
Chapter 13 – Data Quality WITH Analytics

13.1 Introduction

This chapter shows how data quality can benefit from analytics. Chapters 3-9 in the first part of this book have shown that analytics poses special requirements on data quality. This chapter will show that analytics also provides special capabilities to data quality.

These capabilities include both the profiling of the data quality status and the improvement of the data quality.

Chapter 10 and 11 have already shown an important capability of analytics, namely to deal with missing values. This chapter will focus on additional features of analytics for data quality. The main topic here will be detection of outliers, which includes classical outlier detection, outlier detection with predictive models, time series models and cluster analysis.

Missing value imputation and complex outlier detection are the main two important capabilities of analytics in data quality control, however there are also additional features that will be named in this context.

13.2 Benefit of Analytics in General

The following points give an overview over the typical features of analytics that are relevant for data quality.

Outlier Detection

- Simple profiling of univariate data. Analytics plays an important role to detect outliers based on statistical measures like standard deviations or quantiles.
- Outlier detection with methods of cluster analysis and distance metrics. These methods allow the identification of outliers in the data from a multivariate viewpoint.
- Individual outlier detection with predictive models and time series methods. These methods allow calculating validation limits and most likely correction value on an individual basis. Not an average value over all analysis subjects and time periods is being used, but a value that for example compares against peer groups averages.
- Analytics allows not only profiling and identifying outliers and non-plausible values, it also can provide a most probably value that should be entered here.

Missing value imputation

- Analytics can deliver replacement values for missing values. In chapter 10 and 11 it has been shown how missing values in one-row-per-subject data marts and time series data marts can be imputed.
- These imputation methods range from average based imputation values up to analysis-subject-individual-imputation-values, which are based on analytic methods like decision trees or spline interpolations for time series.

Data standardization and de-duplication

- Identification and elimination of duplicates in database where no unique key for the analysis subjects is available, is based on statistical methods that describe the similarity between records.
- These methods provide a measure of closeness and similarity between records which is based on information like address, name, phone number, account number and others.

Handling of data quantity

- Analytics allows planning the optimal number of observations for a controlled experiment with sample size and power calculation methods.
- In the case of small samples or small number of events in predictive modeling, methods for the modeling of rare events are provided.
- In the case of time series forecasting so called “Intermittent Demand Models” are provided that model time series that only have occasional non-zero quantities.

Analytic transformation of input variables

- Analytical methods are used to transform variables to a distribution that is suited for the respective analysis method. Log and square root transformations are for example used to transfer right skewed data to a normal distribution.
- For variables with many categories, analytics provides methods to combine categories. Here the combination logic for these categories depends on the number of observations in each category and the relationship to the target variables. Examples for these methods include decision trees or weight of evidence calculations.
- Text mining allows converting free form text into structured information that can then be processed by analytical methods.

Variable selection

- Various methods for variable selection allow the identification of the subset of variables which have a strong relationship with the target variable in predictive modeling. These methods include simple metrics like the R-square and advanced metrics like LARS, LASSO, compare also [16].
- Many analytical methods allow different methods of variable selection in the respective analysis model itself. Consider for example the forward, backward and stepwise selection in linear regression.

Assessment of model quality and what-if analyses

- Analytical tools are often designed to assist the analysis in the model creation and model validation process. In predictive modeling for example it is often important to get a quick initial insight in the predictive power of the available data (Rapid Predictive Modeling).
- These tools also provide measures to assess model quality very quickly and to provide features for what-if-analyses.
- What-if-analyses are especially important for the determination of the importance of variables or groups of variables. Here the consequences on the predictive power shall be estimated if certain variables are not available.

13.3 Classical Outlier Detection

Ways to define validation limits

The purpose of outlier detection is to identify those observations that are considered to be outside of the typical value range for a variable. The definition of the “typical range” can be either done

- on statistical measures of the variable itself or
- on rules from a business or domain specific point of view.

If business rules are available to define upper or lower limits these are usually applied. For example for variable “age in years” it usually makes more sense to define domain specific limits then calculating limits based on the mean and standard deviation of the age values.

If from a business or domain specific point of view no validation limits can be defined or if the respective definition requires too much effort or is unfeasible at all, statistical validation limits can be used to decide about the outlier status of a value for an observation.

