(1) &= (1 - 1.563) / 1.315 = -0.563 / 1.315 = -0.428 \Rightarrow \text{Not an Outlier} \\ Z\text{-Score}(0) &= (0 - 1.563) / 1.315 = -1.563 / 1.315 = -1.189 \Rightarrow \text{Not an Outlier} \end{aligned}$$

**b. The IQR method.**

The value set is  $\{0,1,2,3,4,5,6,7,8,9\}$ . Since there is an even number of values, we partition between values 4 and 5 as  $\{0,1,2,3,4\}$  and  $\{5,6,7,8,9\}$ . This yields the four quartiles as follows:

$$\begin{aligned} Q1 &= \{0,1,2,3,4\} = 2 \\ Q3 &= \{5,6,7,8,9\} = 7 \end{aligned}$$

We then calculate  $IQR = Q3 - Q1$  as follows:

$$IQR = 7 - 2 = \mathbf{5}.$$

for download complete version of solution (all chapter 1 to 28 ) click here.

Using the IQR, we calculate the upper and lower boundaries as follows:

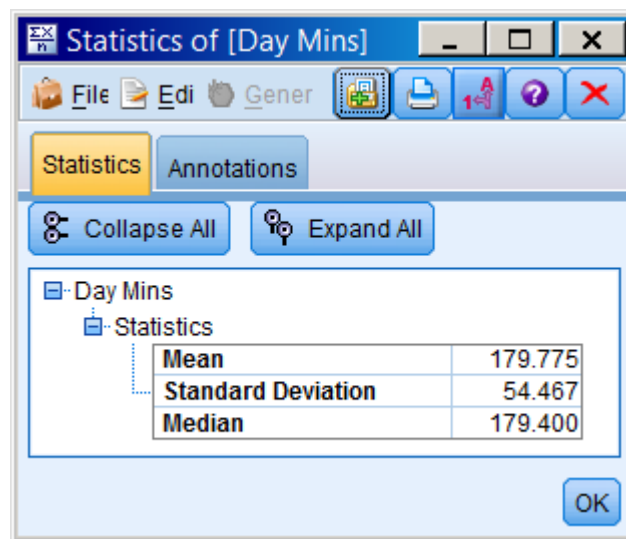
$$\text{LowerBound} = Q1 - 1.5\text{IQR} = 1 - 1.5(5) = 2 - 7.5 = -5.5$$

$$\text{UpperBound} = Q3 + 1.5\text{IQR} = 7 + 1.5(5) = 7 + 7.5 = 14.5$$

The IQR method does not identify any outliers as all values lie within the upper and lower quartile boundaries.

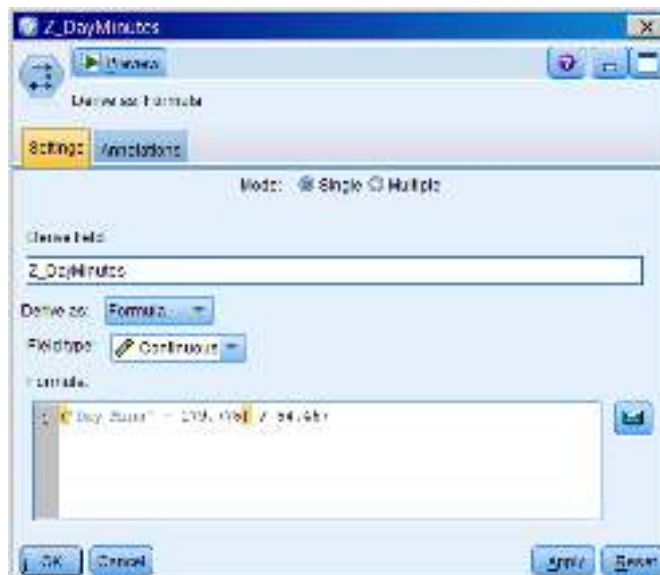
### 37. Transform the day minutes attribute using Z-score standardization.

