×

-

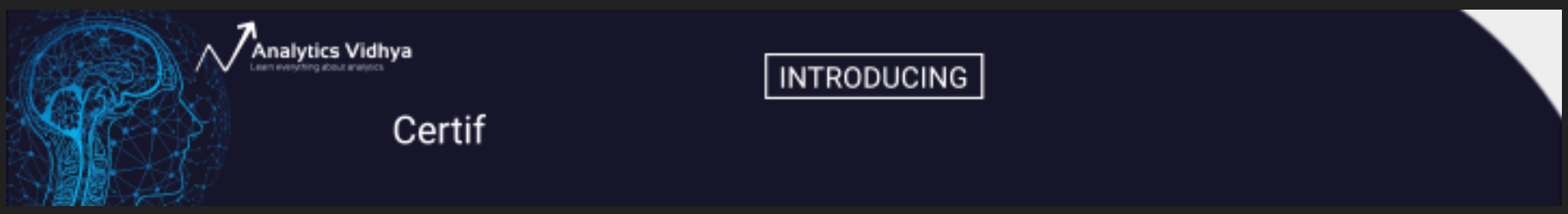
(http://play.google.com/store/apps/details?id=com.analyticsvidhya.android)

×

-

(http://play.google.com/store/apps/details?id=com.analyticsvidhya.android)

Anniversary Special Sale: Certified AI & ML BlackBelt Program (Beginner to Master) - Enroll Today @ 20% Discount | **Coupon: AV6YEARS** (https://courses.analyticsvidhya.com/bundles/ai-blackbelt-beginner-to-master?utm_source=blog&utm_medium=flashstrip)



(https://courses.analyticsvidhya.com/bundles/ai-blackbelt-beginner-to-master?utm_source=AVtopbanner&utm_medium=display)

BUSINESS ANALYTICS (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/BUSINESS-ANALYTICS/)

Reply

A Comprehensive Guide to Data Exploration

SUNIL RAY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/AUTHOR/SUNIL-RAY/), JANUARY 10, 2016 LOGIN TO BOOKMA...

Introduction

There are no shortcuts for data exploration. If you are in a state of mind, that machine learning can sail you away from every data storm, trust me, it won't. After some point of time, you'll realize that you are struggling at improving model's accuracy. In such situation, data exploration techniques will come to your rescue.

I can confidently say this, because I've been through such situations, a lot.

I have been a Business Analytics professional for close to three years now. In my initial days, one of my mentor suggested me to spend significant time on exploration and analyzing data. Following his advice has served me well.

I've created this tutorial to help you understand the underlying techniques of data exploration. As always, I've tried my best to explain these concepts in the simplest manner. For better understanding, I've taken up few examples to demonstrate the complicated concepts.

(https://www.analyticsvidhya.com/wp-content/uploads/2016/01/de.jpg)

Table of Contents

- How to detect outlier ?
- How to remove outlier ?

4. **The Art of Feature Engineering**
    - What is Feature Engineering ?
    - What is the process of Feature Engineering ?
    - What is Variable Transformation ?
    - When should we use variable transformation ?
    - What are the common methods of variable transformation ?
    - What is feature variable creation and its benefits ?

Let's get started.

1. Steps of Data Exploration and Preparation

Remember the quality of your inputs decide the quality of your output. So, once you have got your business hypothesis ready, it makes sense to spend lot of time and efforts here. With my personal estimate, data exploration, cleaning and preparation can take up to 70% of your total project time.

Below are the steps involved to understand, clean and prepare your data for building your predictive model:

1. Variable Identification
2. Univariate Analysis
3. Bi-variate Analysis
4. Missing values treatment
5. Outlier treatment
6. Variable transformation
7. Variable creation

Finally, we will need to iterate over steps 4 – 7 multiple times before we come up with our refined model.

Let's now study each stage in detail:-

Variable Identification

First, identify **Predictor** (Input) and **Target** (output) variables. Next, identify the data type and category of the variables.

Let's understand this step more clearly by taking an example.

Example:- Suppose, we want to predict, whether the students will play cricket or not (refer below data set). Here you need to identify predictor variables, target variable, data type of variables and category of variables.

