International Journal of Engineering & Technology, 7 (4.6) (2018) 209-213

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



Data Analytics: Why Data Normalization

DR. K. Dhana Sree[1], Dr. C. Shoba Bindu[2]

[1]Professor, Dept of IT, Vardhaman college of Engineering, Hyderabad
[2]Professor, Dept of CSE, JNTUCEA, Anantapur
**Corresponding author E-mail:drdhanasreek@gmail.com*

Abstract

The two maestros Artificial Intelligence and Machine learning are ruling the data filled world with good analytics. Many of these domain skills are used in the industry to analyze and interpret the data beyond what it actually is. Supporting the known saying find the horse before the cart is ready is what it mean to normalize the data before getting it analyzed. This article focus on what normalization actually is, why normalization is needed before data analysis and how data normalization is done.

Keywords: *Artificial Intelligence, Machine Learning, Data Science, Normalization, Decision Science, Clustering.*

1. Introduction

Current computer technology is running behind the data analytical tools as it has tasted the fruitful decisions taken on top of this analyzed data. Decision making is an on demand policy of any business or market problem. Before the advent of Data analytics rare are the correct decisions as the data is very messy; as a result the market has lost to the competitors who also survived with decisions that are not satisfiable. The past business decisions are made on the visible relations of the data; at the point where the decision makers are unaware of the hidden relationships beneath data. This hidden relationships of the data has completely mislead the decision path of the decision makers. The business market has sensed the actual culprits are their own decisions made on the visible relationships and are responsible for the fall of their market. The drastic changes in market and the varied demands of the customers has flagged decision making as vital in any business field. The business market in order to enhance its productivity and efficiency has realized the worth of putting in its basket the efficient decision making tools of the data science technology.

Data science is a field where the data is nourished and studied at more-finer levels, digging many useful relationships that are hidden beneath the actual data. Data science can be simply talked of as a platform for accurate decision making. Data science technology is emerging as key field for many business companies throwing its insights into many hidden flavors of the data that are not yet tasted by the business. The central theme of data science technology is accurate decision making and the approaches supporting them.

The initial startup procedure for accurate decision making as identified by data science makers is Data preprocessing. Data preprocessing is where data is cleaned before getting analyzed. Data building approaches generally end up with messy data where frequently some data is missing, data is not uniform on a data scale, irrelevant data relationships which doesn't have any impact on decision making. The AI and ML approaches uses model training before they make the decisions. If data is uncleaned knowledge discovery at training phase for accurate decision making becomes difficult and as such the model will be trained to take misleading decisions. Data cleaning approaches prepare the data by addressing the above issues on top of which accurate decision making can be done.

Data science has unveiled some vital data preprocessing techniques:

- Data cleaning.
- Data Normalization.
- Data transformation.
- Data selection.

Data cleaning is an initial approach where the data sets are cleaned to identify any missing data, remove the noisy data and get the data ready for analyzing. Data cleaning is needed to address the data quality problem. The data quality problem is where the analytics may go wrong on a messy data. Data normalization is where the data is scaled to uniformity. Data normalization is needed to study the best features of the data. Data transformation is transforming the data into types suitable for the analytical model. All the data of the dataset need not drive to useful information. Using data selection the important data features which render to useful patterns are extracted from the data. Unused features may be discarded.

This paper focuses on normalization of data before analytics; what is normalization and why normalization is needed for any analytics. The rest of the paper is organized as follows: in section 2 we discuss the literature survey, section 3 discusses briefly about normalization; section 4 discusses the analysis on normalized data and finally section 5 concludes the paper.

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

2. Literature

Normalization is a data preprocessing stage where data is scaled to needed interval. Data scaling or normalization scales the data to study the smaller insights and useful relationships. Many data preprocessing techniques emphasis on data normalization for if the data is not normalized the statistics of the data like mean, median etc.., may deviate more from the actual value.

Data normalization is the foremost step of many advanced computing technologies like soft computing, cloud computing, Data science and Machine Learning. Literature has provided with many Data normalizing techniques.

Min-Max normalization [1][2] is a linear transformation on the data to scale it to an analytical range. The method generally transforms the data to fall within the range 0 to 1. The min-max normalization technique is used when the mean and standard deviation of the data are not known. The normalization technique preserves the relationships between the original data even after the transformation.