We treat the Day Minutes as Continuous, which gives us a mean of **179.775**, and an SD of **54.467** as depicted in the Summary Statistics below.



Statistics of [Day Mins]	
Mean	179.775
Standard Deviation	54.467
Median	179.400

We then derive a new field called **Z\_DayMinutes** specifying the Z-score calculation as depicted below.





**38. Work with skewness as follows:**

**a. Calculate the skewness of day minutes.**

Recall that skewness is formulated as follows:

$$\text{Skewness} = 3 [\text{Mean}(X) - \text{Median}(X)] / \text{SD}(X)$$

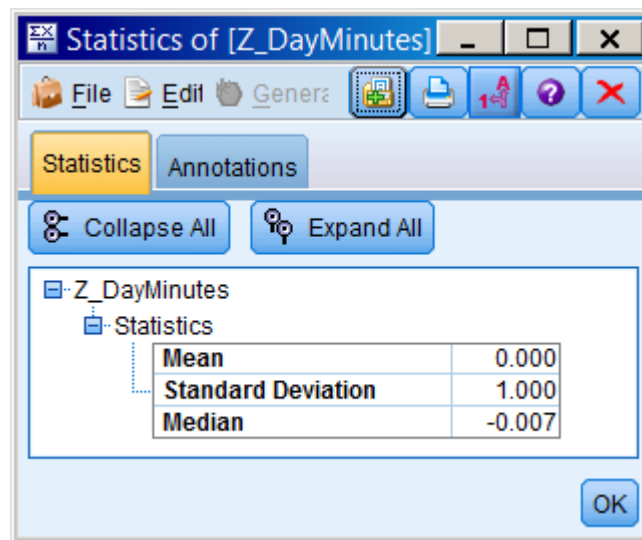
From Exercise #37 we have mean **179.775**, median **179.400**, and SD **54.467**, and calculate the skewness of Day Minutes as follows:

$$\text{Skewness}_{\text{DayMinutes}} = 3 [179.775 - 179.400] / 54.467 = [1.125] / 54.467 = \mathbf{0.021}.$$

**b. Then calculate the skewness of the Z-score standardized day minutes.**

**Comment.**

We calculate the summary statistics of the derived variable Z\_DayMinutes as depicted below.



We calculate the skewness of Z\_DayMinutes as follows:

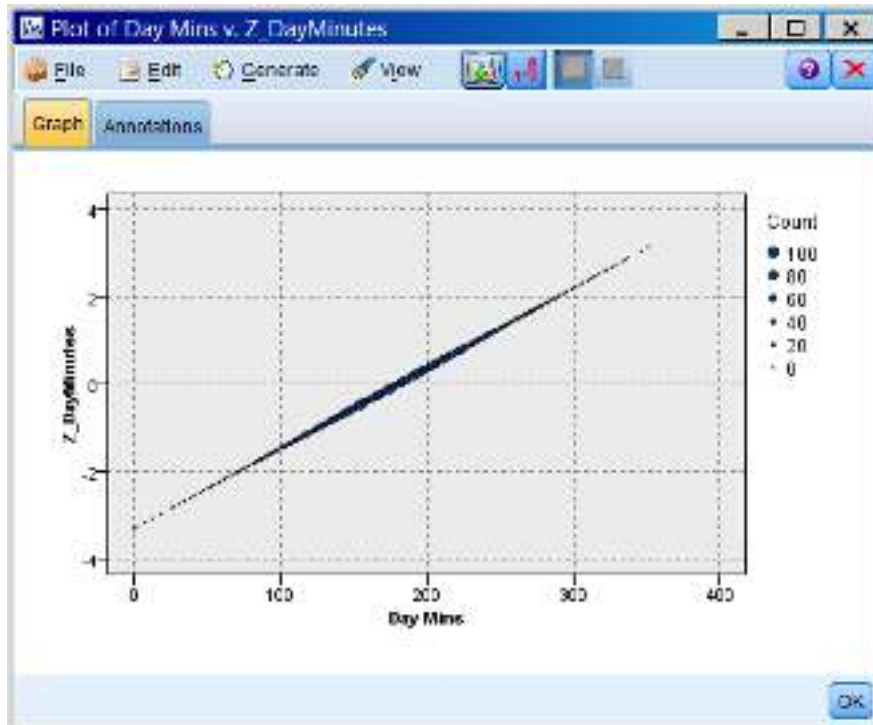
$$\text{Skewness}_{\text{Z\_DayMinutes}} = 3 [0.000 - (-0.007)] / 1.000 = [0.007] / 1.000 = \mathbf{0.007}.$$