(https://www.analyticsvidhya.com/wp-content/uploads/2015/02/Data_exploration_11.png)Below, the variables have been defined in different category:

(https://www.analyticsvidhya.com/wp-content/uploads/2015/02/Data_exploration_2.png)

Univariate Analysis

At this stage, we explore variables one by one. Method to perform uni-variate analysis will depend on whether the variable type is categorical or continuous. Let's look at these methods and statistical measures for categorical and continuous variables individually:

Continuous Variables:- In case of continuous variables, we need to understand the central tendency and spread of the variable. These are measured using various statistical metrics visualization methods as shown below:

(https://www.analyticsvidhya.com/wp-content/uploads/2015/02/Data_exploration_31.png)**Note:** Univariate analysis is also used to highlight missing and outlier values. In the upcoming part of this series, we will look at methods to handle missing and outlier values. To know more about these methods, you can refer course descriptive statistics from Udacity (https://www.udacity.com/course/ud827).

Categorical Variables:- For categorical variables, we'll use frequency table to understand distribution of each category. We can also read as percentage of values under each category. It can be be measured using two metrics, **Count** and **Count%** against each category. Bar chart can be used as visualization.

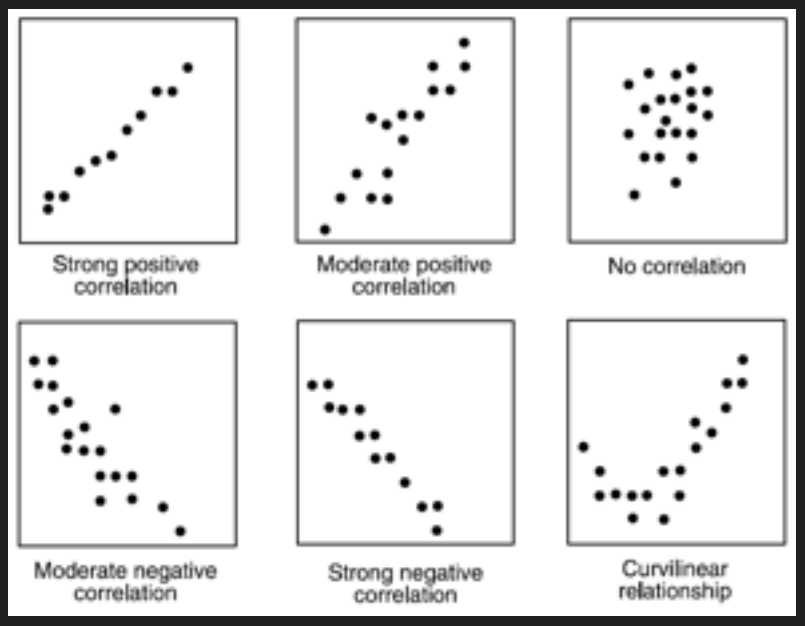
Bi-variate Analysis

Bi-variate Analysis finds out the relationship between two variables. Here, we look for association and disassociation between variables at a pre-defined significance level. We can perform bi-variate analysis for any combination of categorical and continuous variables. The combination can be:

Categorical & Categorical, Categorical & Continuous and Continuous & Continuous. Different methods are used to tackle these combinations during analysis process.

Let's understand the possible combinations in detail:

Continuous & Continuous: While doing bi-variate analysis between two continuous variables, we should look at scatter plot. It is a nifty way to find out the relationship between two variables. The pattern of scatter plot indicates the relationship between variables. The relationship can be linear or non-linear.



(https://www.analyticsvidhya.com/wp-content/uploads/2015/02/Data_exploration_4.png)Scatter plot shows the relationship between two variable but does not indicates the strength of relationship amongst them. To find the strength of the relationship, we use Correlation. Correlation varies between -1 and +1.

- -1: perfect negative linear correlation
- +1:perfect positive linear correlation and (https://www.analyticsvidhya.com/wp-content/uploads/2015/02/Data_exploration_4.png)
- 0: No correlation

Correlation can be derived using following formula:

Correlation = Covariance(X,Y) / SQRT(Var(X)* Var(Y))

Various tools have function or functionality to identify correlation between variables. In Excel, function CORREL() is used to return the correlation between two variables and SAS uses procedure PROC CORR to identify the correlation. These function returns Pearson Correlation value to identify the relationship between two variables:

(https://www.analyticsvidhya.com/wp-content/uploads/2015/02/Data_exploration_51.png)

In above example, we have good positive relationship(0.65) between two variables X and Y.

