

CHAPTER – 3

DATA PREPARATION

The building block of any analytic exercise is a structured database. As we discussed, it is usually sourced from the operational systems of a business. For advanced analysis, this is just a starting point as we may require many other data outside the systems like competitor's information, media, macroeconomic data, credit bureau information etc. These databases are not ready for analysis due to many issues like missing information, illogical values etc. In this section, we will discuss the processes required to make data ready for analysis.

We will start with discussing the levels of measurement (scales of measurement), which is the use of number with various levels of quality attached to it. This is important as the process we are able to perform on a variable depends on the type of measurements.

3.1 LEVELS OF MEASUREMENT

Variables that form the database could be of different types like number, text, date etc. Even while the variable is numerical, we should not proceed with any analysis unless we are clear how it is measured. Many times, a number may stand as flag with no value attached or in otherwords, it is very much textual in nature. In terms of usage of number for measurement, there are 4 levels of usage with varying levels of weightage namely, Nominal, Ordinal, Interval and Ratio (this typology is proposed by Stevens, 1946). The type or sophistication of analysis depends on the level of measurement.

Nominal:- In this scale, usage of number is only as a flag as it doesn't use the value of a number at all. It is used for classification or categorization of entities. Examples include Gender (1=Female, 2=Male), Region (1=North, 2=South, 3=East, 4=West) etc. While using nominal scale variables, we should ensure the usage is consistent with this type of measurement.

Nominal scale is the lowest level of measurement as it is a categorization. We can count the value that is occurring with maximum frequency (mode) as a measure of central tendency. There are few other statistical techniques developed to handle this kind of scale. We will discuss this on a case basis while discussing statistical techniques.

Ordinal:- Here, the usage of number is to rank order entities which helps us to decide which one is higher, smaller, taller etc. Some of the examples are below.

Rank order the position in a hierarchy. Rating of service provided by clients like 1=high, 2=medium, 3=low. Here we know rating of low is the lowest compared to other two.

Ranking of customers based on their household income like:-

Rating	Income Range
1	< 5000
2	5001 – 15000
3	15001 – 30000
4	>30000

Note that the use of number is to provide relative position of the entity. Hence, in the example above, we can conclude that a customer with rating of 2 earns more than 1 and less than 3. We cannot use the value of rating number and conclude that the customer with rating of 2 earns twice as that of 1.

Compared to nominal scale, more statistical analyses can be conducted if the measurement is ordinal. The entities can be arranged in order and the middle value can be determined (median) as a measure of central tendency.

Likert Scale:- Another common usage of ordinal scale is for attitudinal measurements. Usual responses like strongly agree, agree etc. are given a numerical scale that helps rank order the responses. In social science, this type of scale is called ‘Likert Scale’ and it can be 5, 7 or 9 point scales. Example of a 5 point scale is given below.

Rating	Agreement to Policy Change
1	Strongly Agree
2	Agree
3	Neutral/ No Opinion
4	Disagree
5	Strongly Disagree

Although it is ordinal scale, Likert scale is an **exception** as calculation of mean is allowed. In the above example, let us say you have conducted a survey to understand the employees attitude towards agreement to a policy change. You can calculate the mean across all employees. If it is 4.85, it can be interpreted that in general employees are tending more towards strong disagreement.

Nominal and ordinal scales are associated with entities that can be observed but difficult to measure like colour, attitudes, types, positions etc. Hence, it is referred as *qualitative data*.

Interval:- Ordinal scale helped us to order or rank entities. Interval scale takes the measurement one level up by having the difference meaningful. Hence, in the interval scale, equal numerical difference represents equal quantities. Unlike in the case of ordinal

scale, the unit is well defined and standardized so that it doesn't vary by time, place or researcher.

All measurement of time are in interval level scale. Consider the dates of start (say 15/Jan/2001) and end (15/Mar/2001) of a project. The absolute measurements are not important and what matters is the days between the dates (interval). The same logic hold true for measurement of time too. In both cases, you may notice that the origin is arbitrary or historic phenomenon (it does not have a natural zero). Zero or origin doesn't mean absence of the phenomenon.

Another commonly used interval scale is temperature measurement. Here too the origin is arbitrary (freezing point of water if measured in °C) and what matters is the difference. A quick check of interval scale is to take a ratio and see if it is meaningful (eg. Year of start of an event is 1990 and year of close was 2000. Ratio between the years 1.005 doesn't make any sense, while the difference of 10 years is meaningful).

Most of the statistical tests and procedures can be applied on interval scaled data. However, analyst should be careful about simpler operations like addition, averaging, ratio etc.

Ratio:- Ratio scale is the highest level of usage of a number. In ratio scale, the interval (difference) is meaningful as well as the quantity (origin is meaningful). Since the quantity is meaningful, the ratio is meaningful and hence the name. The examples of ratio scale include physical measures like length, weight, age etc. as well as commercial measures like revenue, income, turnover etc.

If the measurement is on ratio scale, all statistical analysis can be conducted. Hence, we would like to have the measurements in ratio scale as much as possible. A measurement in ratio scale can be converted to ordinal scale if required. For example, income of customers can be grouped into High, Medium and Low. This type of categorization will result in loss of information but may be required for better presentation of data. The income may be presented as a bar chart of High, Medium and Low income groups.

The scales are critical as it will determine the type of summarization and statistical techniques that are allowed. In terms of summarization, nominal and ordinal scale will have to adopt proportions while interval and ratio scale can be summarized by taking averages. There is a whole spectrum of statistical tests related to the scales of dependent and independent variables. The table below shows the type of variables, test type and examples of hypotheses. Nominal and Ordinal scales are grouped under categorical as both are serving the purpose of categorization.

Variable1	Variable2	Test	Example of Hypothesis
Interval/ Ratio	Categorical (Two categories)	T-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$
Interval/ Ratio	Categorical (more than two categories)	ANOVA	$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 \dots$ $H_1: \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \dots$
Categorical (Two categories)	Categorical (Two categories)	T-test	$H_0: p_1 = p_2$ $H_1: p_1 \neq p_2$
Categorical (Two or more categories)	Categorical (more than two categories)	Chi-Square Test	$H_0: X \text{ and } Y \text{ are independent}$ $H_1: X \text{ and } Y \text{ are dependent}$

Table 3.1 Variable Types and Hypothesis Tests

The tables shows that if our interest is to test a hypothesis comparing the means of two entities, the test is t-test. Obviously in this case, the dependent variable will have to be interval or ratio scaled (the scales that allow calculation of means) and independent variable should be categorical (ordinal/nominal that divide the sample into two groups like male/female, high/low etc.). Similarly when both the variables are categorical in nature, the relationship can be tested using Chi-Square test. Examples include relationship between color preference and region.

Besides hypothesis testing, we are also interested in evaluating functional relationship between variables. The techniques we choose integrally depend on the type of scales.

Let us now discuss how to clean the data for making it ready for analysis. Most common issues faced are the existence of outliers and missing values. We will start with the discussion of treating outliers.

3.2 DETECTING AND TREATING OUTLIERS

Outlier detection and treatment is a critical step in the Analytics process. It is defined as an observation or a limited set of observation that do not conform the expected or normal behavior of the data. Barnett and Lewis (1994) defined it as *an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data*. Outliers can occur in data due to many reasons.

- **Human Error:-** This could be a case of data entry error if some section of business process rely on manual entry of data.
- **Environmental:-** The environment may consists of a small group of entities that can create data that looks like outliers. Hence, it is not an error.
- **Fraud:-** The unusual data pattern may be due intentional malicious activities. Hence, outlier detection is an important component of fraud detection process.

We need to be bothered about outliers due to many reasons. First of all it is an issue of data quality as it is likely to be an error. Hence, it makes sense to correct it before it is stored as a historical record. However, for analytics our interest is its impact on the results of analyses. If the outlier is present, statistical estimates can be biased and can be away from the unbiased estimate. In regression analysis, it can affect estimates as well as R^2 and hence, the decision based on the model. In decision tree model, the outliers will lead to increase in branches associated with the variable. In turn this will lead to reduced accuracy of the decision tree.

Impact of outliers can be lessened by the appropriate choice of statistic as some are more robust compared to others. For measures of central tendency, mean is less robust as it is moderately influenced by outliers. Median is the best measure as outliers got limited impact. As a measure of variation, standard deviation is severely influenced by outliers. Hence measures like inter-quartile range which avoids extreme values is the best measure. Hence, if you would like to evaluate central tendency and variation of a data before treating for outliers, mean and standard deviation are not the obvious choices.

Detecting Outliers

Manual inspection of scatter plots of each variable is a starting point of outlier detection. Initially variables are examined on its own and then it is examined along with other variables. Hence, it is broadly divided into univariate , bivariate and multivariate methods.

Univariate methods:- Each of the variable is examined and outliers at high as well as low levels are detected. A useful chart for this is bar chart. It is more meaningful if it is standardized to the corresponding z-values. The advantage of z-value is that it is unit independent and can be compared across all the variables. Usually if the absolute value of z is more than 2.5, it is a suspect and any value more than 4 is usually considered as an outlier with certainty.

The chart below shows the histogram of distribution of income of a group of customers under analysis.

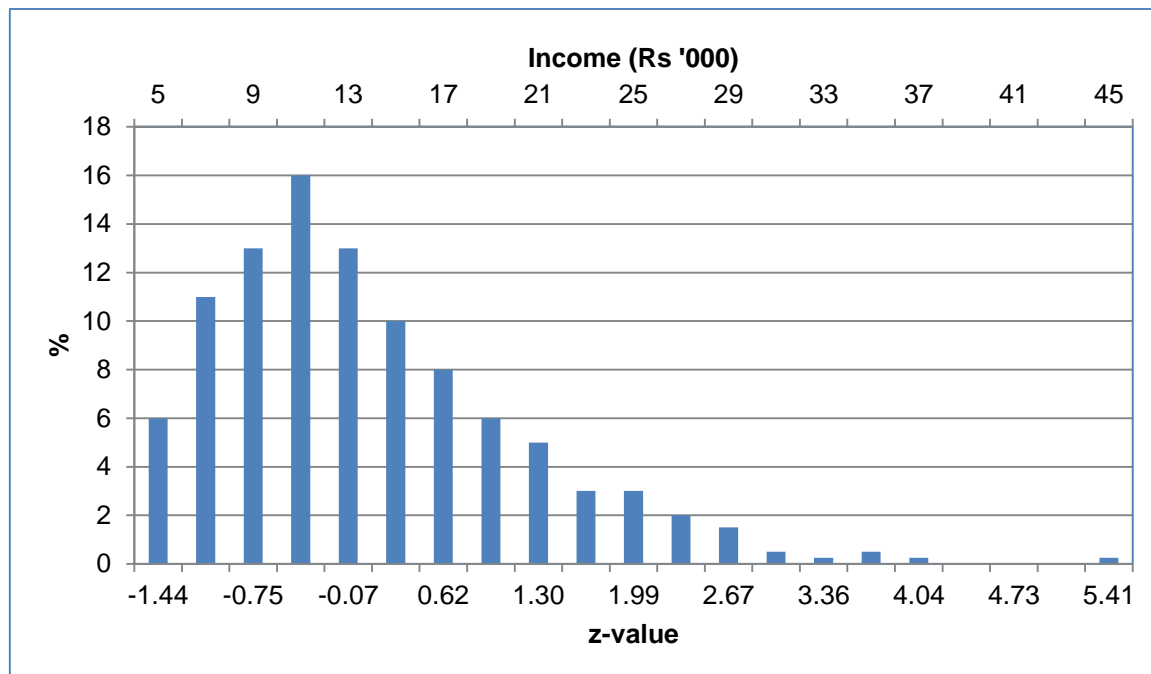


Figure 3.1 Histogram of Distribution of Income

The distribution shows that there are outliers as it too far from the average and also unlike other values. Another criteria to consider is z-value and it is 5.41, much above the typical cutoff value. We also consider the break of the outlier from the rest of the data. The chart shows that the curve tapers to an income level of 37000 gradually and then there is a distinct break with values and then the outlier appears. Hence, the outlier exhibits a clear break from the pattern exhibited by rest of the data. To summarize, we considered following aspects to detect outliers.