**c. Based on the skewness value, would you consider day minutes to be skewed or nearly perfectly symmetric?**

We conclude that the Day Minutes values are nearly perfectly symmetric based on the calculated skewness value.

**39. Construct a normal probability plot of day minutes. Comment on the normality of the data.**

The normal probability plot of day minutes is calculated by first sorting the data set in ascending order by DayMins and then plotting  $Z_{\text{DayMinutes}}$  by DayMins as depicted below. We observe that there is an almost perfect linear relationship between the Z-score quantiles and the raw data quantiles indicating that the DayMins values are almost perfectly normally distributed. Please note that this aligns well with the near-zero skewness observed in Exercise #38.

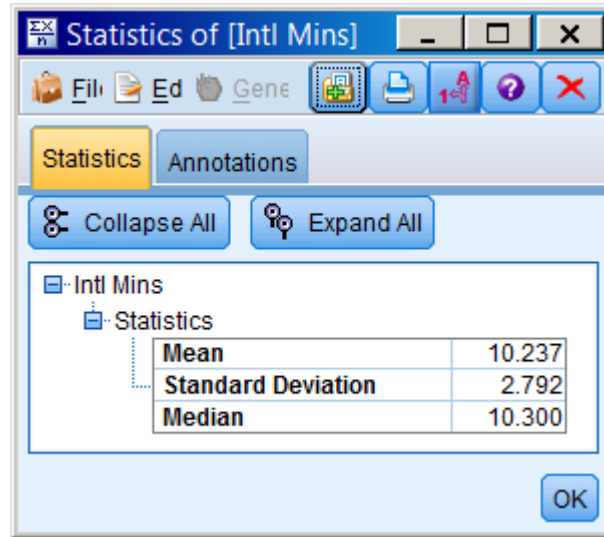


for download complete version of solution (all chapter 1 to 28 ) click here.

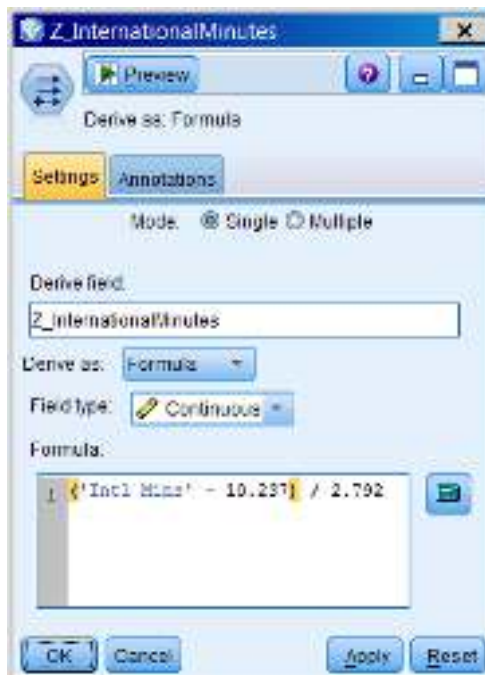
40. Work with international minutes as follows:

- a. Construct a normal probability plot of international minutes.