Purpose of outlier detection

The purpose of defining these limits and checking these values is mainly to make sure that only valid and plausible observations are used for the analysis. From a statistical point of view, it is for some methods however also advisable to identify observations as outliers even if their values are plausible and correct. Many analytical methods for example expect interval data to have a distribution that is close to the normal distribution.

- For example in a distribution of the usage duration of a certain service for example most values are around 120 hours. The distribution has two extreme outliers with 2430 hours and 4302 hours. From a business point of view these values are correct and plausible, they will however be most likely filter or shifted for the analysis in order to achieve a more well-shaped and central distribution.

If outliers are identified in data there are two strategies to deal with the outliers.

- Observations with outliers can be excluded from the analysis. This is also called filtering.
- Alternatively the values that are considered as outliers are replaced by appropriate replacement values. Often these replacement values are located closer to the center of the distribution. This is then called shifting of values.

Statistical methods

Statistical outlier detection usually uses information from the whole observations base to calculate validation limits. These limits are then applied for each observation independently. Statistical methods to define outliers include the following methods:

- Statistical measures like the mean of the distribution +/- the standard deviation multiplied by a factor.
 - Here the calculation of the mean and the standard deviation can be performed on the whole data itself or on a “central” subset of the data in order to avoid a biasing of the calculation of the mean and standard deviation by outliers.
 - The determination of the number of standard deviations is domain specific. 3 standard deviations for example will result in approximately 1 % of outliers for a normal distributed. SAS® Enterprise Miner for example uses 3 standard deviations in its default settings.

- Quantiles like the 1% and the 99 % quantile
- Special forms of trimmed means, robust estimators or other quantile based measures like the 1.5 fold interquartile distanced added to the 3rd quartile and subtracted from the 1st quartile. Note that this is definition which is also used by many statistical software packages for outliers in box plots.

Implementation

The above methods can easily be calculated in SAS and applied to data marts. SAS offers a wide range of functions and analytical procedures to calculate virtually any statistical validation range. Compare Cody [9] for coding examples how standard deviation and percentile based limits can be calculated.

SAS/QC also provides methods for statistical quality control. Here PROC SHEWHART provides many features to define outliers based on statistical measures. These methods can be used for data quality control as well. Compare also [17], Svolba, for examples how to use methods of statistical quality control for clinical trials, which also has a relation to data quality control.

Outlier detection with analytic methods

The following sections will show examples for outlier detection with more advanced analytical methods, which will include

- Outlier detection with predictive modeling
- Outlier detection in timer series analysis
- Outlier detection in cluster analysis

13.4 Outlier Detection with Predictive Modeling

General idea

An extension to the method described above is to determine the validation limits not from the respective variable in a univariate ways but to calculate a most likely value based on other variables. These other variables in this case can be demographic variable like age, sex, region or other variable from the base table which makes sense to individualize the validation limits.

Based on these variables predictive modeling is applied. The predictive model uses a set of base variables as input variables and uses them to predict as target variable the variable that shall be validated. This leads to the fact that an expected value (reference value) for the variable based on the values of the input variables is created

The individualization can go that far that for each analysis subject an individual reference value is calculated. If only a few categorical variables are used the reference value will correspond to the so defined segments. This can also be understood as a peer group approach, where similar analysis subjects get the same reference value.

This reference value can then be compared against the actual value. The deviation between the reference value and the actual value is then used to judge whether the value shall be considered as outlier.

The reference value is then used to define the individual validation limits. The methods to do this include the following:

- Limits from a business point of view in absolute numbers

- Limits from a business point of view, which are calculated as absolute or relative differences between the values.
- Calculating the upper and lower limits based on the standard deviation for the predicted value or the deviation.

In any case the decision whether a value is an outlier is based on the reference value which is based on variables of the respective observation.

Methods in SAS

Analytical methods to calculate individual validation limits usually include:

- Linear model like linear regression or a general linear model. The advantage of this method is that for the prediction and the residual a standard deviation is calculated as well. This method can be performed in SAS® Enterprise Miner with different regression nodes and in SAS®STAT for example with the REG or the GLM procedure.
- Decision trees. They have the advantage that they can automatically detect interactions in the data and thus provide very specific reference values, however the set of prediction (expected) values is not continuous as only that many different predictive values are available as the number of leafs in the tree. This method is available in SAS® Enterprise Miner. See also [3] Schubert.