- Visual examination that indicated existence of outliers.
- Noted that the z-value is more than 4 that confirms it.
- Chart also indicated a break from the pattern.

Since Analytics deal with large number of variables, it is unlikely than an analyst will be able to spend enough time to individually examine all the variables. A practice of the industry is to give individual attention to important variables (based on prior analysis or industry knowledge) and use an automated algorithm for the rest. The automated process

Outlier but Important:- Analysis of transaction amount of users of a **retailer** credit card (credit card that can be used only with a particular retailer) customers showed few but very large values. The initial conclusion of analyst was that it is an outlier. On further examination it was uncovered that these were transactions made by contractors (small business owners), not an error at all. Still these customers were removed from analysis as the target of analysis was retail customers.

Hence, it is not wise to jump to any conclusion without careful examination.

may use a z-value cutoff to detect outliers. For large samples, a suggested cutoff is ± 4 and for small samples it is ± 2.5 (Stevens, 2009).

Bivariate/ Multivariate Methods:- In addition to univariate analysis, pairs of variables will be considered together to detect outliers in relation to other variables. It is usual to consider the dependent variable in relation to each of the independent variable for this. The chart below shows the relationship between salary and experience. The scatter plot, estimated line, and confidential estimate at $z = \pm 2.5$ is also shown.

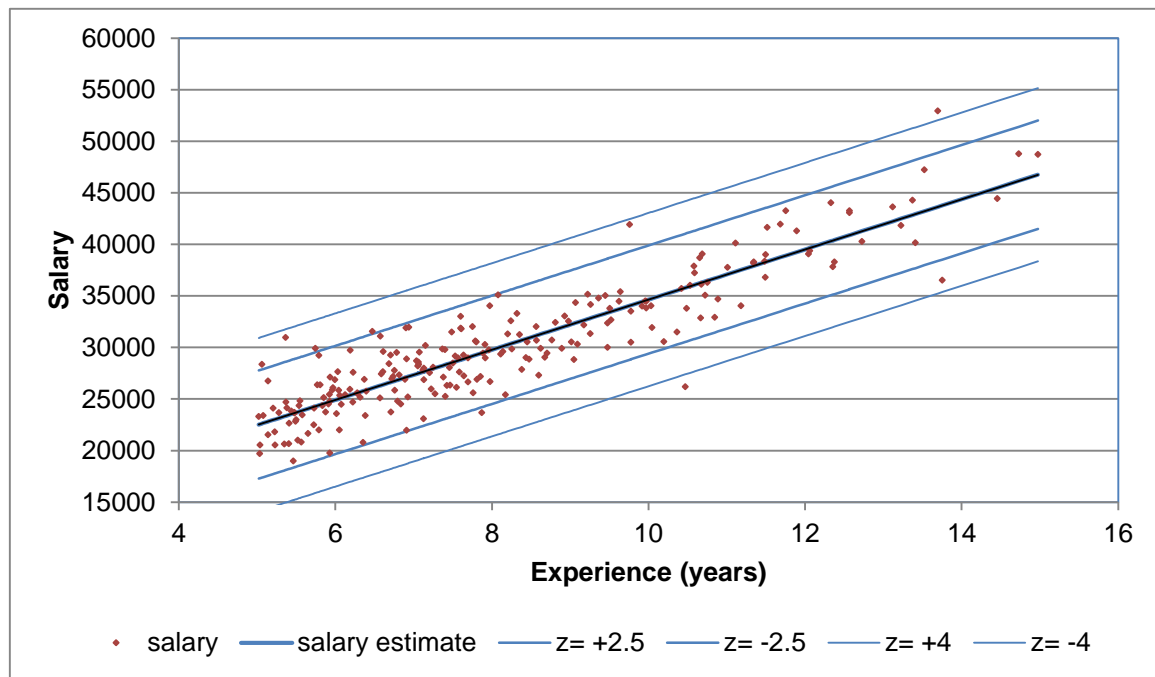


Figure 3.2 Histogram of Distribution of Income

To identify the outliers we have used the standard error of the estimate (salary) which is nothing but standard deviation of the error. The chart shows the boundary marked using a $z=2.5$ and 4 above and below the estimated line. Using z table we can observe that $z=\pm 2.5$ will bind 98.76% of the observations and $z=\pm 4$ will bind 99.996% of the observations. Hence, the observations outside this bound is a possible outlier. Examining in relation to experience, there are precisely two observations that can be outliers for sure. Some of these are at low level of salary which would not have been considered as outlier through univariate analysis.

Modeling involve more than one independent variables in almost all cases. Hence, it makes sense to conduct this analysis as a multivariate scenario using all the variables that will be part of the model. This way, we need not evaluate numerous pairs. However, here we will not be able to plot the scatter to decide. Instead, there are robust measures that

can be used to identify outliers. Cook's distance (Cook, R.D.,1977) is one of such measures that can be used and it is available as part regression analysis output in most statistical packages. It provides an overall measure of the impact of an observation on the estimated regression coefficient. Usual cutoff for deciding the outlier is $4/N$ where N is the number of observations.

Obvious query for an analyst is which method to follow given a problem. Typical practice in the industry is to use univariate detection due to many reasons. Many a times the analytic problem may not require application of multiple linear regression. Hence, there is no opportunity to use multiple regression and identify the outlying observations.

Treatment Of Outliers

Once the outliers are identified, next step is to consider what to do with the outliers. In situations where the analytic process involves building a multiple linear regression, it make sense to identify outliers using multivariate approach and remove the observations. Since analytics deal with large number of observations, this is a safe approach. However, usual approach is to identify outliers through univariate approach. If we decide to eliminate observations, we stand the risk of loosing a significant number of observations as number of variables are large in a typical analytic project. Hence, a method adopted is to force the outlying values to the maximum in the data (capping).

In the chart below, there are two outliers at lower and upper level. Both the values are forced to the next level in the data. This will ensure that the observations are not lost while treating outliers.

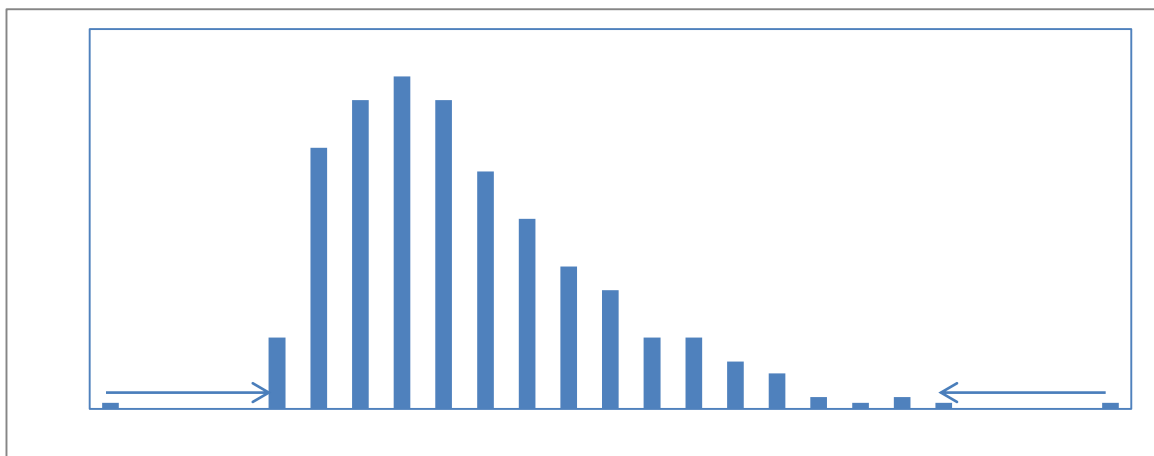


Figure 3.3 Outlier Treatment

There exists a number of other methods of treating outliers without removing. Replacing it with mean is one such approach. However, this will tend to reduce the spread of the

Caesars Entertainment, US

Caesars Entertainment is one of the world's largest diversified casino entertainment companies and operates hundreds of outlets on nearly 40 properties in 20 cities in North America, primarily under the Harrah's, Caesars and Horseshoe brands. It is well known for its industry leading customer relationship management program. The program evolved over the years and currently known as Total Rewards; which won Master of Enterprise Loyalty award in 2013. The program deliver a personalized experience for Caesars guests and resulted in 20% growth in members in 2012¹.

The design and operation of this program is data centric. The company creates detailed profile of each of the guest based on the transactional and demographic data it collects. Employees then use this information throughout the company to make better operational decisions. For example, the staff can use this information to personalize nearly every aspect of a guest's stay, from how she is greeted on arrival, to how her room is made up. The company's marketers can create and target special offers with precision. It is common in this industry to lavish attention on Big Spenders. However, at Caesars technology allow them to personalize service to larger group of customers.

The company was able translate the learnings from this program into mobile platform too. The mobile interactions start with an option to check-in through SMS bypassing the registration line. It is also location aware and the company can offer appropriate offers. For example, if the customer is at Paris, a free admission to Eiffel tower may be given. The mobile App named myTR put more of its services at guests' fingertips during their stay. This allows Total Rewards members to keep track of special offers, manage reward points and even book rooms. At some locations, it provide access to a mobile concierge, real-time event information, in-room dining, and even wake-up calls. In this industry, most decisions by guests are taken while they are on the floor and that is when the company should reach out to them. Mobile phone with the guest is a unique opportunity to reach out to them².

*1 <https://colloquy.com/colloquy-recognizes/caesars-entertainment-2013/>

*2 [http://www.capgemini.com/resource-file-access/resource/pdf/The Digital Advantage How Digital Leaders Outperform their Peers in Every Industry.pdf](http://www.capgemini.com/resource-file-access/resource/pdf/The_Digital_Advantage_How_Digital_Leaders_Outperform_their_Peers_in_Every_Industry.pdf)

population. A more complex approach is multiple imputation that involves replacing

outliers with possible values (Elliott & Stettler, 2007).