We treat the International Minutes as Continuous and calculate the mean **10.237** and the SD **2.792** as depicted in the summary statistics below.

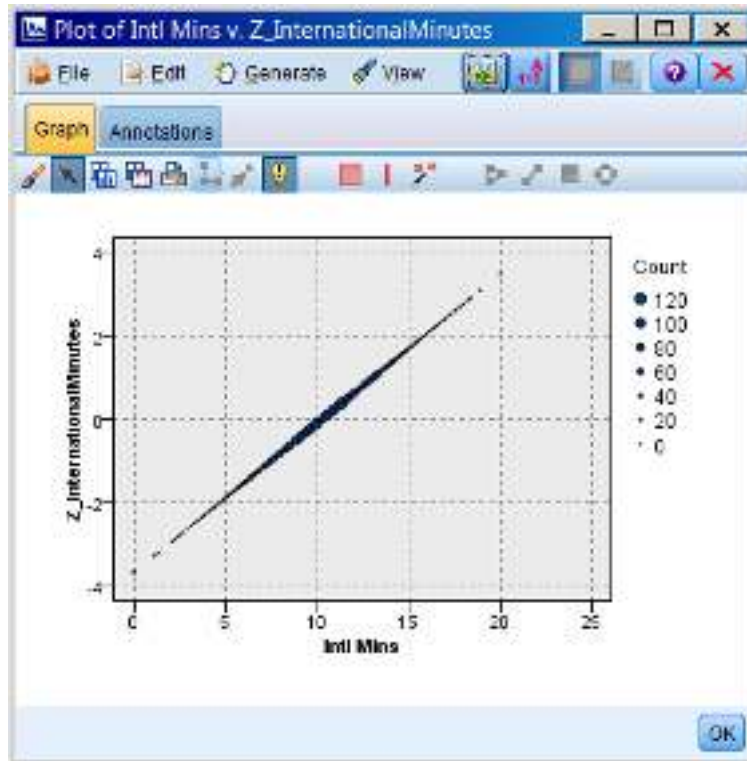


We then derive a new variable named **Z\_InternationalMinutes** specifying the Z-score calculation as depicted below.



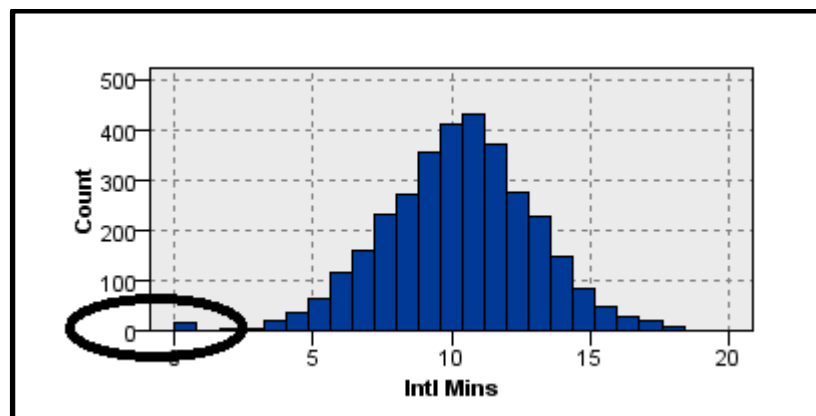
for download complete version of solution (all chapter 1 to 28 ) click here.

Finally, we construct a Normal Probability Plot by first sorting the data set by International Minutes values in ascending order and plotting  $Z_{\text{InternationalMinutes}}$  by IntlMins as depicted below.