Example for laboratory data

In order to illustrate the concept described above, a laboratory dataset from a clinical trial is used. The following variables are collected for each patient:

- Age (4 groups)
- Sex
- Weight (3 groups)
- Melanoma Stage (2 stages)
- Trial Center ID (8 centers)

For each patient, data is collected over time at different visits. For illustrative purposes the cholesterol value is used here to show how a predictive model based on the above variables is used to calculate individual reference values. As all input variables are categorical the maximum number of different reference values is $4 \times 2 \times 3 \times 2 \times 8 = 384$. In the example data that are shown below the 403 patients fall into 181 different categories with respect to the grouping shown above. For these 403 patients, 3154 measurements have been made. An excerpt of the analysis data is shown in table 13.1.

Table 13.1 – Excerpt of the analysis data.

PATNR	Age_Grp	SEX	Weight_Grp	STAGE	CENTERNR	VisitDate	CHOL
232	19-30	0	80+	1	1	16SEP1999	237.00
232	19-30	0	80+	1	1	08OCT1999	193.00
232	19-30	0	80+	1	1	24DEC1999	194.00
232	19-30	0	80+	1	1	23FEB2000	205.00
232	19-30	0	80+	1	1	24MAY2000	217.00
232	19-30	0	80+	1	1	14MAR2001	211.00
232	19-30	0	80+	1	1	20JUN2001	223.00
232	19-30	0	80+	1	1	19SEP2001	218.00
191	60+	0	80+	2	1	11AUG1999	139.00
191	60+	0	80+	2	1	06OCT1999	166.00
191	60+	0	80+	2	1	10NOV1999	166.00
605	46-60	1	66 -	1	4	29DEC1999	169.00
605	46-60	1	66 -	1	4	01MAR2000	188.00
605	46-60	1	66 -	1	4	06SEP2000	185.00
605	46-60	1	66 -	1	4	06DEC2000	158.00
605	46-60	1	66 -	1	4	13JUN2001	174.00
225	46-60	0	80+	1	3	23SEP1999	211.00
225	46-60	0	80+	1	3	29OCT1999	185.00
225	46-60	0	80+	1	3	21MAR2000	193.00
225	46-60	0	80+	1	3	05SEP2000	203.00

Based on this data a predictive model is created that uses CHOL as target variable (y) and the other variables (except PATNR and VISITDATE) as categorical input variables. Proc GLM is used to calculate the model.

```
proc glm data=labor_chol_data;
  class sex centernr stage age_grp weight_grp ;
  model chol = age_grp sex weight_grp centernr stage;
  output out=pred_chol p=reference r=residual
         stdi=stdi stdr=stdr stdp=stdp;
run;
quit;
```

Table 13.2 shows the output table, PRED_CHOL, which is generated by PROC GLM. In addition to the input variables this table also holds the

- Predicted value for CHOL (variable REFERENCE)
- The residual between the predicted value and the actual value
- The standard deviations of the individual, the predicted and the residual values

Table 13.2: Output table (PRED_CHOL) with the results of PROC GLM

PATNR	Age_Grp	SEX	Weight_Grp	STAGE	CENTERNR	VisitDate	CHOL	reference	residual	stdi	stdr	stdp
144	46-60	0	80+	1	2	24AUG2001	232.00	205.87601735	26.123982652	36.494115985	36.400139366	1.8507235971
144	46-60	0	80+	1	2	02NOV2001	220.00	205.87601735	14.123982652	36.494115985	36.400139366	1.8507235971
144	46-60	0	80+	1	2	07FEB2002	244.00	205.87601735	38.123982652	36.494115985	36.400139366	1.8507235971
144	46-60	0	80+	1	2	03MAY2002	241.00	205.87601735	35.123982652	36.494115985	36.400139366	1.8507235971
144	46-60	0	80+	1	2	04AUG2002	218.00	205.87601735	12.123982652	36.494115985	36.400139366	1.8507235971
144	46-60	0	80+	1	2	08NOV2002	236.00	205.87601735	30.123982652	36.494115985	36.400139366	1.8507235971
144	46-60	0	80+	1	2	08JAN2003	253.00	205.87601735	47.123982652	36.494115985	36.400139366	1.8507235971
144	46-60	0	80+	1	2	08MAY2003	230.00	205.87601735	24.123982652	36.494115985	36.400139366	1.8507235971
144	46-60	0	80+	1	2	08AUG2003	230.00	205.87601735	24.123982652	36.494115985	36.400139366	1.8507235971
145	60+	0	80+	1	3	05NOV1999	276.00	202.50040776	73.499592235	36.563135081	36.330810622	2.9099008993
146	60+	0	80+	1	6	28NOV2000	276.00	203.511552	72.488447999	36.538977428	36.355106601	2.5887347805
146	60+	0	80+	1	6	28DEC2000	230.00	203.511552	26.488447999	36.538977428	36.355106601	2.5887347805
146	60+	0	80+	1	6	01FEB2001	231.00	203.511552	27.488447999	36.538977428	36.355106601	2.5887347805
146	60+	0	80+	1	6	01MAR2001	230.00	203.511552	26.488447999	36.538977428	36.355106601	2.5887347805
146	60+	0	80+	1	6	31MAY2001	190.00	203.511552	-13.511552	36.538977428	36.355106601	2.5887347805
146	60+	0	80+	1	6	06SEP2001	191.00	203.511552	-12.511552	36.538977428	36.355106601	2.5887347805
146	60+	0	80+	1	6	13DEC2001	220.00	203.511552	16.488447999	36.538977428	36.355106601	2.5887347805