Outlier identification and treatment is quite an evolved topic in statistics. For the sake of brevity and practicality, we covered only topics that are important from the industry perspective. Some of the areas we didn't cover are:-

- Accurate identification of z-score while the distribution is not exactly normal. This would involve treating the data to make it normal and then applying the cutoff.
- There are quite a few complex methods for treating outliers. These techniques are broadly under multiple imputation and the objective is to replace outliers with most likely values.
- There is a large body of work on identifying outliers for categorical variables. Identifying outliers for such variables are not usually practiced as such variable may not carry any useful information (phone numbers etc.).

3.3 IDENTIFYING AND TREATING MISSING VALUES

Missing values refers to a situation where valid values are missing in the database. It is so common that it is unlikely that any analysis can be done without facing this issue. The objective of missing value treatment is to address this issue such that analysis result is unbiased and as accurate as possible.

Missing data can occur due to a number of reasons:-

- It could be because the subject of a longitudinal study might have dropped out. Examples include data collections from stores on point of sale data. The store management may decide to stop providing the data. In studies considering consumers as unit of analysis, there will be missing values if they stopped subscription to the service.
- In survey research, missing data can occur due to non-response. Respondents may not answer personal questions like income, age or education etc.
- There could be missing values due to operational issues. These are occurrences at random due reasons unknown or difficult to uncover.

The impact of missing values depends on the level of missings and if it is distributed randomly. If the missing values are these for only 5% or less observations and if it is distributed randomly, it can be considered as trivial. In such situations, it is a safer option to remove those observations without any serious impact on results.

For multivariate analyses that consider large number of variables, dropping observations may not be the best approach. Here, even a moderate level of missing in each of the variables will result in serious reduction in sample size. Another issue is the impact of non-randomness of the missing values. If the missing values are distributed non-

randomly, ie it depend on dependent variable or any of the independent variable, it will introduce biasness to the result. Hence, a number of methods are developed to treat the issue without dropping observations.

Methods Of Handling Missing Values

Replace with zero:- In many instances, missings are provided following the data policy to indicate an information. A closer examination will reveal if that is the case. The table below shows typical coding scheme used by financial institutions for a variable ‘outstanding loan’.

- If the customer has not availed a loan so far, it will carry a missing value (indicating ‘not applicable’).
- Otherwise it will carry the actual amount outstanding and it could be zero too if the loan is paid off.

This column cannot be used for analysis straight away because of missing values. Treatment should depend on the objective of the analysis. If the requirement is to have a variable that will indicate the level of debt for customers, then it makes sense to convert the missing into zero. If the requirement is to have a variable that indicate a customer has ever availed a loan, then the variable can be converted into binary (yes/no).

Customer ID	Loan Amount		Customer ID	Ever availed a loan	Outstanding Amount
221001	15642.00		221001	Yes	15642.00
221002	26445.00		221002	Yes	26445.00
221003	.		221003	No	0.00
221004	.		221004	No	0.00
221005	0.00		221005	Yes	0.00
221006	32000.00		221006	Yes	32000.00

Table 3.3 Missing Value Imputation

In many instances, missing is an indicator of being ‘not applicable’ and care should be taken before converting the missing into zero.

- Performance rating of an employee joined recently in the HR information system is coded as missing (conversion to zero would mean the appraisal was conducted and it was zero).

- Satisfaction rating of customers who made calls to a service center is coded as missing if a customer never made a call (conversion to zero would mean the calls were made and rating was zero).
- It is also a standard practice to fill missing values using flags like '99', '999' etc. Hence, if there are customers of age 999 in the dataset you are analysing, it only means the age was not available!!! Care should be taken to ensure if flags exist in the dataset you are analyzing.

There are many variables like income, age etc. for which replacing with zero is not an option all. In such cases, we require a method to treat the issue.

Delete Observations: Deleting observations is often the appropriate approach if the missings are at random. If the data do not meet the assumption of randomness, this may yield biased estimates. Depending on the biasness, it may exaggerate some effects or underestimate others. It may even reverse the direction of effects. However, this is not a good strategy when there is limited sample as it will lead to a loss of statistical power. Moreover, for analytics problems dealing with large number of variables, this can lead to sizeable loss of sample size.

In summary, even while the missing values are distributed at random and there is sufficiently large sample, listwise deletion is not an appropriate strategy. But note that this is the default strategy for all statistical software as it will not consider the observations that contain any missing value in the variable.

Mean Substitution:- Substituting with mean is the most commonly used approach. It is based on the fact that values around mean have got highest probability in random sample. Here too, if the missings are not at random, it is not a good approach. Another issue to consider is the proportion of missings. If it is large (say more than 10%), it may have undesirable impact on the analysis outcome. Since, a large proportion of values are going to be a constant, it will reduce variation. To illustrate the impact of mean substitution, an experiment is presented below (similar to Acock, A.C., 2013). In order to understand the impact we have evaluated the distribution of a variable (income) as it appears in a customer database. In the next step we have created missing values at random for 20% of observations and evaluated the distribution. In the final step, we have replaced the missing values with the mean and then evaluated the distribution. Charts were created to understand the impact. The chart below shows the distribution with no missing values.

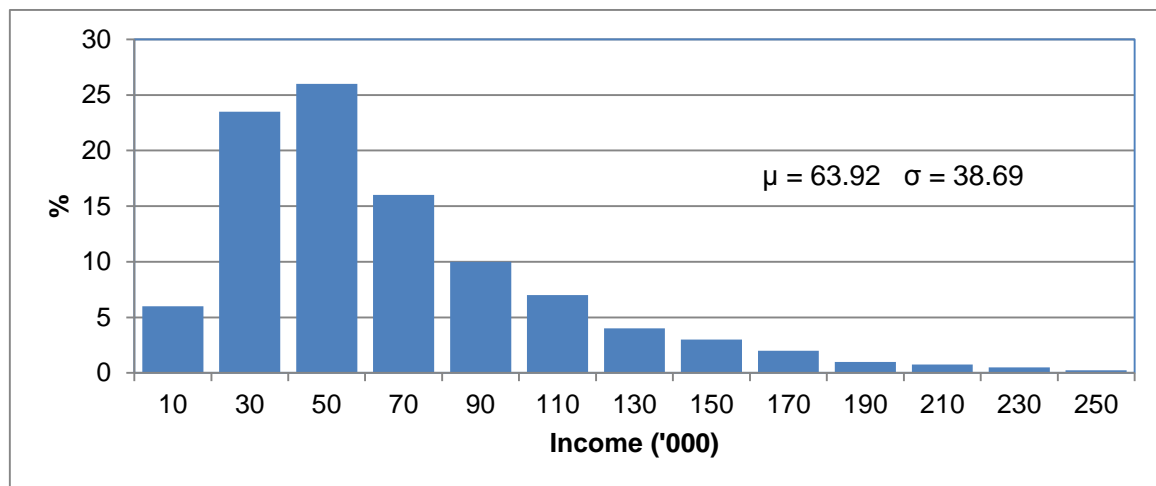


Figure 3.4 Distribution with no Missing Values

The chart shows that original distribution is right skewed with mean of 63.92 with a standard deviation of 38.69. The chart below shows the distribution after forcing 20% observations missing at random.

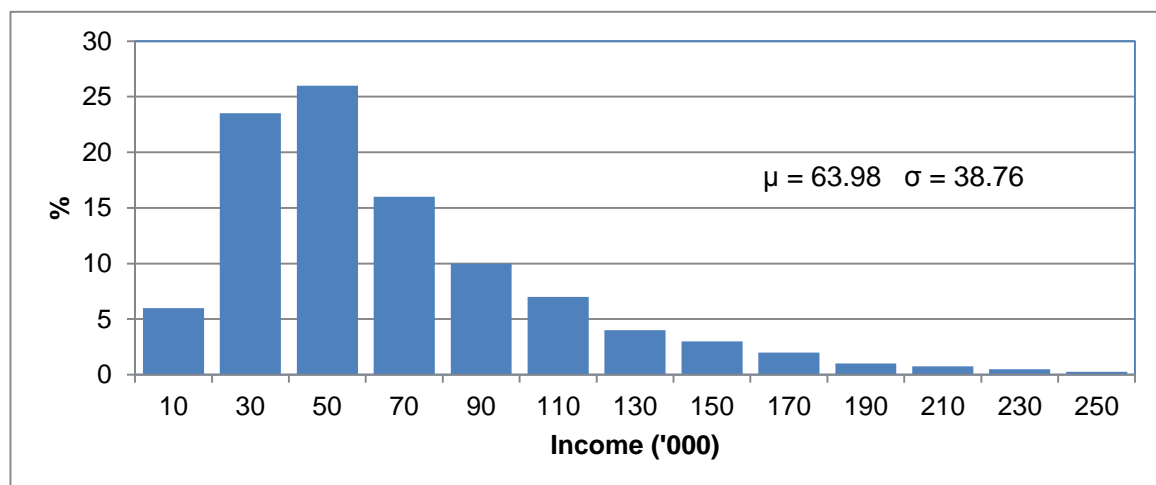


Figure 3.5 Distribution with 20% missing values

As we can observe, this didn't make any significant change to the nature of distribution or the parameters. This emphasized the contention that missing values if at random is not harmful.

The chart below shows the impact of replacing all missing values with mean.

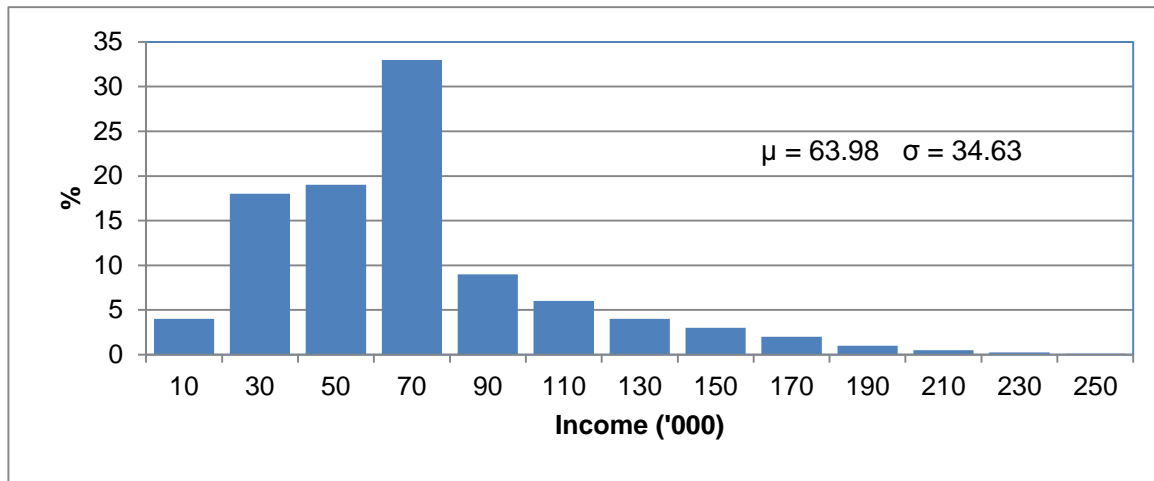


Figure 3.6 Distribution with missing values replaced by Mean

The chart shows that replacing missing values with mean made a big difference to the nature of distribution and also reduced the variation.