A standard normalization technique called the Z-Score is presented in [3] [4]. The normalization is also called statistical normalization technique. This technique is used when the mean and standard deviation of the data are given. The Z-score standardization transforms data to approximate to 0 mean and unit standard deviation. Normalization by decimal scaling is discussed in [5]. The decimal scaling normalization scales the decimals depending on the absolute value of an attribute.

Discussions presented in [6] shows how the histograms with normal distribution can be used to assess data. The work showed if the data histograms are not symmetrical around the bell curve then the data is not normally distributed. There arises the need for data transformation.

The need for data normalization is discussed in [7]. Normalization scales the data to fall into uniformity and attempts to give all data attributes equal weights. Normalization is useful when we have no prior knowledge of the data. In most of the distance based clustering mechanisms one attribute value may be very large compared to other attribute values. With such differences the deviation may be very high with which the output obviously may not converge to the actual value and there comes the need for normalization. Many of the classification mechanisms of data mining restrict on the first stage of data normalization before going into the exact analytics of the data.

3. Data Normalization

The whole world is today behind the available data trying to dig the useful patterns from it. To extract such patterns from the data, the data should be initially cleaned or preprocessed. Analyzing patterns from uncleaned data will also derive patterns but such patterns may mislead the decision makers. To avoid all these inconsistencies, analysts insist on initial data preprocessing stage before data being analyzed. In data preprocessing the data is cleaned making it right for analytics stage. Normalization is very vital data preprocessing technique without which analytics are dumped into solutions arriving at inconsistencies.

Real world data is enormous in a way analyzing at a glance is quite difficult and we may miss some of the interesting relationships between the data attributes. This has led to the incorporation of many statistical approaches using which one can dig deeper into these relationships. One such approach is studying the probability distributions of the data. While working with statistical data, distributions are useful in picturing the hidden patterns of the data attributes. Most of the statistical analytics rely on two types of basic distributions: Discrete and Continuous. As real world business data is mostly assumed to be continuous the analytics follow the approaches related to continuous data. Today major emphasis is on Standard Normal distribution which a continuous distribution.

The analytical methods that are prevailing today has mandatorily made normality test as the basic data preprocessing step. Normality test on a data set checks whether the data is normally distributed or not. The normal distribution is a standard reference in observing the normal or outlier nature of the data. The normal distribution has become popular as it follows central limit theorem. The central limit theorem is based on comparing the sampling statistics with population statistics.

3.1 Standard Normal distribution

A normal distribution occurs more often in many of the real time situations. Normal distribution is used to compare various data distributions.

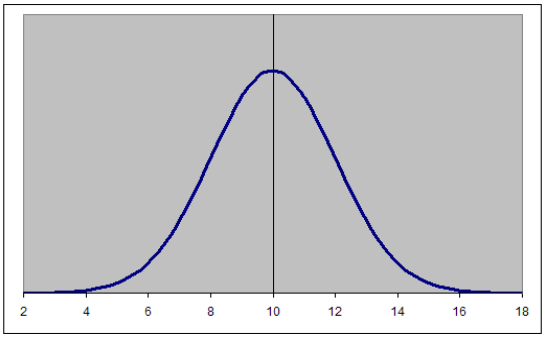


Figure 1: Standard Normal Distribution

Figure 1 shows a standard normal curve which is commonly bell shaped. Generally normal distributions are transformed to standard normal distributions to study more general characteristics of data variables. The normal distribution is standardized to zero mean and unit standard deviation and universally referred as standard normal distribution. The standard normal distribution scales the data to fall into a uniform interval [0,1]. With such standardization the proximity of throwing deeper insights into the data is more.

3.1.1 Transformations

Data transformations are techniques to increase data interpretability and data visibility. Data transformations make distributions to be more symmetric and make data to move close towards normal distribution. Data transformation is a mathematical function f, where for each data point Xi, f(Xi) transforms to a value which is more interpretable. Transformations can be used for many reasons. Most of the statistical methods apply transformations to:

- Project the data convenient for interpretation.
- To reduce data skewness.
- To produce equal spreads.
- To observe linear patterns.

General data may be scaled across larger values which may not be convenient for the analyst to have a whole good view of the data. For a convenient view of the data, transformations are applied to scale data into a smaller interval or range. Most of the statistical methods apply transformations to reduce data skewness. Skewness is where the data is not uniformly distributed around the mean. Many data inferences can be done based on the spread of the data. Data transformations are applied to produce equal spreads so that inter data relationships can be studied more properly. Statistical methods use various kinds of data transformations, projecting major ones like:

- Standard Z transformation.
- Reciprocal transformation.
- Logarithmic transformation.