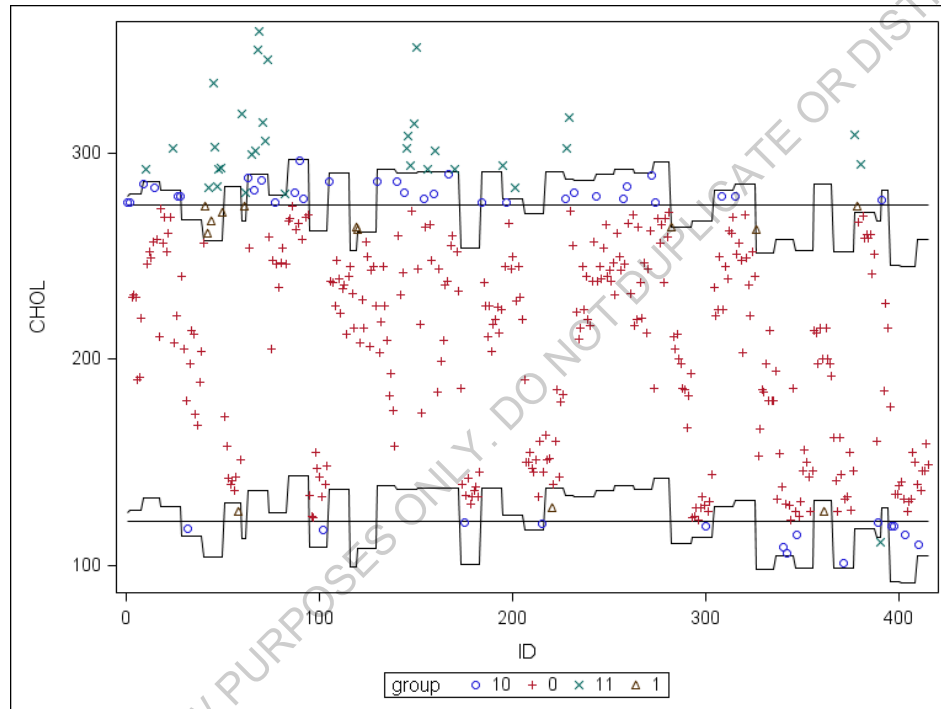
The predicted value (reference value) or the residual values can now be used together with the standard deviations to calculate individual upper and lower validation limits.

For the illustration of the usage of the individual reference values the following sets of validation limits have been defined:

- Overall validation limits UPPER and LOWER have been defined based on the overall mean of 198.34 and the standard deviation of 38.38, with a width of +/- 2 standard deviations.
- Individual validation limits UPPER_I and LOWER_I have been defined based on the individually predicted values (REFERENCE) +/- 2 standard deviations.

The results for selected patients are represented graphically in diagram 13.3.

Diagram 13.3 Scatterplot for cholesterol with different validation limits



Note the following from the graph:

- The graph shows two straight horizontal lines, which are the overall validation limits.
- The two variable black solid lines represent the individual validation limits.
- The cholesterol values are represented on the Y-axis, the ID variable represents an artificial enumeration of the observations.
 - Those observations that are within both validation limits, the general and the individual ones, are represented by a plus sign “+”.
 - Observations that would be outside the general validation limits but within the individual limits are represented as circles ‘o’. These observations are those that are now counted as “regular” observations if the individual limits are applied.
 - Observations that have not been considered as outliers with general validation but are outliers when applying the individual validation limits are represented as triangles.