An improvement over this approach is to use mean evaluated at subgroups. Instead of replacing the missing with a single value of mean, we can replace it with mean evaluated at subgroup levels based on an important, relevant variable. In the previous example, we will evaluate means for various professions of the customer. Then the missing will be replaced by the means according to the profession. This will preserve the distribution better than previous approach.

Regression imputation:- In this method, a regression model is estimated to predict observed values of the variable under consideration using other variables in the database. This model is then used to impute values in cases where that variable is missing.

For example; if we are treating the missing values in income, then we will build a model to predict income based on other variables in the database like occupation, age etc. Then this model will be used to impute income wherever it is missing.

Here, available information for non-missing observations are used to predict the value of the variable in cases where it is missing. A problem is that the imputed value does not have any variation around the fitted line (no residual variance). This causes estimates to have lower variance and falsely indicate greater precision. Hence, the regression model predicts the most likely value of missing data but does not supply uncertainty about that value.

Multiple Imputation:- Substituting with mean is an example of single imputation. It is appropriate if the proportion of missing values is small. As we discussed, a drawback of this approach is drop in variation resulting in an overestimate of precision. Hence, multiple imputation is being considered as a better option. It involves creating multiple

datasets, creating parameter estimates and then pooling it to get the final estimates. It allows analyst to incorporate the uncertainty in imputation through standard errors of the parameter estimates. It follows the steps given below.

The first step involved creating 3-10 datasets using simulation methods. These methods impose a probability model on the complete data (observed and missing values) and impute the missing values with multiple values. One of the most commonly used method is Markov chain Monte Carlo (MCMC). However, depending on the package, there are many options like regression, propensity score method etc.

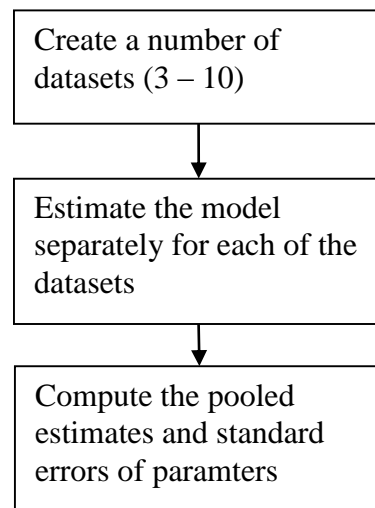


Figure 3.7 Steps of Multiple Imputation

It is natural to suspect if 3-10 imputation are enough for this exercise. Rubin(1987) has evaluated the efficiencies and has shown that there is little advantage in analyzing more than few imputation unless the rate of missing information is very high. However these decisions depend on the original data structure.

Second step involve estimation of the model using each of the datasets created. The estimation method may be any techniques (linear regression, logistic regression etc.) as warranted by the problem. There are some guidelines regarding the choice of techniques to be applied under Multiple Imputation. In general the techniques which are not too sensitive to normality of data (the techniques which depend on means and variances) can be appropriate for data created with simple imputation methods. While analyses that is sensitive to normality would require complex imputation methods.

Third step would require calculating the pooled estimate and the standard error. If the standard error is low, this is an evidence supporting the imputation. As we can see, the process is quite complex and purely manual approach may be not practical. There are

Analytics @ Work

United Parcel Service, US

United Parcel Services had humble beginning as messenger company in 1907 at Seattle, US. Today UPS has become world's largest package delivery company and a leading provider of specialized transportation and logistics services with presence in 200 countries. Analytics, especially Operations Research based solutions played a critical role in achieving this growth. As early as in 1954, the CEO, George Smith stated that "Without operational research, we would be analyzing our problems intuitively only, and we would miss many opportunities to get maximum efficiency out of our operation"¹. Over the period of time UPS operations have become very complex covering the globe in terms geography and all conceivable transport modes covering air, road and water. For example, in 2009, UPS had 263 aircrafts in their inventory making it the ninth largest airline in the world.

UPS continue to use Operations Research in all its main areas which are broadly divided into Package delivery/pickup; Hub (sorting packages); Feeder (over-the-road package transport, hub to hub or hub to delivery center); Airline (package transport via our air network). In all these areas it got custom and even award winning solutions based on OR.

An interesting application is ORION which stands for On-Road Integrated Optimization and Navigation, a data-intensive system that lays out the most efficient routes for individual drivers to deliver their loads via a series of complex algorithms. This is based on data generated from sensors attached to its delivery trucks. By analyzing this voluminous data, the company was able to make significant improvement in the operations by reducing idle time and cutting distance travelled. Moreover, the algorithm could predict the probability of failure of each truck so that preventive maintenance could be conducted².

Analytics will continue to drive the growth story of UPS. As the company's network and business become more globalized and complex, it will rely on Analytics tools and techniques to maintain an efficient, reliable and cost effective service.

^{*1} <http://www.analytics-magazine.org/march-april-2010/154-corporate-profile-analytics-at-ups.html>

^{*2} www.computerworld.com/article/2483847/enterprise-applications/data-analytics--eye-popping-results-from-intel--ups-and-express-scripts.html

many packages that incorporate components that automate some of these steps (MI and MI Analyze of SAS; ICE, MVIS, and MICOMBINE of Stata).

In real world analytics, MI is not very common. Missing value treatment forms only a small part of overall process and hence may not have the flexibility to the level of effort required for MI. It could also happen that the change in process required for MI may not be acceptable to the process owners. Hence, usual practice is to adopt of any of the single imputation techniques.

3.4 DATA TRANSFORMATION

Business process data is usually messy from the point of view of statistical analysis. The messiness is because the operational systems capture huge amount of data but may not carry much useful information. For example, any call made on a mobile phone will generate detailed information about the time, location, phone number, duration etc. at the operational system. All these may not be useful at such a minute level. These variables are called raw variables; not exactly analytical variables (Refaat, M.,2007). For most analytics initiatives, it is enough to get a summarized data at monthly level going back for say two years.

Given the size of this data, it is not practical (consume too much resource) to summarize these variables for each of the projects. Hence, businesses in general summarize the operational datasets at weekly or monthly level and make it available for all analytic initiatives. This data base is usually called Analytic database or analytics data server (ADS). It is specifically designed to support Analytics and Business Intelligence applications. This differentiates it from an operational, transactional or OLTP database, which is used for transaction processing – i.e., order entry and other “run the business” applications. Databases that do transaction processing can also be used to support Analytic and BI applications, but the advantage of analytic database is that it contains summarized information as usually required. Of course some study still would require analyst to reach operational database and fetch information that is not available on Analytic database but such cases are rare.

Analytics in Practice

Single Version of Truth

Single version of truth (or one view of truth) is a term you will hear very often in a professional setting. This refers to the availability of data at one place for Analytics. It is achieved through either having a single database (impractical) or through a distributed synchronized database, which stores all of an organization’s data in a consistent and non-redundant form. It is an issue when undertaking an Analytics initiatives. Many organizations keep information in siloes and bringing these together could be a herculean task. This makes it difficult to deliver consistent information across the organization. Organizations are now starting to consolidate these siloes into Enterprise Data Warehouse (EDW). However, organizations in general have a long way to go in achieving the right level of consolidation. It is a long process with considerable investment.

Even while AD exists, data would require further treatment and summarization to meet the requirements of statistical techniques. Some of the methods of summarization is discussed below.

Time Rooting

The databases store information by date and time stamp. But for most studies, the date or time is not very important but the time duration. The information like duration of an activity leading to an event or vintage of an entity before an action etc. are what is being studied rather than the date/time of the event or activity.

For example, consider the case of a large organization studying the factors influencing resignation of employees. The study is being conducted as of Mar/2012. It would involve evaluating the performance, promotion records, salary, and department etc. of these employees. As in the case of any analytic project, we will set some criteria for including an employee into the analysis.

- An employee (resigned and current) should be on the rolls for a period of 12 months at the least. This requirement ensures that we have a minimum amount of information for analysis.
- Resignation of resigned employees should have happened in the last 12 months. Such a condition is required as the drivers of resignation earlier than this may be different compared to the recent past. Hence, to ensure that the result is relevant for decision making, such a condition is required.

In the table below ‘C’ indicate the month the employee is on the roll and ‘R’ indicate the month of resignation. Status column indicate if the employee is on the rolls (Status=0) or resigned (Status=1) or the employee is not being considered for analysis (Status=D).

The table shows that:-

- Employee-1 joined service some time earlier than Mar/2010 and is continuing in service.
- Employee-2 joined service in April/2010 and resigned during July/2011. We do have 15months history available for this employee and his status is 1.
- Employee-3 joined the service prior to Mar/2010 and resigned during Nov/2010. We are not able to consider this employee for analysis as the resignation happened before the 12-month window (similar to employee-5).
- Employee-8 resigned during the resignation window but doesn’t have 12 months history. Hence, we are not able to consider this employee for analysis.

Emp	Mar-10	Apr-10	May-10	Jun-10	Jul-10	Aug-10	Sep-10	Oct-10	Nov-10	Dec-10	Jan-11	Feb-11	Mar-11	Apr-11	May-11	Jun-11	Jul-11	Aug-11	Sep-11	Oct-11	Nov-11	Dec-11	Jan-12	Feb-12	Status
1	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	0
2		C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	R								1
3	C	C	C	C	C	C	C	C	R																D
4	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	R		1
5	C	C	C	C	C	R																			D
6	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	0
7	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	0
8														C	C	C	C	C	C	C	R				D
9	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	0
10												C	C	C	C	C	C	C	C	C	C	C	C	R	1

Table 3.5 Time Series Data

For analysis, date joining and resignation is not important. What is important is the amount of time lapsed and developments in that period. Hence, the data is time rooted based on the interval as shown below. The dependent variable indicates if the employee is in service or resigned. Months indicate the months before the event. Month-1 is the month before resignation irrespective of the actual date when it happened. For employees in service, it is the latest month in consideration. The employees that doesn't meet our criteria is removed from analysis. The table shows that:-

- For Employee-1, who is in service, Month-1 is the current month of study Feb/12 and Month-2 is Jan/12.
- For Employee-2, Month-1 is the last month of service Jul/2011 and Month-2 is the Jun/2011.

Emp	Month-12	Month-11	Month-10	Month-9	Month-8	Month-7	Month-6	Month-5	Month-4	Month-3	Month-2	Month-1	DV
1	C	C	C	C	C	C	C	C	C	C	C	C	0
2	C	C	C	C	C	C	C	C	C	C	C	C	1
4	C	C	C	C	C	C	C	C	C	C	C	C	1
6	C	C	C	C	C	C	C	C	C	C	C	C	0
7	C	C	C	C	C	C	C	C	C	C	C	C	0
9	C	C	C	C	C	C	C	C	C	C	C	C	0
10	C	C	C	C	C	C	C	C	C	C	C	C	1

Table 3.6 Time Series Data-Time rooted

Hence, out of 10 employees we started with, 7 met our criteria and the data is time rooted so that chronology is removed. For analysis, we will collect data about these employees for each of the months (Month-1 to Month-12) like salary, appraisal data, increment, promotion etc.