A. Z Transformation

Z transformation is a transformation technique where data with different levels and spread is adjusted to a standard level and spread. The transformation is given by:

Z=X-mean(X)/ Standard deviation(X);

Where Z is the standard normal variable, X is continuous random variable. Z standardization makes the mean (levels) to be 0 and standard deviation (spread) to be 1. On standardization data is made unit less. Statistical approaches uses standardization to compare data of various units.

B. Reciprocal Transformation

Sometimes the original data may not throw useful snapshots than a reciprocal of it. Reciprocal transformations are mostly applied for positive data. For example a data analyst is analyzing the population density data so as to infer: how much living area needed for the heavy raise of population. If X is a random variable denoting the population density showing the population per certain area. X may not show any useful information at the crisis of heavy population growth. He then studies the reciprocal of X which is the amount of living area needed per population, and the reciprocal may thus throw useful insights into the population growth.

C. Logarithmic transformation

A myth followed in statistic field is many of momentary events are log normally distributed i.e the log of the data is normally distributed and taking the log values may restore the normality of the data. Log transformations are used when data is of varied and wide ranges. Log transformations are used when a multiplicative statistical model is used.

3.2. Normality and Testing for Normality

Testing for data normality is observing whether data follows standard normal distribution or not. Normality of the data infers the normal characteristics between the attributes of the data and data being not normal shows outlier characteristics of the data attributes. If the data is not normal then transformations are applied to move data closely towards normality. Analysts follows many prior normality tests which are either graphical or statistical that are based on probability distributions of the data. Here in this section we discuss the q-q plot a graphical based and the shapiro-wilk test a statistical based.

3.2.1 q-q plot

The mathematical significance of collinear points is that they lie on the same straight line with slope zero and shows positive correlations between the variables satisfying them. The q-q plot is a graphical based test method to test whether the data points lie on a straight line. If more of data points lie on the straight line then the data is normally distributed and it will be easy for the analyst to perform predictions of data characteristics. If few points lie on the straight line then we have to apply the data transformation techniques on the attributes with larger values and then apply q-q plot.

3.2.2 Shapiro-Wilk test for Normality

The shapiro wilk test is statistical. The graph based methods are wells suited for small data and large data cannot be more precisely studied using them. The statistical methods are well suited for large data. Since the statistical methods are based on probabilities where data is inferred through mathematical calculations major of the analytical methods use statistical approaches for normality check. The shapiro-wilk test is hypothetical test using the principles of hypothesis testing: the Null (H0) and the alternate hypothesis (H1). The null hypothesis assumes the data is normally distributed. The hypothesis are based on .05 levels of confidence; and if the observed probability is greater than .05, the H0 is accepted otherwise, the H1 is accepted.

4. Experimental Results

This section presents the experimental results of data with and without transformation applied. A normal bell-shaped curve is referenced while comparing these transformations. If the distributions follow bell shaped curve then the data is normally distributed. Otherwise data need to be transformed. Try to analyze the cases on the following data.

Considered Data frame
Data<-
c(1,1.3,1.1,1.2,1.2,2.5,2.3,2.6,4.1,5.0,10.0,4.0,4.1,4.2,4.1,5.1,4.5,5.0,15.3,10.0,20.0,1.1,1.2,1.6,2.2,3.0,4.0,10.5)

Experimental observations are done under two cases.

4.1.1 Case 1: Before data normalization

Here let us observe data skewness. Try to plot the normal histogram plot in R using the following command:

plotNormalHistogram (Data, xlab="Data")

The plot is shown in Figure 2.

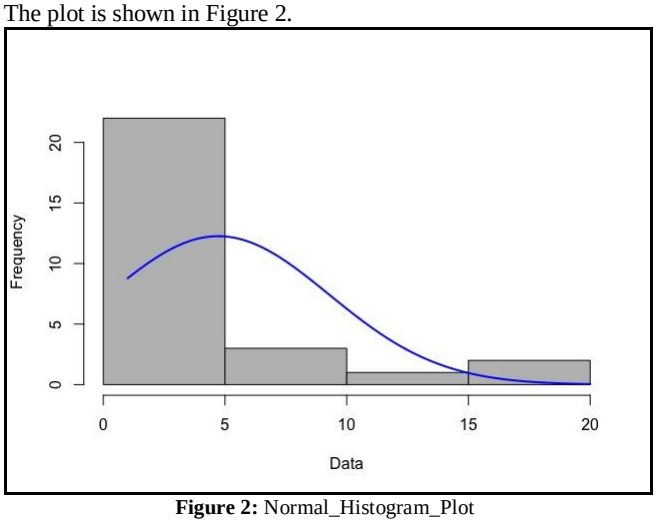


Figure 2: Normal_Histogram_Plot

The Normal histogram plot of data shown in figure 2 presents the histograms of the data which are not completely fit into the bel-shaped curve. This shows the data is skewed. Skewness shows asymmetries of data and which is not normal.