- Crosses 'x' mark those observations that are outliers in both cases.
- It can be seen that the limits on an individual basis provide much more specific information about the data.

The above graph has been created using PROC SGPLOT with the following code:

```
proc sgplot data = pred_chol;  
  series x=id y=upper;  
  series x=id y=lower;  
  series x=id y=upper_i;  
  series x=id y=lower_i;  
  scatter x=id y=chol / group = group MARKERATTRS=(size=2);  
run;  
quit;
```

Note the simplicity of the code that creates a very detailed and illustrative picture.

Extension of this method

The above method can also be extended with a time dimension. Here not a static reference value for each individual group is calculated but a possible time trend of the values is considered.

13.5 Outlier Detection in Time Series Analysis

General

The decision whether a specific value is an outlier or not is in many cases not only based on the value itself. An example of predictive modeling has been shown in the previous subsection. This method could now be extended to include time series information as well. For example if it can be assumed that the values underlie seasonal patterns it will make sense to differentiate the expected values by season, for example calendar month, as well.

Time series models

If the data that shall be quality controlled are taken from a process over time, it is more advisable to perform the “predictive” model as a time series model which includes typical time series elements like seasons, trends, cycles and shifts.

In this case the modeling will be performed on basis of the historic time series. The decision whether a value is an outlier or will not only be based on static limits but will be based on the deviation of the forecasted value in the time series.

This can account for example for the fact, that a sales number of 251.000 pieces would be an outlier in February, but would be a normal value in December. The individual validation intervals will result in less “false alarms”.

Outlier detection with ARIMA(X) models

The detection of outliers is a very important topic in time series analysis. ARIMA(X) models can be used to filter the effect of detected outliers. The process in this case is as follows:

- An initial time series model is fit based on the available data.
- Those observations that are considered as outliers as they are too distant from the predicted values are flagged as outliers.

- A new model is fit that also includes the dummy variables for the detected outliers.
- Based on the new model new observations will maybe considered as outliers, and the process is iterated again.

SAS®Forecast Studio provides this functionality for ARIMA models. The option can be set in the DIAGNOSTICS tab of the FORECAST SETTINGS. A partial screenshot is shown in graph 13.4.

Graph 13.4 – Outlier detection setting in SAS®Forecast Studio

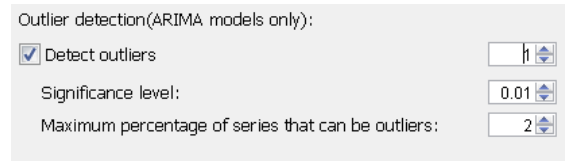
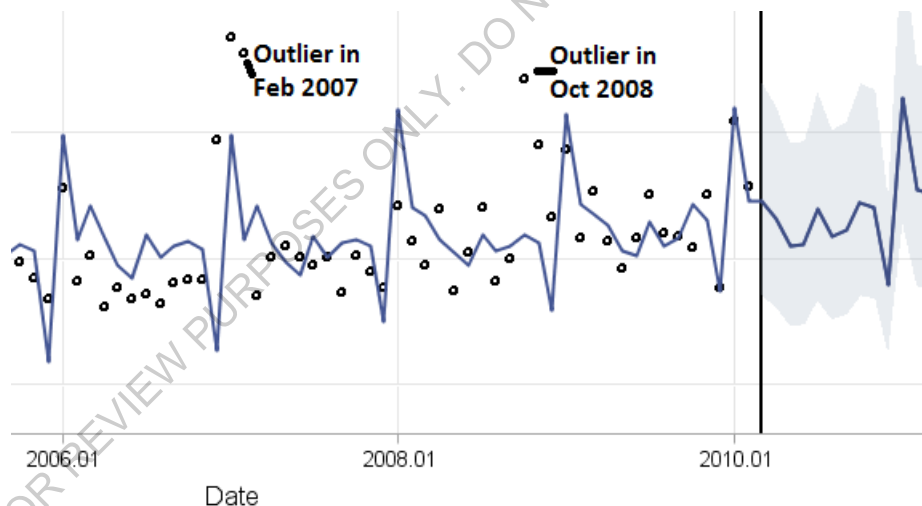


Diagram 13.5 shows a screenshot of SAS®Forecast Studio of a time series. Here the historic observations are show as circles and the forecast is shown as a solid line. It appears that some observations have very high values.