Summarization Of Time Series Data

As discussed earlier, businesses store operational data at the level of activity and it is too granular. Even when it is summarized at monthly level, it is too voluminous. Consider the study using 2 years data for say 10 variables. This means we will have 240 variables to handle. Hence, this data is summarized at regular time interval for analysis.

For example, consider a portion of data of a typical credit card analytic database provided below.

pymt1 2	pymt1 1	pymt1 0	pymt 9	pymt 8	pymt 7	pymt 6	pymt 5	pymt 4	pymt 3	pymt 2	pymt 1

Table 3.7 Summarization of Time rooted data

Usually it will contain variables like purchase, payment, balance, interest etc for a period of 48 months. As we discussed, it is too disparate for analytic applications. Hence, usual method is to sum up these values at 3months **overlapping** time periods (last 3months/ 6months/ 9months etc.). When these values are linked to the event (response to a campaign/ switching of service etc.), it is likely to have an impact rather than individual monthly values.

Another method is to sum this up distinctly at regular intervals (usually at 3monthly or 6monthly levels). These sums are directly comparable as it is summed at same intervals.

It can be used to calculate the change over the last few quarters,

pymt1 2	pymt1 1	pymt1 0	pymt 9	pymt 8	pymt 7	pymt 6	pymt 5	pymt 4	pymt 3	pymt 2	pymt 1
Sum of payment for 10-12 months (SumQ4)			Sum of payment for 7- 9 months (SumQ3)			Sum of payment for 4-6 months (SumQ2)			Sum of payment for last 3 months (SumQ1)		

Table 3.8 Summarization of Time rooted data

$$\text{GrowthQ1_Q2} = (\text{SumQ1} - \text{SumQ2}) / \text{SumQ2}$$

$$\text{GrowthQ1_Q3} = (\text{SumQ1} - \text{SumQ3}) / \text{SumQ3}$$

$$\text{GrowthQ1_Q4} = (\text{SumQ1} - \text{SumQ4}) / \text{SumQ4}$$

These growth rates are quite useful for Analytics as it can usually warn about customers leaving the service and thus could turn out be a good predictor variable.

The discussion above was illustrated using the sum of the variables but it needs to be modified as mean for variables for which sum is not relevant like price. Depending on the data and the problem a number of meaningful variables can be generated.

- Total or Average purchase value over a certain period (month, quarter etc.)
- Total or Average amount of outstanding over a certain period (month, quarter etc.)
- Total or Average number of transactions over a certain period (month, quarter etc.)

Hence, depending on the requirement, the derived value could be based on Sum, Average, Min, Max, Standard Deviation, Count etc.

The predictive ability of these variables can be further enhanced by taking ratio of the variables. Some examples of the ratios are provided below.

- Ratio of payment in time period to outstanding in certain period.
- Ratio of total purchases in a time period to available credit in the same period
- Ratio of Advertisement spend in a time period to Total for the Year
- Ratio of Total debt purchase in time period to Credit available



Analytics to Determine Character!!

Lending decisions are usually taken based on financial information like credit history, credit scores etc. This immediately works against recent graduates and students who lack credit card; mortgages or car payments; the things that normally earn good or bad credit scores. The innovative startup 'Upstart (<https://www.upstart.com/>)' looks at other information like SAT score, colleges they attended, majors and grade-point averages (GPA). Using this the company is not just trying to evaluate the job prospects, but the personality as well.

According to Paul Gu, the co-founder, "If you take two people with the same job and circumstances, like whether they have kids, five years later the one who had the higher G.P.A. is more likely to pay a debt. It's not whether you can pay. It's a question of how important you see your obligation."^{*1}

The idea is that these indicators stand for certain behavioral characteristics that give preference to honoring their debts (as evidenced by data) even if they are in financial difficulty. In other words, they give utmost importance to fulfilling their obligations.

In the financial terms, Upstart can be considered as a peer-to-peer lending platform. It was started by former Google employees (Dave Girouard, Paul Gu and Anna Mongayt). The motivation behind the venture was to help the youngsters at the early stages of their career so that they can choose an entrepreneurial track rather than a corporate career. Hence, an investment in people.

^{*1} http://bits.blogs.nytimes.com/2015/07/26/using-algorithms-to-determine-character/?_r=0

Standardization and Normalization

The data of a typical analytics project consists of a large number of variables with widely different nature. The levels of measurement and the range of values could be extremely different. We don't want such differences to influence the outcome of the results as some techniques are sensitive to it.

Examples include techniques that use distance measure (Euclidean distance) like cluster analysis. A variable like credit limit might have a range of 10000 to 500000 while age might have a range of 18 to 100. In a distance measure using these two variables, contribution of age to the distance will get swamped by credit limit. Hence, it is important to treat the variable so that the variability determines the importance and not the range. Following are some of the standardization techniques.

Normalization:- Normalization will scale all values within [0,1] with at least one expected value at each of the end points. The formula used is given below.

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (3.1)$$

Here, x_{norm} is the normalized value of x . x_{min} and x_{max} are the minimum and maximum value of x in the database.

A disadvantage of this approach is that it depends on extreme values and those may be outliers. Since outliers are quite common, analyst should be careful about applying it.

Standardization:- Standardization uses a transformation of the data to zero mean and unity variance.

Tips on Technique

Disaggregation of Data

Although not as common, you should be prepared to disaggregate data too. Disaggregation involves converting the data into a lower form. This is required for bringing the data into same level as the master dataset. Unlike aggregation, disaggregation will require external information to complete the task. Consider few examples below.

- In a Marketing Analytics study the master data (sales data) is at weekly level. The advertisement data could be at monthly level. In order to merge this data with sales data, either we should summarise sales data into monthly level or convert monthly advertisement data into weekly data. Our will not prefer to summarize the sales data as summarization would result in loss of information (also drastic reduction in sample size). The advertisement data can be disaggregated into weekly by assuming equal amount by week. If you have any basis to think that it is not equal across week, that information may be used to split the data. For example if you have weekly advertisements at macro level, it may be used as an index to divide.
- Similar situation may happen at geographical level too. You may have data at higher level like country, and the requirement may be to disaggregate the data into state level. Here it doesn't make sense to convert equally between the states given the large size difference. Hence, you will have to use some information as an Index for this. It could be population of the state, sales of relevant product etc.

$$x_{std} = \frac{x - \mu}{\sigma} \quad (3.2)$$

Here x_{std} is the standardized value of x . μ is the mean and σ is the standard deviation. The main disadvantage of standardization is that it is not bounded like normalization. The standardized value is nothing but z-score and it indicates how many standard deviation away a value lies. Hence, it is also called *z-score transformation*. A z-score of 1 means the value is $(\mu + 1*\sigma)$ and z-score of -1.5 is $(\mu - 1.5*\sigma)$.

Since clustering is sensitive to units, most of the statistical packages do standardization before clustering while others expect standardized input. It is critical to evaluate the requirement of a package before the analysis. This will help you to standardize the input before submission if the package expects it. Also avoid standardizing twice if the standardization is inbuilt.

The table shows values of a variable and the standardized and normalized values. As expected, the normal values are bounded between 0 and 1.

Variable ($\mu=49.67, \sigma=15.45$)	Normalized value	Standardized value
36	0.25	-0.88
61	0.73	0.73
23	0.00	-1.73
63	0.77	0.86
34	0.21	-1.01
32	0.17	-1.14
67	0.85	1.12
57	0.65	0.47
59	0.69	0.60
63	0.77	0.86
53	0.58	0.22
53	0.58	0.22
74	1.00	1.64
27	0.08	-1.47
66	0.83	1.06

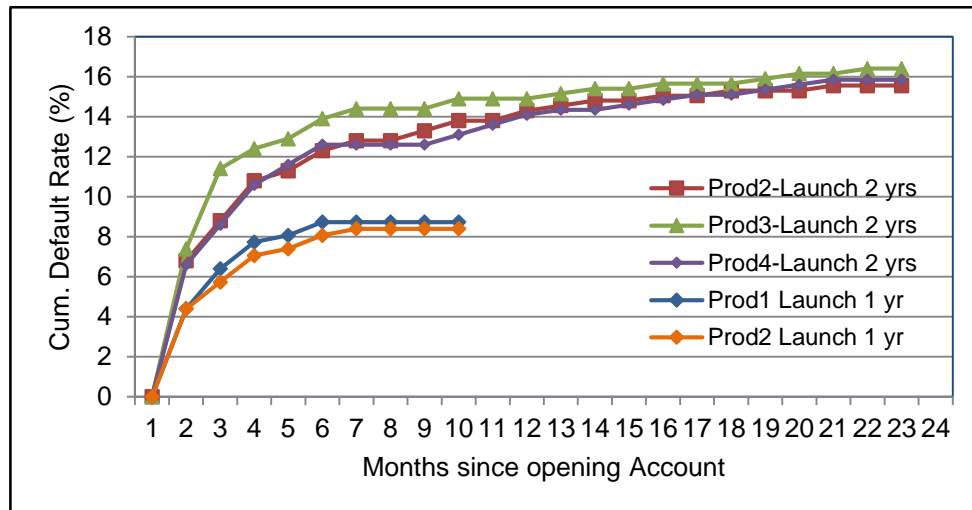
Table 3.10 Normalized and Standardized Values

Tips on Technique

Vintage Analysis

Vintage analysis is a simple but powerful technique for evaluating segments of customers based on the origin of commencement. Here, the entities are grouped based on the vintages and any critical behavior is assessed. This is quite useful where information is collected on time series.

The first step in vintage analysis is to root the time to month of enrolment as explained in the chapter. In the example below we are considering two groups of products launched one year and two years back. In the case of products launched two years back, there were consumers with maximum vintage of 2 years. The default level of these customers were evaluated and summarized at product level. Similarly, this was repeated for products launched one year back.



The chart shows that cumulative default rate increases at faster rate till about 6 months since opening the account. After this, the default slows down. Both groups products exhibit similar behavior, but default rate products launched 2 years back is much higher than products launched 1 year back.

The chart below shows comparison between normalized and standardized values.