In data analytics skewed data will be identified as outliers. Clearly from the histogram of figure 2 the data points 15.3 and 20.0 fall into right skewed tail of the distribution and will be identified as outliers. This is because the distance between the min value=1 and

the max value=20 of the data is more. Also the two points are not close to the mean value=4.7 of the data. So without analyzing the data there are chances that these two points will be identified as outliers and may be removed from the data.

Now going for inter similarity cluster analysis, the mean of the Data is 4.7. Assume mean of the Data as center of single cluster, the distance between the point 15.3 and mean is 10.6; the distance between the mean and the point 20.0 is 15.3; the rest of the data points have distance <=6; Taking 6 as the clustering parameter centered at mean all the data points will fall into the defined cluster and the two points 15.5 ad 20.0 will not fall into the defined cluster as these two points fail two meet the inter cluster distance parameter. Thus these two points will be identified as outliers and have more chances to be removed without analyzing.

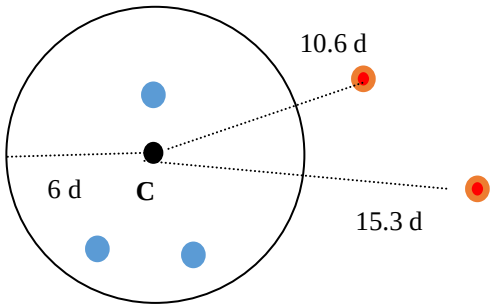


Figure 3: Clustering before transformation

Now observe the plot of the data points in R using the following command:

qqnorm(Data,ylab="sample Quantiles for Data")

qqline(Data,col="red")

The q-q plot is used to identify the linearity of the data points. If more number of points are on the q-q line then the data is normally distributed without more number of outliers.

The normal q-q plot of the data is shown in figure 4. Maximum data points are not on the normal line showing the data is not normally distributed.

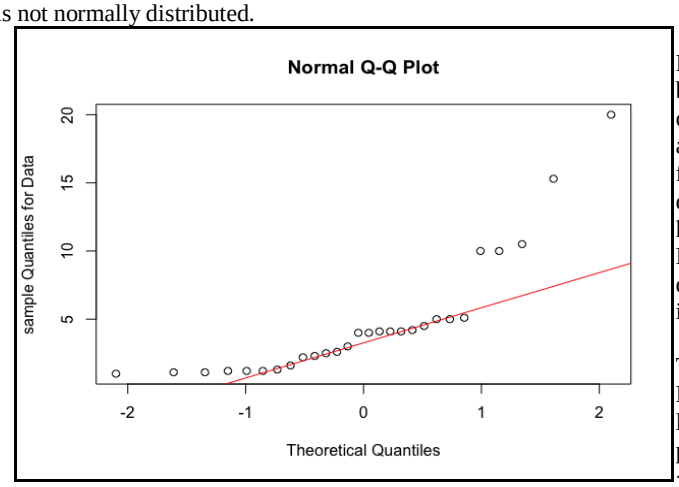


Figure 4: Normal q-q plot

4.1.2 Case 2: After data normalization

Data can be normalized by applying various transformations. Here consider the case of applying log transformation on the data.

Log Transformation:

Tran1=log(Data)

table(Tran1)

0.09531017980432
0.18232155679395
0.26236426446749
0.47000362924573
0.78845736036427
0.83290912293510
0.91629073187415
0.95551144502743
1.09861228866811
1.38629436111989
1.41098697371026
1.43508452528932
1.50407739677627
1.60943791243410
1.62924053973028
2.30258509299405
2.35137525716348
2.72785282839839
2.99573227355399

The table command in R presents the sorted values of the log(data). Now plotting the histogram of the transformed data

Library(rcompanion)

plotNormalHistogram(Tran1, xlab="log Data")