The ARIMAX model that has been built for these data with the “Detect Outlier” Option identifies the observations in February 2007 and October 2008 as outliers.

Diagram 13.5 – Line plot of a time series in SAS®Forecast Studio



The graph annotates the observations that are considered as outliers. It can be seen that the value in January 2007 (upper left to the outlier in February 2007) is not flagged as an outliers, although the value is larger. The reason is that in January values are higher in general and such a value is not seen as “too high” whereas in February usually not such high values occur.

Similar to the predictive model that has been shown in the previous section; analytic models here allow a more intelligence judgment of the outliers status of certain values.

13.6 Outlier detection with cluster analysis

General

Another method to profile data is multivariate outlier detection. Different to the situation of univariate outlier detection where a rule is defined on the individual value only, multivariate

outlier detection identifies observations as outliers that are outliers in a multivariate sense. The rule is based on the combination of the values of two or more variables.

For example statistical cluster analysis can identify outliers that are not found in a univariate analysis. From univariate point of view, each variable may be within its validation limits. From a multivariate point of view however it may appear that for a few cases special combinations of variables occur that make an observations to an outlier as these combinations are unusual.

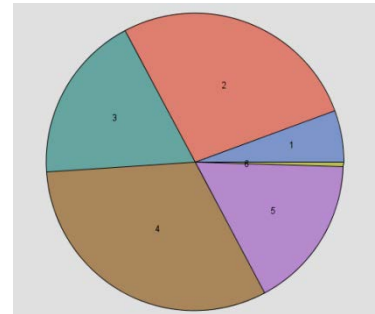
The advantage of this method is that outliers are found that would otherwise stay hidden. On the other hand side the definition of an outlier is not that straightforward as with univariate rules and validation limits.

This method is not only applied for outlier detection but also to identify suspicious cases in fraud detection, see also [3] Schubert.

Diagram 13.6 shows the results of the clustering in SAS®Enterprise Miner of insurance customers based on variables like age, income, bluebook value of the car and other factors. It can be seen that the observations are grouped into six clusters. Cluster 6 is a rather small cluster with 61 observations only (from 10303). From a univariate perspective the observations in this cluster have average values for example for age and income.

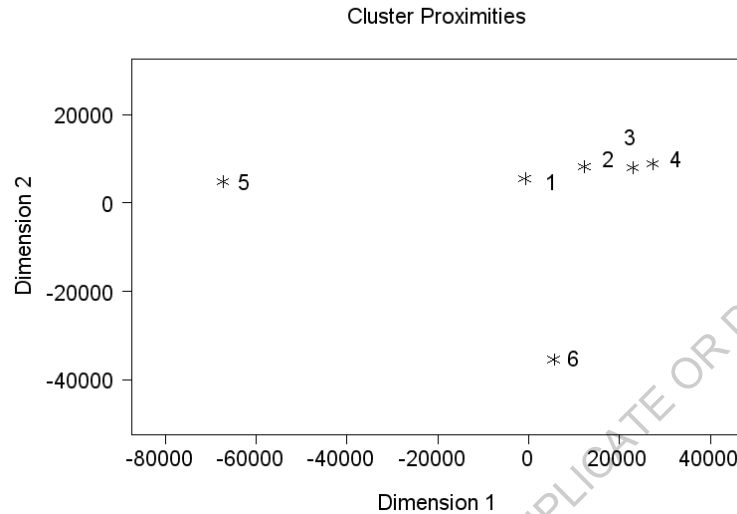
Diagram 13.6 – Clustering results of insurance customers

Segment Id ▼	Frequency of Cluster	Age	Income
6	61	44.19672	62171.06
5	1722	46.30081	133086.8
4	3265	38.83691	39514.4
3	1868	41.47185	43807.37
2	2810	52.6306	54005.21
1	577	47.37674	66922.92



From a multivariate point of view however it can be seen in the “Cluster Proximities” chart in Diagram 13.7, cluster 6 is distant from the other clusters.

Diagram 13.7 – Cluster Proximities chart of the insurance customer clustering.