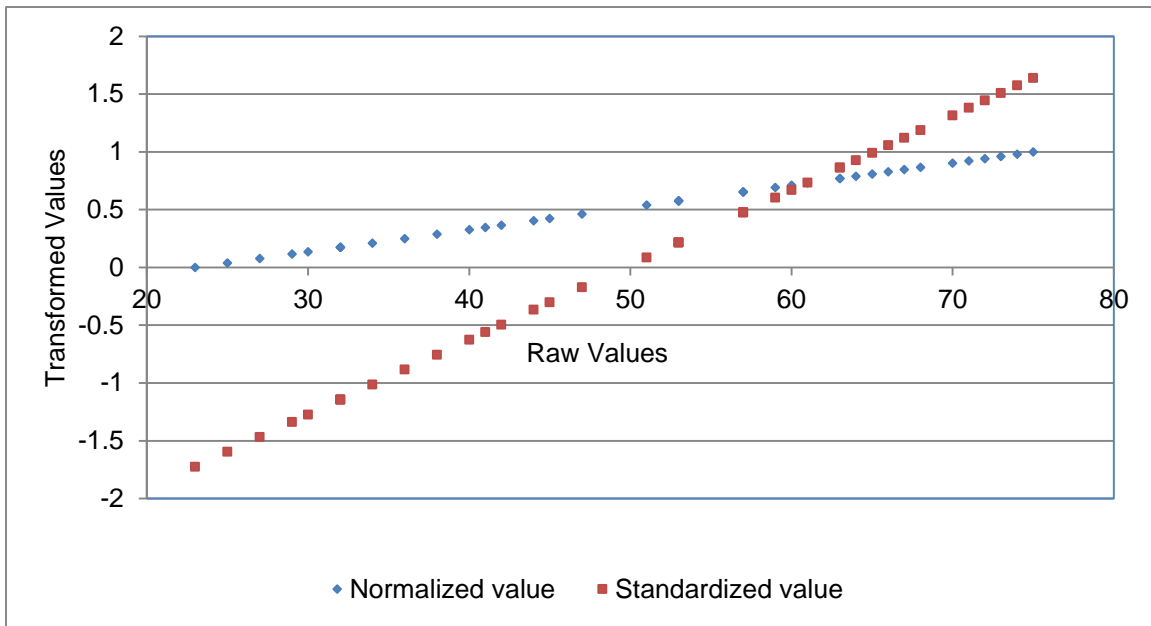


Figure 3.8 Normalization and Standardization

The chart shows that normalized values are between 0 and 1 while standardized values range from -1.73 to 1.12.

Transformation of Data

Many a times the variables will not be in the form that is right for analysis. There could be problems with extreme values, clustering around certain values, non-linearity, asymmetry etc. These issues make visual representation of the data or relationship with another variable not effective. Another issue is that in this form, it may be violating certain assumptions like linearity required for the statistical techniques. The objective of data transformation is to apply a mathematical function to the data to modify the distribution.

Common transformations include calculating reciprocals, logarithms, and raising variables to positive integral powers and taking roots (square root, cube root, etc.). These are applied depending on the objective of the transformation. The examples below will illustrate some of these transformations.

Log Transformation To Improve Visual Presentation:- Log transformation is useful to make highly skewed distribution less skewed.

Consider the plot of two variables below. Since it is highly skewed towards lower values, the relationship is not clear. Taking log on both values will reduce the skewness and it shows that there is a strong relationship between the two (although not linear).

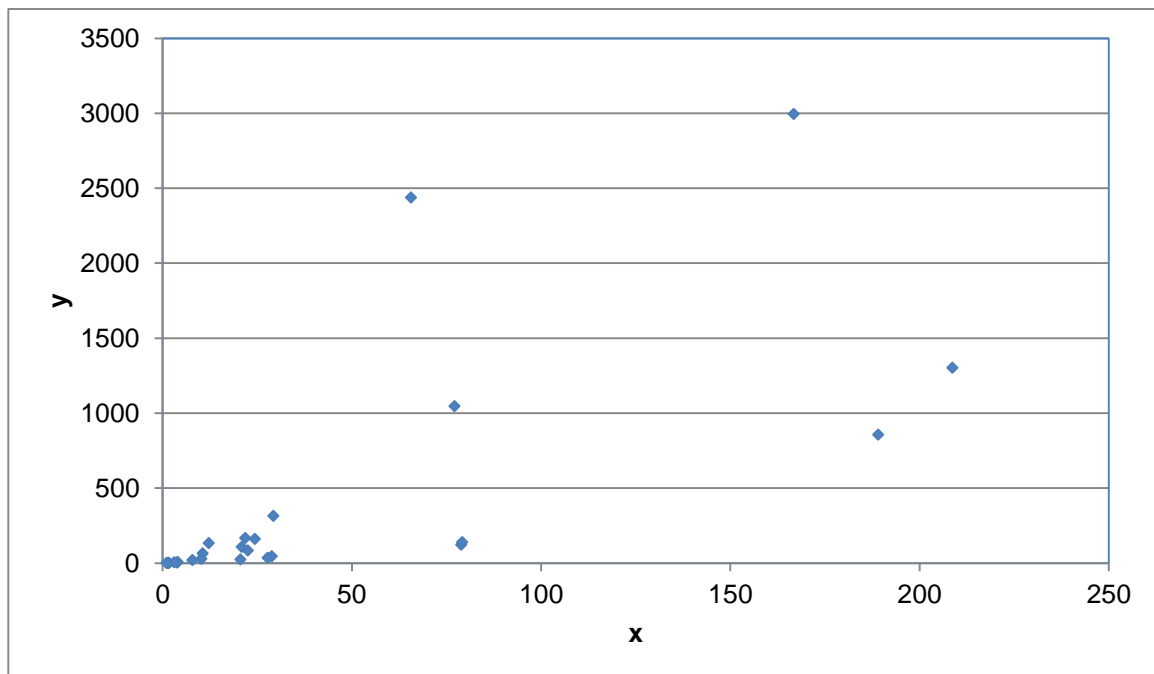


Figure 3.9 Scatter plot two Raw Variables

The chart below shows the relationship between two variables after log transformation.

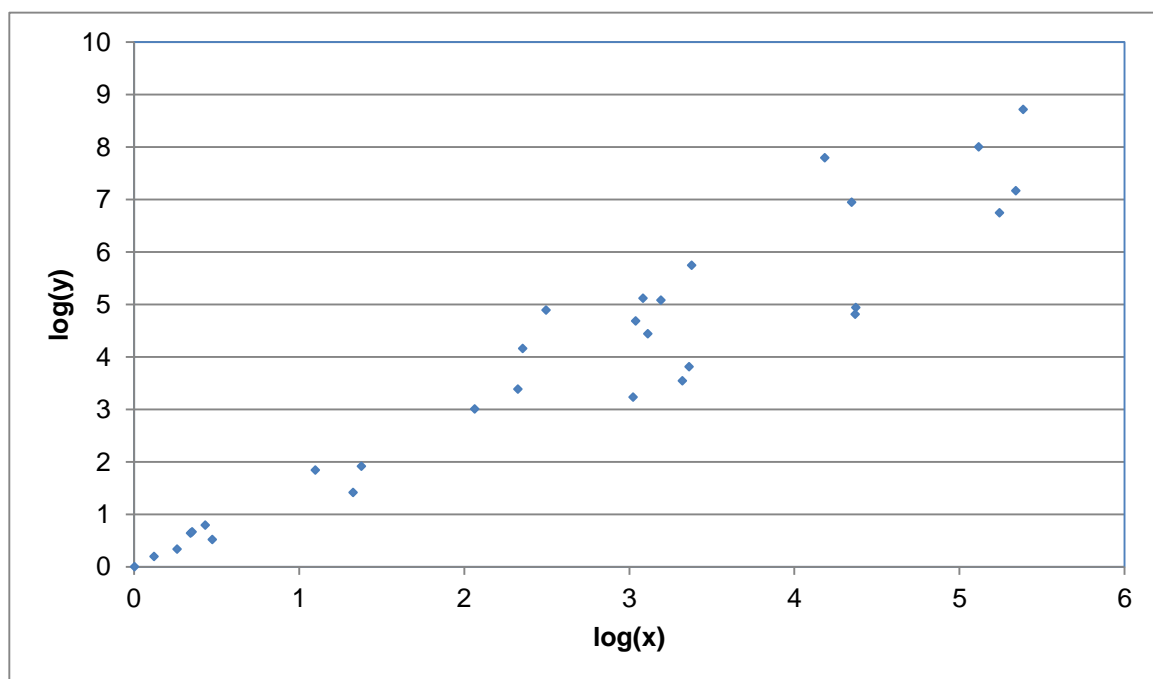


Figure 3.10 Scatter plot of treated variables

Transformation to remove non-linearity:- Some of the statistical techniques would expect linear relationship between dependent and independent variables in parameters. When the relationship is not inherently so, transforming the variable helps estimation.

For example consider the non-linear model given below.

$$S = CP^\alpha \quad (3.3)$$

Where S = sales, C=constant, P = price and α =price elasticity

By taking log on both sides we get

$$\log S = \log C + \alpha \log P \quad (3.4)$$

This can be re-written as;

$$s = a + \alpha p$$

Here a and α can be estimated by regressing s on p. What we have done is to transform the original non-linear model into a model that is linear in parameters by taking log transformation of variables. The transformed equation can be easily estimated.

The chart below shows a relationship between two variables that are not linear. It shows that the incremental value of y keeps decreasing at higher and higher levels of x. Between 50 and 100, y increased by about 3 and between 150 and 200, y increased by about 2. Hence, the function governing the relationship shrink the value x progressively.

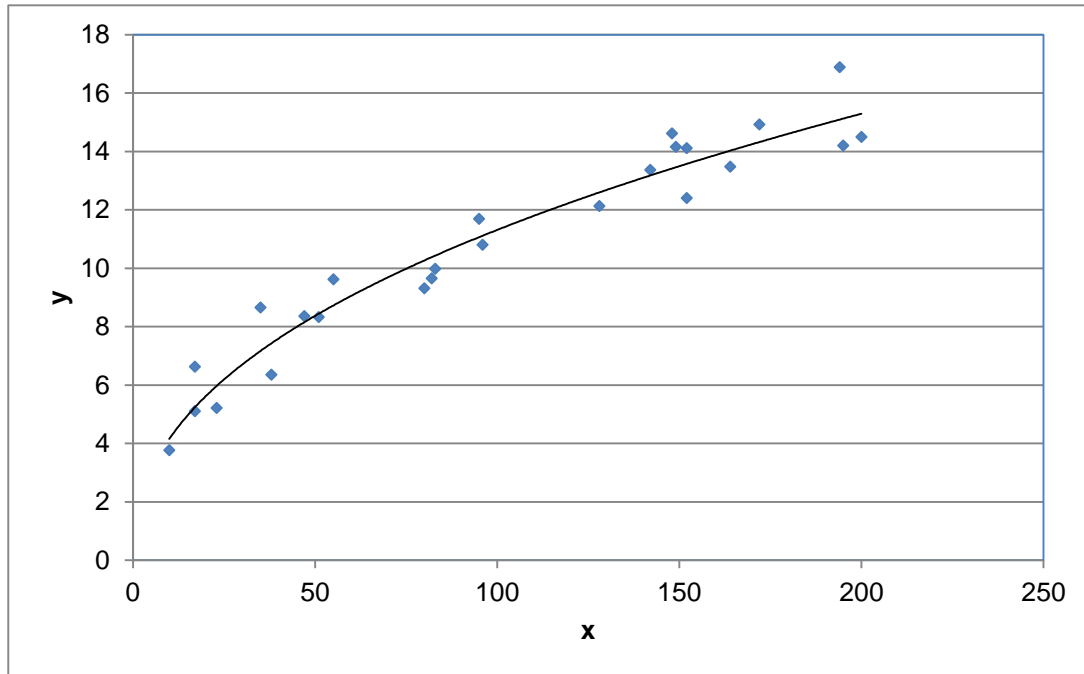


Figure 3.11 Scatter plot of x and y

Hence, we tried to transform x values by calculating the square root which will match such a behavior. The chart below shows the plot between y and \sqrt{x} . As noticed the transformation converted the relationship into linear.