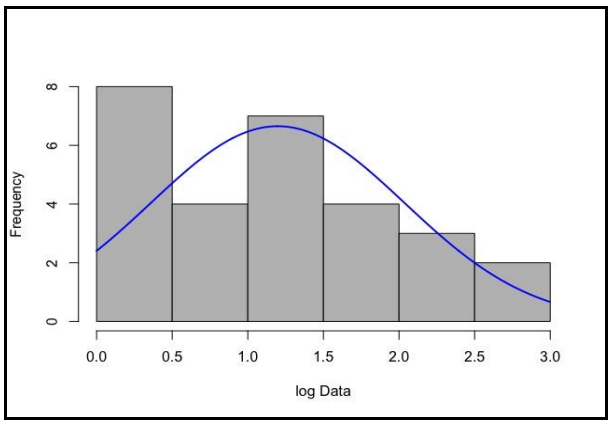


Figure 5: Log transformation of the data

From figure 5 it is clear that maximum histograms has fit into the bell-shaped curve after the data is log transformed. Now the mean of the transformed data is 1.197. The transformed values of 15.3 and 20.0 are 2.72 and 2.99 respectively. Now these two transformed values are close two the mean of the transformed data; the differences are less. Hence an analytical statement can be raised here "transformation of the data moves the data close to the mean". Figure 4 clearly shows data density at both the tails. So once the data is transformed the points 15.3 and 20.0 which are initially identified as outliers are now identified as points of information.

The mean of the transformed Data is 1.197. With mean of log Data as center of single cluster, the distance between the point log(15.3) and mean is 1.5; the distance between the mean and the point log(20.0) is 1.7; the rest of the data points have distance <=2; Taking 2 as the cluster distance parameter centered at mean all the data points will fall into the defined cluster, including the transformed points log(15.5) and log(20.0). This shows that the log transformation on the two points has made them to fall in the defined cluster, making the two points to have much inter similarity with other points. With transformation the two points 15.3 and 20.0 are made to fall in the cluster and can be analyzed.

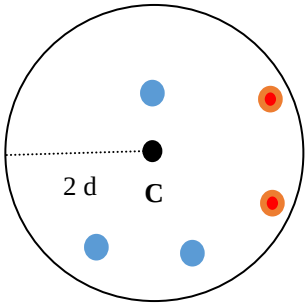


Figure 6: Clustering after Transformation

From the above illustration it is clear that data transformation makes the points informative. In general many of the analytical approaches skip the transformation stage of preprocessing for their approaches miss to study data properly. So before data analytics it is very much needed to study the data set carefully by applying possibly all the preprocessing stages. Data preprocessing is the crucial stage for accurate and correct decisions.

5. Conclusion

Data analytics is a science where the data is analyzed and studied at more-finer levels, digging many useful relationships that are hidden beneath the actual data. Data analytics can be simply talked of as a platform for accurate decision making. The central theme of data science technology is accurate decision making and the effective use of methods supporting them. The initial startup procedure for accurate decision making as identified by data analytics makers is Data preprocessing. Data preprocessing is where data is cleaned before getting analyzed. This paper focuses on data normalization before analytics; what is normalization and why normalization is needed for analytics. Normalization is very vital data preprocessing technique without which analytics are dumped into solutions arriving at inconsistencies. Experimental results shown in section 4 shows data sets should be clearly observed before analyzing the data.

Refferences

[1] SB Kotsiantis, D Kanellopoulos, PE Pintelas."Data preprocessing for supervised learning". International Journal of Computer Science. Vol 1 (2), 111-117.

[2] S Patro, KK Sahu."Normalization: A Preprocessing Stage". arXiv preprint arXiv:1503.06462

[3] L Al Shalabi, **Z** Shaaba. "Data Mining: a preprocessing Engine". Journal of Computer Science. Vol 2 (9), 735-739, 2006.

[4] H.E.Barbaree, D.J.K.Mewhort. "The effects of the z-score transformation on measures of relative erectile response strength: A reappraisal". Journal of behavior research therapy. Vol 32 (5), 547-558, June 1994.

[5] Ismail Bin Mohamad, Dauda Usman. "Standardization and Its Effects on K-Means Clustering Algorithm". Research Journal of Applied Sciences, Engineering and Technology. Vol 6 (17), 3299-3303, 2013.

[6]https://www.epa.gov/sites/production/files/2016-06/documents/normality.pdf

[7] K. Y. Yeung W. L. Ruzzo. "Principal component analysis for clustering gene expression data". Bioinformatics, Vol 17 (9), 763–7741, Sep 2001,