Conclusion

Analytic methods have helped here to profile the data from an advanced perspective and to surface relationships in the data. In data quality profiling, these methods can be applied to identify a subset of observations that differ from the others. It has then to be decided, whether this combination is considered to be within usual business interpretation or shall be checked for possible biases or errors in the data.

13.7 Recognition of Duplicates

General

The recognition of duplicates is an important topic in data quality control. Records in a table or in related tables shall be checked for duplicates. This task is simple as long as unique keys for the records are available. In the absence of unique keys however different attributes of the record like address, name, bank account or phone number are used to create so called surrogate keys to identify the records.

The recognition of duplicates is usually preceded by a standardization of the values in these fields. The standardized version can then be used to find out how similar (or close) the respective records are to each other. Records that have a high similarity may be candidates for duplicates.

Contribution of analytics

Analytics provides methods to define the similarity between records. These similarity measures can be based on fuzzy matching algorithms. Fuzzy matching methods are based on fuzzy sets and mirror the concept of degrees of membership. Here an observation is not only assigned to a single set but is assigned multiple sets with a respective probability. Other similarity methods include for example Euclidean distances or clustering methods like k-means clustering or hierarchical clustering.

13.8 Other Examples of Data Profiling

General

Even if data does not have any missing values and all values fall into the predefined ranges data may not be useful for a particular reasons. Data may be falsified or artificially created. A sub discipline in fraud detection, forensic data analysis, deals with the analysis of the process of data creation and the checking whether data contains unusual pattern.

Thus there is a link between data quality check and detection of data anomalies

Benford's law for checking data

One feature that is frequently checked in the fraud detection context is the distribution of the first (largest) digits of numbers. In 1938 Frank Benford [18] stated, based on the work of Simon Newcomb, that the numbers from many real-life situations follow a specific distribution. This distribution is non-uniform. Moreover the distribution follows the logarithmic scale. This fact is known as Benford's law.

Benford's law can be formulated as follows: $P(d) = \log_{10} (1-1/d)$.

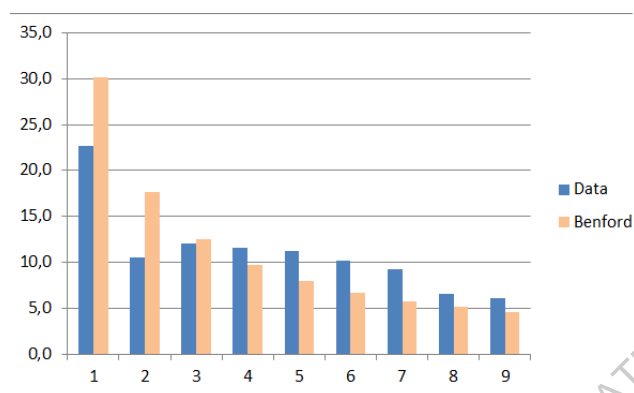
This means that the probability that the largest digits equals d, decreases with increasing d. A leading digit of '1' has a probability of 30.1 %, while a leading digit of '2' has a probability of 17.6 %.

Table 13.8 – Benford probabilities for digits 1-10.

digit	prob
1	0.30103
2	0.17609
3	0.12494
4	0.09691
5	0.07918
6	0.06695
7	0.05799
8	0.05115
9	0.04576
10	0.04139

Table 13.9 shows an example of the comparison of the distribution of the first digits on income information for customers (dark bars), with the expected percentage following Benford's law (bright bars).

Table 13.9 Bar chart for frequency of the first digit of income values



It can be seen that the true distribution does not follow the expected distribution. Analytical methods like the χ^2 -test can be applied to assess the deviation of the observed data from the expected distribution.

13.9 Conclusion

This chapter has shown how analytics can improve data quality. An overview over different areas of analytics that help to improve data quality or can be used in the data quality process has been presented. Some of these topics will be discussed in the next chapter where an overview over SAS Analytic Tools with the respect to data quality is given.

The main focus of this chapter was to show how analytics can perform an advanced definition of outliers. Methods have been presented for calculating individual reference values and to retrieve individual validation limits. Methods like predictive modeling or time series analysis have been presented in this context. These methods allow an individualization of outlier detection and reduce the number of false alarms.

Another method that has been presented for outlier detection is the cluster analysis. Here complex patterns of outliers can be found that would be overseen in a univariate analysis.