Categorical & Categorical: To find the relationship between two categorical variables, we can use following methods:

- **Two-way table:** We can start analyzing the relationship by creating a two-way table of count and count%. The rows represents the category of one variable and the columns represent the categories of the other variable. We show count or count% of observations available in each combination of row and column categories.
- **Stacked Column Chart:** This method is more of a visual form of Two-way table.

(https://www.analyticsvidhya.com/wp-content/uploads/2015/02/Data_exploration_6.gif)

- **Chi-Square Test:** This test is used to derive the statistical significance of relationship between the variables. Also, it tests whether the evidence in the sample is strong enough to generalize that the relationship for a larger population as well. Chi-square is based on the difference between the expected and observed frequencies in one or more categories in the two-way table. It returns probability for the computed chi-square distribution with the degree of freedom.

Probability of 0: It indicates that both categorical variable are dependent

Probability of 1: It shows that both variables are independent.

Probability less than 0.05: It indicates that the relationship between the variables is significant at 95% confidence. The chi-square test statistic for a test of independence of two categorical variables is found by:

$$X^2 = \sum (O - E)^2 / E$$

(https://www.analyticsvidhya.com/wp-content/uploads/2015/02/Data_exploration_7.png)where *O* represents the observed frequency. *E* is the expected frequency under the null hypothesis and computed by:

$$E = \frac{row\ total \times column\ total}{sample\ size}$$

(https://www.analyticsvidhya.com/wp-content/uploads/2015/02/Data_exploration_8.png)
From previous two-way table, the expected count for product category 1 to be of small size is 0.22. It is derived by taking the row total for Size (9) times the column total for Product category (2) then dividing by the sample size (81). This is procedure is conducted for each cell. Statistical Measures used to analyze the power of relationship are:

- Cramer's V for Nominal Categorical Variable
- Mantel-Haenszed Chi-Square for ordinal categorical variable.

Different data science language and tools have specific methods to perform chi-square test. In SAS, we can use **Chisq** as an option with **Proc freq** to perform this test.

Categorical & Continuous: While exploring relation between categorical and continuous variables, we can draw box plots for each level of categorical variables. If levels are small in number, it will not show the statistical significance. To look at the statistical significance we can perform Z-test, T-test or ANOVA.

- **Z-Test/ T-Test:-** Either test assess whether mean of two groups are statistically different from each other or not.

$$z = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

(https://www.analyticsvidhya.com/wp-content/uploads/2015/02/ztestformula1.jpg)If the probability of Z is small then the difference of two averages is more significant. The T-test is very similar to Z-test but it is used when number of observation for both categories is less than 30.

(https://www.analyticsvidhya.com/wp-content/uploads/2015/02/ttest.png)
- **ANOVA:-** It assesses whether the average of more than two groups is statistically different.

Example: Suppose, we want to test the effect of five different exercises. For this, we recruit 20 men and assign one type of exercise to 4 men (5 groups). Their weights are recorded after a few weeks. We need to find out whether the effect of these exercises on them is significantly different or not. This can be done by comparing the weights of the 5 groups of 4 men each.

Till here, we have understood the first three stages of Data Exploration, Variable Identification, Uni-Variate and Bi-Variate analysis. We also looked at various statistical and visual methods to identify the relationship between variables.

Now, we will look at the methods of Missing values Treatment. More importantly, we will also look at why missing values occur in our data and why treating them is necessary.

2. Missing Value Treatment

Why missing values treatment is required?

Missing data in the training data set can reduce the power / fit of a model or can lead to a biased model because we have not analysed the behavior and relationship with other variables correctly. It can lead to wrong prediction or classification.

(https://www.analyticsvidhya.com/wp-content/uploads/2015/02/Data_Exploration_2_11.png)

Notice the missing values in the image shown above: In the left scenario, we have not treated missing values. The inference from this data set is that the chances of playing cricket by males is higher than females. On the other hand, if you look at the second table, which shows data after treatment of missing values (based on gender), we can see that females have higher chances of playing cricket compared to males.

Why my data has missing values?