**b. What is stopping this variable from being normally distributed?**

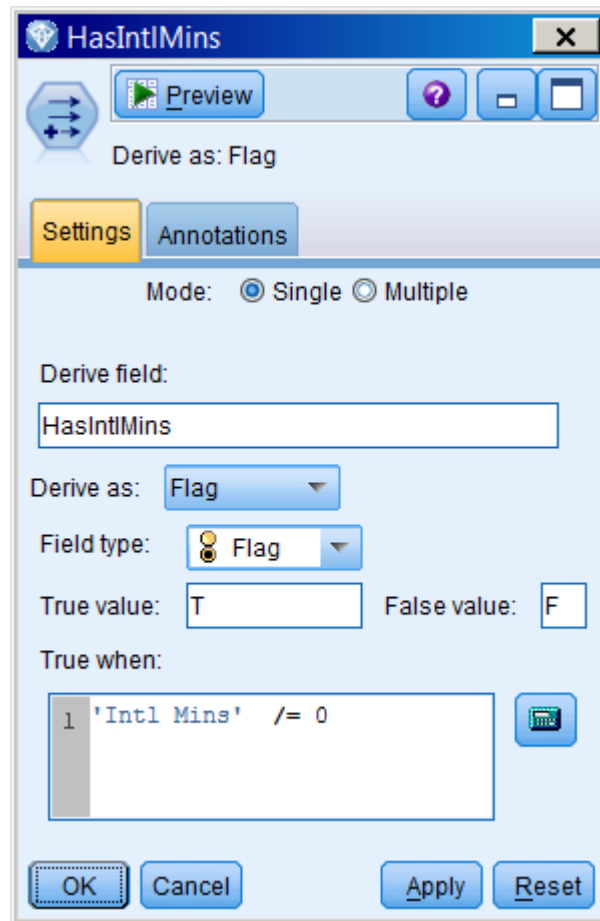
There are 18 records in the data set that have zero International Minutes, and this is preventing this variable from being normally distributed as depicted below.



for download complete version of solution (all chapter 1 to 28 ) click here.

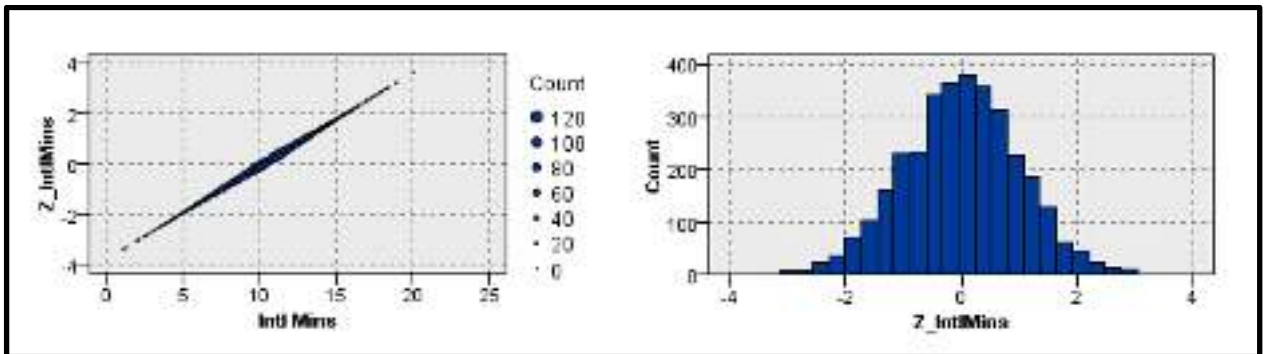
c. Construct a flag variable to deal with the situation in (b).

In order to perform analyses requiring normally distributed data, we can construct a flag variable that we can use to partition the variable into the subset that has international minutes and the subset that has no international minutes. The HasIntlMins flag is set to true when the number of minutes is non-zero and false otherwise as depicted below.