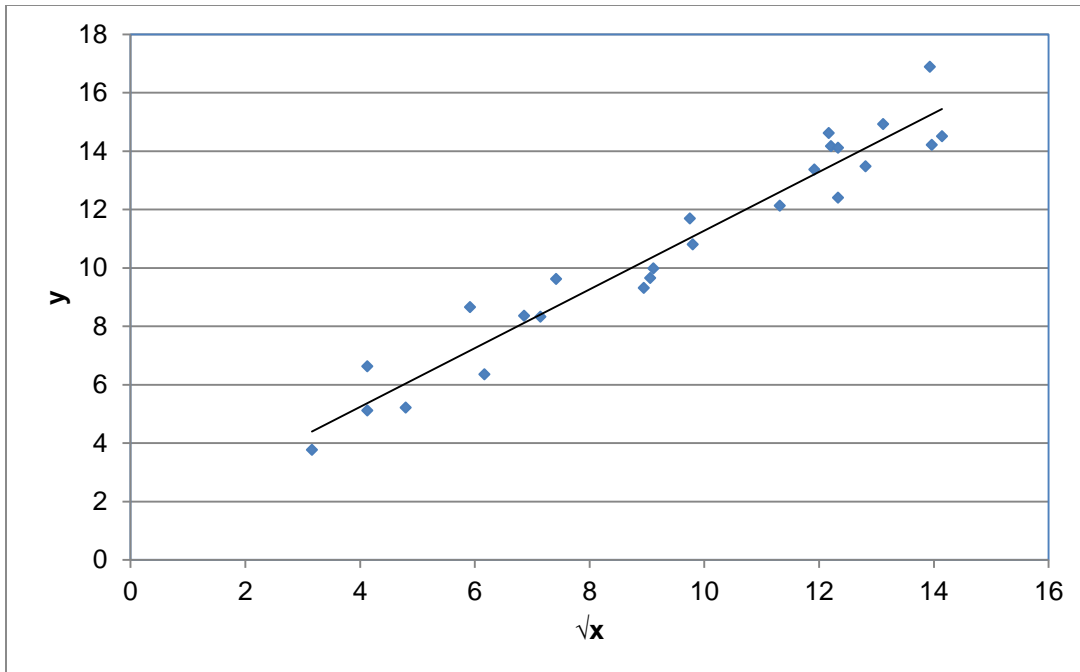


Figure 3.12 Scatter plot of transformed x and y

You will appreciate that there are other transformations that we could try to achieve the shrinkage. We could try $\log(x)$, or $1/x$ and might get similar behavior, atleast visually. However, the choice should be based on robust criteria and following points may be considered.

- Examine any theoretical basis of relationship between the variables. For example, the concept of price elasticity suggest that the relationship between price and quantity is following the functional form illustrated above (3.3). Hence, this basis can be used for appropriate transformation.

Also by logic, growth models in time are represented as exponential model. For example, if income grows $r\%$ every year, it can be represented as an exponential model.

$$I_t = I_0(1+r)^t \quad (3.5)$$

Here I_0 is the starting income, r is the yearly growth rate, I_t is the income at year 't'.


- It is worthwhile to check the industry practice of handling the relationship between certain variables. There are many advantages in following the practice. Most importantly, the results will be comparable to other studies. You will notice many such accepted relationships in the areas of media, risk analysis etc.

Transforming Categorical Variables

‘Categorical variables’ by definition carries the information of multiple categories of a variable. This information is qualitative in nature and most techniques expect quantitative information. Following are the typical coding approaches for this conversion.

GLM Coding:- Examples of typical categorical variables include gender, occupation, region, usage plans etc. In order to prepare such variables for analysis, dummy coding is used. In this approach, a number of new variables are created to represent each of the level in the category. Under GLM Coding a new binary variable is created for each of the category.

Customer ID	Occupation	Gender
345006	Business	Male
345007	Service	Male
345008	Professional	Female
345009	Business	Female
345010	Business	Male
345011	Service	Female
345012	Professional	Male
345013	Education	Male
345014	Education	Female
345015	Service	Female



Customer ID	Business	Service	Professional	Education	Male	Female
345006	1	0	0	0	1	0
345007	0	1	0	0	1	0
345008	0	0	1	0	0	1
345009	1	0	0	0	0	1
345010	1	0	0	0	1	0
345011	0	1	0	0	0	1
345012	0	0	1	0	1	0
345013	0	0	0	1	1	0
345014	0	0	0	1	0	1
345015	0	1	0	0	0	1

Table 3.11 GLM Coding of Categorical Variable

This is used in Proc GLM and the parameter estimates of classification main effects estimate the difference in the effects of each level compared to the last level.

Reference Coding:- In the above coding scheme it may be noticed that all the variables are not required for representing the levels as all levels can be uniquely determined by one variable less. For example to represent gender, we do not need 2 variables as when one is zero other is always one. This means we need only the variable ‘Male’ or ‘Female’. If we choose Female as the variable, 1 indicates means Female and zero indicates Male. In this scheme, Male is being used as the reference level. Hence, this coding scheme is named Reference coding.

Reference coding example of the data above is shown below. Here, Male is the reference level for gender and hence it is zero when the gender is Male. For occupation, Education is the reference and hence it is zero for all occupation variables when it is Education.

Customer ID	Occupation	Gender	Customer ID	Business	Service	Professional	Female
345006	Business	Male	345006	1	0	0	0
345007	Service	Male	345007	0	1	0	0
345008	Professional	Female	345008	0	0	1	1
345009	Business	Female	345009	1	0	0	1
345010	Business	Male	345010	1	0	0	0
345011	Service	Female	345011	0	1	0	1
345012	Professional	Male	345012	0	0	1	0
345013	Education	Male	345013	0	0	0	0
345014	Education	Female	345014	0	0	0	1
345015	Service	Female	345015	0	1	0	1

Table 3.12 Reference Coding of Categorical Variable

Generalizing this; if there are k levels, $k-1$ variables are required to dummy code it. Statistical techniques like linear regression require reference coding for analysis. This coding is most commonly used.

Deviation Coding:- Here too we have only $k-1$ dummy variables to represent k levels in the category of a variable. It is different from reference coding as the reference level is filled with -1.

Customer ID	Occupation	Gender	Customer ID	Business	Service	Professional	Female
345006	Business	Male	345006	1	0	0	-1
345007	Service	Male	345007	0	1	0	-1
345008	Professional	Female	345008	0	0	1	1
345009	Business	Female	345009	1	0	0	1
345010	Business	Male	345010	1	0	0	-1
345011	Service	Female	345011	0	1	0	1
345012	Professional	Male	345012	0	0	1	-1
345013	Education	Male	345013	-1	-1	-1	-1
345014	Education	Female	345014	-1	-1	-1	1
345015	Service	Female	345015	0	1	0	1

Table 3.13 Deviation Coding of Categorical Variable

The advantage of this coding is that while obtaining regression result, it is possible to obtain an estimate of a coefficient for the reference level (negative sum of coefficient for other categories). It is important as otherwise interpretation of the result is not straight forward.

3.5 REDUCING DIMENSIONS - PRINCIPAL COMPONENT ANALYSIS (PCA)

It is quite common for Analytics projects to deal with numerous but similar variables. If the variables are time series in nature, we could handle this easily by summarizing at appropriate level as discussed above. However, there could be large number of distinct variables with certain amount of redundancy between some of the variables. Redundancy occurs when few of the variables are measuring more or less same phenomena and will be correlated to each other. The objective of Principal Component Analysis is to reduce

such variables to a smaller number of components that will account for most variation in the data. Hence, PCA is quite useful when,

- There are too many variables
- There are strong correlation between few groups of variables.

It is quite common to have such variables when the business is trying to measure qualitative factors like customer or employee satisfaction, evaluation of a service experience, quality evaluation etc (Kulcsár E., 2010; Boitor, A.B. et al 2011; Gupta, S. et al 2010) . There are many applications of this technique in areas of finance like equity research, financial planning and company valuation as well (Novosyolov, A et al, 2008; Pearson, M.J. 2009, Tolmasky, C., 2002; Fengler, M.R. et al 2003). PCA is commonly applied to reduce the number of variables in such cases.

PCA Example from Financial Sector

The example is part of a research study conducted to evaluate categorization of financial ratios. The objective was to test the common categorization on whether it matches with common practice. Some of the ratios considered are provided below.

ID	Var	Ratio
V1	NFATCE	Net Fixed Assets to Capital Employed
V2	CPTTA	Cash profit to total assets
V3	SHFTTL	Share-holders' Fund to Total Liabilities
V4	CPTAVCE	Cash Profit to Avg. Capital Employed
V5	CPTTI	Cash profit to total income
V6	CPTAVSHF	Cash Profit to Avg. Shareholder's Fund
V7	TDTTA	Total Debt to Total Assets

Table 3.14 Variable List of Finance Example

The analysts have the option to analyze each of these ratios individually. However, through PCA, we are trying to reduce the number of variables to 2 or 3 which will account for most of the variation in the data.

An indication of any redundancy between the variables is reflected in the correlation table. The table shows that few variables are correlated to each other that raise suspicion that they could be measuring same phenomenon. V4, V5 and V6 are correlated to each other and all are indicators of the strength of cash position in the organization. While the correlated pair of V1 and V3 are measures of capital structure.

Correlations							
	V1	V2	V3	V4	V5	V6	V7
V1	1						
V2	0.038	1					
V3	0.414	0.030	1				
V4	0.046	0.105	0.072	1			
V5	0.233	0.301	0.224	0.580	1		
V6	0.207	0.211	0.213	0.625	0.453	1	
V7	0.067	0.154	0.260	0.147	0.426	0.244	1

Table 3.15 Correlation Matrix

Also we notice that the variables in these two groups are not correlated between them. Hence, these are two distinct components and can be named 'Cash Position' and 'Capital Structure'. To formalize this, we need an approach to identify such variables, measure about how strong is the interlinkage and a method to combine these correlated to variable to create a new variable. The steps in Principal Component Analysis achieve these.

Extracting Components:- Principal components are a linear combination of the variables and the components are identified by variables with higher loadings. These are extracted in such a way that the components are orthogonal to each other. Loadings of each of these variables on the extracted components are provided below.

	Component	
	1	2
V1	0.069	<u>0.768</u>
V2	<u>0.534</u>	0.022
V3	0.094	<u>0.784</u>
V4	0.338	0.027
V5	<u>0.663</u>	0.413
V6	<u>0.737</u>	0.403
V7	<u>0.755</u>	0.227

Table 3.16 Component loading of Variables

Usually all loadings greater than 0.5 is considered for naming. Variables like V2, V5, V6 and V7 are loaded on the first component and all are related to Cash of the business. Hence, it can be named as 'Cash Position'. While V1 and V3 are loaded on component 2 and both are Capital Structure related variables. Hence, component 2 can be named as 'Capital Structure'.