We looked at the importance of treatment of missing values in a dataset. Now, let's identify the reasons for occurrence of these missing values. They may occur at two stages:

1. **Data Extraction**: It is possible that there are problems with extraction process. In such cases, we should double-check for correct data with data guardians. Some hashing procedures can also be used to make sure data extraction is correct. Errors at data extraction stage are typically easy to find and can be corrected easily as well.
2. **Data collection**: These errors occur at time of data collection and are harder to correct. They can be categorized in four types:
    - **Missing completely at random:** This is a case when the probability of missing variable is same for all observations. For example: respondents of data collection process decide that they will declare their earning after tossing a fair coin. If an head occurs, respondent declares his / her earnings & vice versa. Here each observation has equal chance of missing value.
    - **Missing at random:** This is a case when variable is missing at random and missing ratio varies for different values / level of other input variables. For example: We are collecting data for age and female has higher missing value compare to male.
    - **Missing that depends on unobserved predictors:** This is a case when the missing values are not random and are related to the unobserved input variable. For example: In a medical study, if a particular diagnostic causes discomfort, then there is higher chance of drop out from the study. This missing value is not at random unless we have included "discomfort" as an input variable for all patients.
    - **Missing that depends on the missing value itself:** This is a case when the probability of missing value is directly correlated with missing value itself. For example: People with higher or lower income are likely to provide non-response to their earning.

Which are the methods to treat missing values ?

1. **Deletion:** It is of two types: List Wise Deletion and Pair Wise Deletion.
    - In list wise deletion, we delete observations where any of the variable is missing. Simplicity is one of the major advantage of this method, but this method reduces the power of model because it reduces the sample size.
    - In pair wise deletion, we perform analysis with all cases in which the variables of interest are present. Advantage of this method is, it keeps as many cases available for analysis. One of the disadvantage of this method, it uses different sample size for different variables.

      (https://www.analyticsvidhya.com/wp-content/uploads/2015/02/Data_Exploration_2_2.png)

      (https://www.analyticsvidhya.com/wp-content/uploads/2015/02/Data_Exploration_2_2.png)
    - Deletion methods are used when the nature of missing data is "**Missing completely at random**" else non random missing values can bias the model output.

2. **Mean/ Mode/ Median Imputation:** Imputation is a method to fill in the missing values with estimated ones. The objective is to employ known relationships that can be identified in the valid values of the data set to assist in estimating the missing values. Mean / Mode / Median imputation is one of the most frequently used methods. It consists of replacing the missing data for a given attribute by the mean or median (quantitative attribute) or mode (qualitative attribute) of all known values of that variable. It can be of two types:-
    - **Generalized Imputation:** In this case, we calculate the mean or median for all non missing values of that variable then replace missing value with mean or median. Like in above table, variable "**Manpower"** is missing so we take average of all non missing values of "**Manpower"** (**28.33**) and then replace missing value with it.
    - **Similar case Imputation:** In this case, we calculate average for gender "**Male"** (29.75) and "**Female**" (25) individually of non missing values then replace the missing value based on gender. For "**Male**", we will replace missing values of manpower with 29.75 and for "**Female**" with 25.

3. **Prediction Model**: Prediction model is one of the sophisticated method for handling missing data. Here, we create a predictive model to estimate values that will substitute the missing data. In this case, we divide our data set into two sets: One set with no missing values for the variable and another one with missing values. First data set become training data set of the model while second data set with missing values is test data set and variable with missing values is treated as target variable. Next, we create a model to predict target variable based on other attributes of the training data set and populate missing values of test data set.We can use regression, ANOVA, Logistic regression and various modeling technique to perform this. There are 2 drawbacks for this approach:

1. The model estimated values are usually more well-behaved than the true values
2. If there are no relationships with attributes in the data set and the attribute with missing values, then the model will not be precise for estimating missing values.

4. **KNN Imputation:** In this method of imputation, the missing values of an attribute are imputed using the given number of attributes that are most similar to the attribute whose values are missing. The similarity of two attributes is determined using a distance function. It is also known to have certain advantage & disadvantages.
    - **Advantages:**
        - k-nearest neighbour can predict both qualitative & quantitative attributes
        - Creation of predictive model for each attribute with missing data is not required
        - Attributes with multiple missing values can be easily treated
        - Correlation structure of the data is taken into consideration

    - **Disadvantage:**
        - KNN algorithm is very time-consuming in analyzing large database. It searches through all the dataset looking for the most similar instances.
        - Choice of k-value is very critical. Higher value of k would include attributes which are significantly different from what we need whereas lower value of k implies missing out of significant attributes.

After dealing with missing values, the next task is to deal with outliers. Often, we tend to neglect outliers while building models. This is a discouraging practice. Outliers tend to make your data skewed and reduces accuracy. Let's learn more about outlier treatment.

3. Techniques of Outlier Detection and Treatment

What is an Outlier?