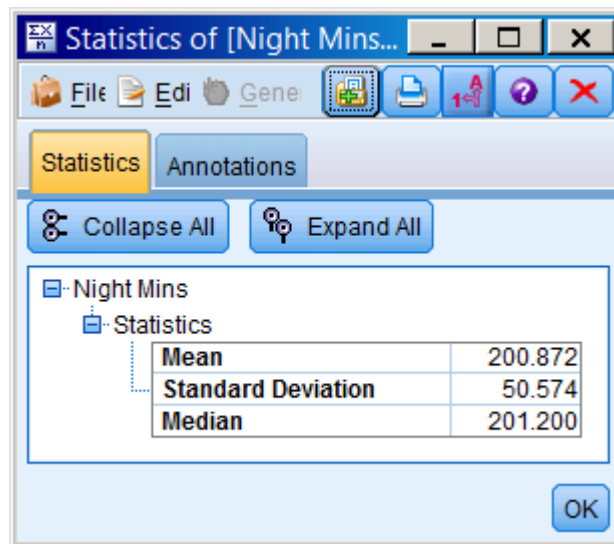
d. Construct a normal probability plot of the derived variable *nonzero international minutes*. Comment on the normality.

We sort the partition where HasIntlMins is true by the number of International Minutes in ascending order, calculate the Z-scores for each and create the Normal Probability Plot as depicted below. We have also included a histogram of the Z-scored non-zero international minutes in order to observe the near-perfect normality.



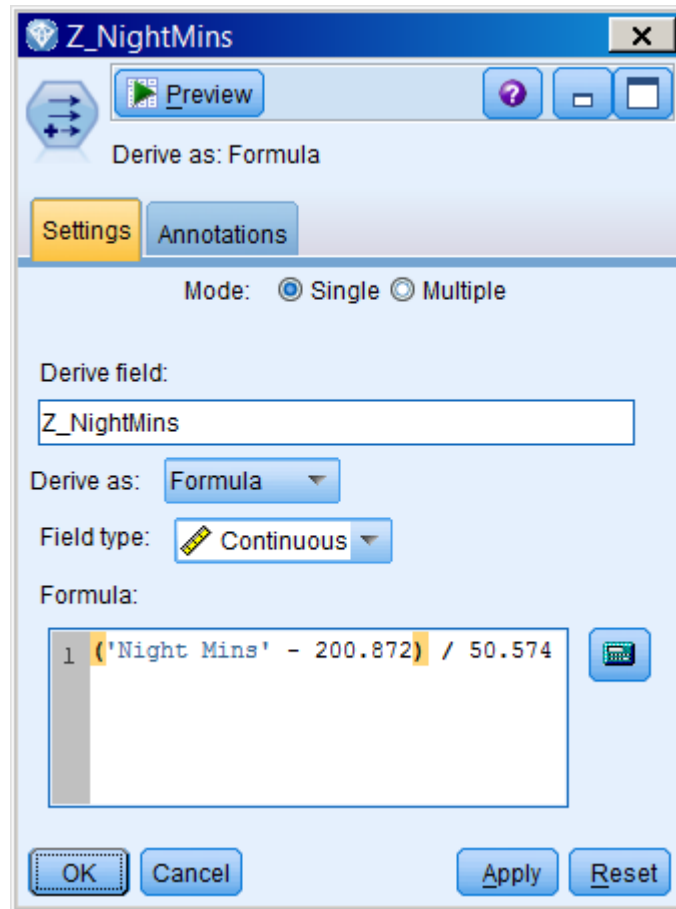
41. Transform the night minutes attribute using Z-score standardization. Using a graph, describe the range of the standardized values.

We treat night minutes as Continuous and calculate the mean **200.872** and SD **50.574** as depicted in the summary statistics below.



for download complete version of solution (all chapter 1 to 28 ) [click here.](#)

We then derive a new variable named **Z\_NightMins** specifying the Z-score calculation as depicted below.



We examine the range of Z\_NightMins both graphically using a histogram and quantitatively using summary statistics as depicted below. Observe in the histogram how the range of the data appears to expand into what is typically considered as the **outlier space**. We confirm this quantitatively noting that the minimum value is less than -3 and the maximum value is greater than 3.