First component is extracted such that it accounts for maximum variance of variables. Rest of components account for lesser and lesser variance. Conceptually there are as many components as the number of variables. Considering all this is meaningless as our interest is to consider few components that account for maximum variation. Table below gives the variance extracted by each of the components.

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.016	28.806	28.806	2.016	28.806	28.806
2	1.538	21.976	50.781	1.538	21.976	50.781
3	.967	13.813	64.595			
4	.894	12.777	77.371			
5	.703	10.045	87.416			
6	.511	7.296	94.712			
7	.370	5.288	100.000			

Table 3.17 Total Variance Explained

Since the variables are standardized, the variance of each of the variable is equal to one. The table shows that first components accounted for variance of 2.016 which is 28% of total variance in the data. Hence, the component that doesn't account for variance (initial

Tips on Technique

Principal Component Analysis vs Principal Factor Analysis

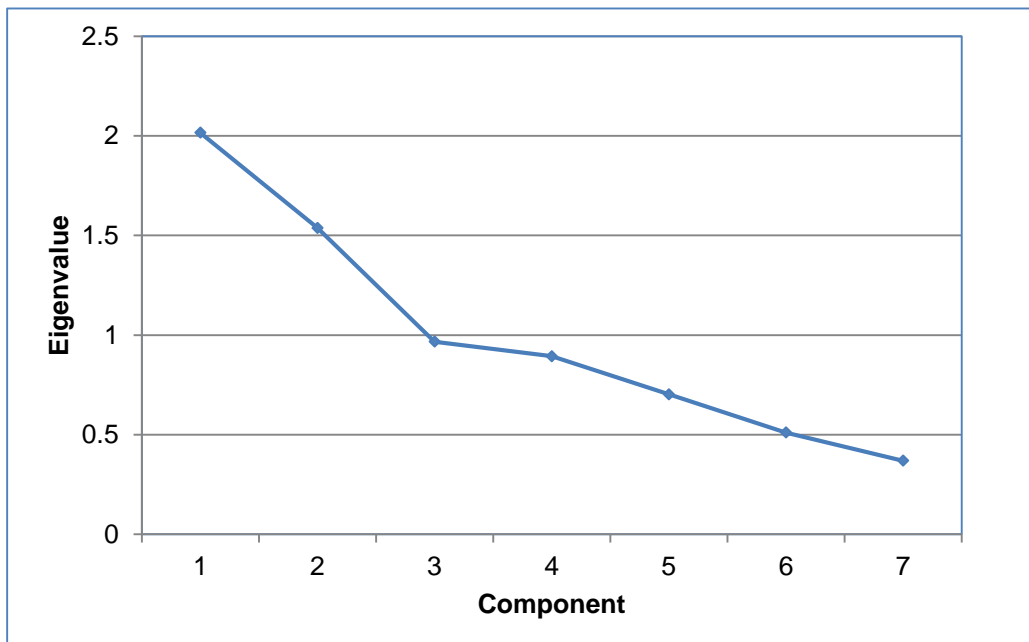
It is quite usual to confuse Principal Component Analysis (PCA) with Principal Factor Analysis (PFA) as there are many similarities between them (both are variable reduction methods). Even results in most cases will be similar. However, there are some conceptual difference between these methods.

PFA presume that certain latent factors exist that exert causal influence on the observed variables. It helps to identify the number and nature of these latent factors. PCA makes no assumption about an underlying causal model. It is simply a variable reduction procedure that results in a relatively small number of components that account for most of the variance in a set of observed variables.

PCA is a variance-focused approach in which components reflect **both common and unique variance** of the variable. While PFA is a correlation-focused approach in which the factors represent the **common variance** of variables, excluding unique variance".

Hence, PCA is generally preferred for data reduction and PFA is preferred when the purpose is detecting data structure (i.e., latent constructs or factors).

eigenvalue) at least equal one is not considered (the component is not even carrying the variation of one variable). The table above shows that only two components (1 and 2) meet this requirement.



The chart below is the plot of the eigenvalues commonly known as *scree plot*.

Figure 3.13 Eigenvalues of Components

When we are considering large number of variables that generate many components with eigenvalues larger than one, this plot could be used to select the components. Usually the scree plot will display an ‘elbow’ and all components on the left side are considered as important. In the chart below, elbow is visible at 3rd component and the first 3 component should be selected. However, in this case we didn’t select 3rd as eigenvalue was lower than 1.

Principal Component Scores:- Principal component scores are a linear combination of variables and it will be calculated for each of the subject. The coefficient used to

Tips on Technique

Principal Component Analysis / Principal Factor Analysis vs. Cluster Analysis

It is quite common to confuse between PCA/PFA and Cluster analysis. Although both are used for reduction/ simplification, it is not comparable. PFA/PCA is used for reduction of **variables** while Cluster analysis is used for reduction of **entities**.

For example, if you are dealing with large number of interrelated variables and would like to reduce the complexity by reducing the number of variables, the technique is PCA. On the other hand if your interest is group the subjects you are studying (say customers, companies etc.), the technique is Cluster Analysis.

calculate this is called score coefficients and is estimated using approaches like regression.

The table below shows the raw variables and component scores for a section of data.

Company	V1	V2	V3	V4	V5	V6	V7	Princ1	Princ2
1	0.376	0.401	0.400	0.371	0.383	0.420	0.414	0.0549	-0.7071
2	0.377	0.379	0.288	0.380	0.327	0.381	0.419	-0.5401	-0.178
3	0.428	0.516	0.302	0.371	0.195	0.293	0.326	-1.4921	-0.474
4	0.490	0.413	0.125	0.518	0.429	0.515	0.280	-0.3241	1.7169
5	0.307	0.498	0.317	0.513	0.505	0.293	0.311	0.165	-1.3131
6	0.383	0.391	0.420	0.398	0.375	0.422	0.399	0.0549	-0.7071
7	0.408	0.505	0.281	0.483	0.517	0.482	0.480	1.7837	0.4701
8	0.409	0.414	0.490	0.388	0.385	0.389	0.276	-0.2507	-1.1262
9	0.478	0.490	0.476	0.485	0.487	0.503	0.373	1.5197	0.2108
10	0.410	0.478	0.206	0.398	0.406	0.391	0.379	0.1207	0.2671

Table 3.17 Section of Data with Raw Variables and Component Scores

The table shows that raw value for company-1 for V1 to V7 vary from 0.376 to 0.414. The estimated value for principal component 1 (cash position) is 0.0549 and principal component 2 (capital structure) is -0.7071 for Company-1. It may be noted that values of these components are unlike the variables, although derived from these variables. Some scores are even negative. This is because the scores are calculated on the standardized values of the variables. For further analysis, two derived variables (Cash Position and Capital Structure) may be used instead of the original 7 variables.

The advantage of the approach should be very clear especially when there are numerous variables with interrelationship. This will make the analysis simpler and more focused.

Questions

1. A Human Resource Information System (HRIS) contain following variables. What are the levels of measurement of each of these?

- Name
- Age
- Gender
- Joining Date
- Experience (years)
- Performance Review Score (a scale of 1-5; 1 – poor to 5- Excellent)
- Basic Pay
- Salary Increment

2. In a customer profile dataset, Is the following missing values legitimate? Will you replace with zero?

- Income
- Age
- Loan outstanding
- Debt Ratio

3. A consumer banking business is interested in evaluating delinquency of customers. They are planning to use such a model to approve loan applications. It has collected application data and the record of delinquency of its customers for last few years. DSCH03PREPDLQW.xls contains this data. Prepare the data for analysis. Focus on missing values and outliers. Compare the raw and prepared datasets.

4. A leading credit card business organization sends mails to existing customers on various new products. So far the organization has been following a policy of ‘carpet bombing’ ie send mail to all customers. The management is very keen to reduce the cost of mailing and also not to contact customers too frequently. It is trying to apply Analytics to target customers for next mailings. It has collected response information and customer profile of the last mailing and is available in the file DSCH04EXPLRESW.xls. Prepare the data for analysis by cleaning the database.

5. Convert following equations into equations linear in parameters.

1. $y = \beta_1 + \beta_2 x_i + \beta_2 x_i^2$
2. $y = \beta_1 + \beta_2 x_1 + \beta_3 x_2 + \beta_4 x_1 x_2$
3. $y = \beta_1 + \beta_2 x_1^2 + \beta_3 \sqrt{x_2} + \beta_4 \log x_3$

Research Questions

1. Review the literature and connect type of missing value imputation and the context. Context can include proportion missing, type of application, industry etc.
2. Review literature for a published Analytics studies and document various data preparation strategies adopted. Connect these strategies to the purpose of the study.

References

- Barnett, V. and Lewis, T.: 1994, Outliers in Statistical Data. John Wiley & Sons.,3 edition.
- Boitor, A.B. & Bratucu, G. 2011, "Competition Analysis On The Operating System Market Using Principal Component Analysis", Bulletin of the Transilvania University of Brasov.Economic Sciences.Series V, vol. 4, no. 1, pp. 15-22.
- Cock,A.C., Working with missing values, <http://people.oregonstate.edu/~acock/growth-curves/working%20with%20missing%20values.pdf>, Accessed on 9/Aug/2013
- Cook, R. Dennis (February 1977). "Detection of Influential Observations in Linear Regression". *Technometrics* (American Statistical Association) 19 (1): 15–18.
- Elliott, M.R & Stettler, N. (2007). Using a mixture model for multiple imputation in the presence of outliers: the ‘Healthy for life’ project. *Applied Statistics*, 56, 63-78.
- Erika Kulcsár (2010), Principal component analysis in tourism marketing, *Management & Marketing*, Vol.5, No.2, pp., 151-158
- Fengler, M.R., Hardle, W.K. & Villa, C. 2003, "The Dynamics of Implied Volatilities: A Common Principal Components Approach", *Review of Derivatives Research*, vol. 6, no.3, pp. 179-202.
- Gupta, S. & Mehra, P. 2010, "An Empirical Analysis Of The Factors Influencing The Purchase Behaviour Of Micro-Brands", *Interdisciplinary Journal of Contemporary Research In Business*, vol. 2, no. 6, pp. 433-450.
- Mamdouh Refaat (2007) *Data Preparation for Data Mining Using SAS*, Morgan Kaufmann Publishing: San Francisco, CA.
- Novosyolov, A. & Satchkov, D. (2008), "Global term structure modelling using principal component analysis", *Journal of Asset Management*, vol. 9, no. 1, pp. 49-60.
- Pearson, M.J. 2009, *Statistical inference of securities pricing using Principle Component Analysis*, University of Louisville.
- Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*. J. Wiley & Sons, New York.
- Stevens, S. S. (1946). "On the Theory of Scales of Measurement". *Science* 103 (2684): pp.677–680.

Stevens, J. S. (2009). Applied multivariate statistics for the social sciences. New York, N.Y. : Routledge.

Tolmasky, C. & Hindanov, D. 2002, "Principal components analysis for correlated curves and seasonal commodities: The case of the petroleum market", The Journal of Futures Markets, vol. 22, no. 11, pp. 1019-1035.