Outlier is a commonly used terminology by analysts and data scientists as it needs close attention else it can result in wildly wrong estimations. Simply speaking, Outlier is an observation that appears far away and diverges from an overall pattern in a sample.

Let's take an example, we do customer profiling and find out that the average annual income of customers is $0.8 million. But, there are two customers having annual income of $4 and $4.2 million. These two customers annual income is much higher than rest of the population. These two observations will be seen as Outliers.

(https://www.analyticsvidhya.com/wp-content/uploads/2015/02/Outlier.png)

What are the types of Outliers?

Outlier can be of two types: **Univariate** and **Multivariate**. Above, we have discussed the example of univariate outlier. These outliers can be found when we look at distribution of a single variable. Multivariate outliers are outliers in an n-dimensional space. In order to find them, you have to look at distributions in multi-dimensions.

Let us understand this with an example. Let us say we are understanding the relationship between height and weight. Below, we have univariate and bivariate distribution for Height, Weight. Take a look at the box plot. We do not have any outlier (above and below 1.5*IQR, most common method). Now look at the scatter plot. Here, we have two values below and one above the average in a specific segment of weight and height.

(https://www.analyticsvidhya.com/wp-content/uploads/2015/02/Outlier_21.png)

What causes Outliers?

Whenever we come across outliers, the ideal way to tackle them is to find out the reason of having these outliers. The method to deal with them would then depend on the reason of their occurrence. Causes of outliers can be classified in two broad categories:

1. **Artificial (Error) / Non-natural**
2. **Natural**.

Let's understand various types of outliers in more detail:

- **Data Entry Errors:-** Human errors such as errors caused during data collection, recording, or entry can cause outliers in data. For example: Annual income of a customer is $100,000. Accidentally, the data entry operator puts an additional zero in the figure. Now the income becomes $1,000,000 which is 10 times higher. Evidently, this will be the outlier value when compared with rest of the population.
- **Measurement Error:** It is the most common source of outliers. This is caused when the measurement instrument used turns out to be faulty. For example: There are 10 weighing machines. 9 of them are correct, 1 is faulty. Weight measured by people on the faulty machine will be higher / lower than the rest of people in the group. The weights measured on faulty machine can lead to outliers.
- **Experimental Error:** Another cause of outliers is experimental error. For example: In a 100m sprint of 7 runners, one runner missed out on concentrating on the 'Go' call which caused him to start late. Hence, this caused the runner's run time to be more than other runners. His total run time can be an outlier.
- **Intentional Outlier*:*** This is commonly found in self-reported measures that involves sensitive data. For example: Teens would typically under report the amount of alcohol that they

consume. Only a fraction of them would report actual value. Here actual values might look like outliers because rest of the teens are under reporting the consumption.

- **Data Processing Error:** Whenever we perform data mining, we extract data from multiple sources. It is possible that some manipulation or extraction errors may lead to outliers in the dataset.
- **Sampling error:** For instance, we have to measure the height of athletes. By mistake, we include a few basketball players in the sample. This inclusion is likely to cause outliers in the dataset.
- **Natural Outlier:** When an outlier is not artificial (due to error), it is a natural outlier. For instance: In my last assignment with one of the renowned insurance company, I noticed that the performance of top 50 financial advisors was far higher than rest of the population. Surprisingly, it was not due to any error. Hence, whenever we perform any data mining activity with advisors, we used to treat this segment separately.

What is the impact of Outliers on a dataset?

Outliers can drastically change the results of the data analysis and statistical modeling. There are numerous unfavourable impacts of outliers in the data set:

- It increases the error variance and reduces the power of statistical tests
- If the outliers are non-randomly distributed, they can decrease normality
- They can bias or influence estimates that may be of substantive interest
- They can also impact the basic assumption of Regression, ANOVA and other statistical model assumptions.

To understand the impact deeply, let's take an example to check what happens to a data set with and without outliers in the data set.

Example:

(https://www.analyticsvidhya.com/wp-content/uploads/2015/02/Outlier_31.png)

As you can see, data set with outliers has significantly different mean and standard deviation. In the first scenario, we will say that average is 5.45. But with the outlier, average soars to 30. This would change the estimate completely.

How to detect Outliers?

Most commonly used method to detect outliers is visualization. We use various visualization methods, like **Box-plot**, **Histogram**, **Scatter Plot** (above, we have used box plot and scatter plot for visualization). Some analysts also various thumb rules to detect outliers. Some of them are:

- Any value, which is beyond the range of -1.5 x IQR to 1.5 x IQR
- Use capping methods. Any value which out of range of 5th and 95th percentile can be considered as outlier
- Data points, three or more standard deviation away from mean are considered outlier
- Outlier detection is merely a special case of the examination of data for influential data points and it also depends on the business understanding
- Bivariate and multivariate outliers are typically measured using either an index of influence or leverage, or distance. Popular indices such as Mahalanobis' distance and Cook's *D* are frequently used to detect outliers.
- In SAS, we can use PROC Univariate, PROC SGPLOT. To identify outliers and influential observation, we also look at statistical measure like STUDENT, COOKD, RSTUDENT and others.

How to remove Outliers?

Most of the ways to deal with outliers are similar to the methods of missing values like deleting observations, transforming them, binning them, treat them as a separate group, imputing values and other statistical methods. Here, we will discuss the common techniques used to deal with outliers:

Deleting observations: We delete outlier values if it is due to data entry error, data processing error or outlier observations are very small in numbers. We can also use trimming at both ends to remove outliers.

Transforming and binning values: Transforming variables can also eliminate outliers. Natural log of a value reduces the variation caused by extreme values. Binning is also a form of variable transformation. Decision Tree algorithm allows to deal with outliers well due to binning of variable. We can also use the process of assigning weights to different observations.

(https://www.analyticsvidhya.com/wp-content/uploads/2015/02/Transformation_1.png)

Imputing: Like imputation of missing values (https://www.analyticsvidhya.com/blog/2015/02/7-steps-data-exploration-preparation-building-model-part-2/), we can also impute outliers. We can use mean, median, mode imputation methods. Before imputing values, we should analyse if it is natural outlier or artificial. If it is artificial, we can go with imputing values. We can also use statistical model to predict values of outlier observation and after that we can impute it with predicted values.

Treat separately: If there are significant number of outliers, we should treat them separately in the statistical model. One of the approach is to treat both groups as two different groups and build individual model for both groups and then combine the output.

Till here, we have learnt about steps of data exploration, missing value treatment and techniques of outlier detection and treatment. These 3 stages will make your raw data better in terms of

information availability and accuracy. Let's now proceed to the final stage of data exploration. It is Feature Engineering.

4. The Art of Feature Engineering

What is Feature Engineering?

Feature engineering is the science (and art) of extracting more information from existing data. You are not adding any new data here, but you are actually making the data you already have more useful.

For example, let's say you are trying to predict foot fall in a shopping mall based on dates. If you try and use the dates directly, you may not be able to extract meaningful insights from the data. This is because the foot fall is less affected by the day of the month than it is by the day of the week. Now this information about day of week is implicit in your data. You need to bring it out to make your model better.

This exercising of bringing out information from data in known as feature engineering.

What is the process of Feature Engineering ?

You perform feature engineering once you have completed the first 5 steps in data exploration – Variable Identification, Univariate, Bivariate Analysis (https://www.analyticsvidhya.com/blog/2015/02/data-exploration-preparation-model/), Missing Values Imputation (https://www.analyticsvidhya.com/blog/2015/02/7-steps-data-exploration-preparation-building-model-part-2/) and Outliers Treatment (https://www.analyticsvidhya.com/blog/2015/02/outliers-detection-treatment-dataset/). Feature engineering itself can be divided in 2 steps:

- Variable transformation.
- Variable / Feature creation.

These two techniques are vital in data exploration and have a remarkable impact on the power of prediction. Let's understand each of this step in more details.

What is Variable Transformation?

In data modelling, transformation refers to the replacement of a variable by a function. For instance, replacing a variable x by the square / cube root or logarithm x is a transformation. In other words, transformation is a process that changes the distribution or relationship of a variable with others.

Let's look at the situations when variable transformation is useful.

When should we use Variable Transformation?

Below are the situations where variable transformation is a requisite:

- When we want to **change the scale** of a variable or standardize the values of a variable for better understanding. While this transformation is a must if you have data in different scales, this transformation does not change the shape of the variable distribution

- When we can **transform complex non-linear relationships into linear relationships**. Existence of a linear relationship between variables is easier to comprehend compared to a non-linear or curved relation. Transformation helps us to convert a non-linear relation into linear relation. Scatter plot can be used to find the relationship between two continuous variables. These transformations also improve the prediction. Log transformation is one of the commonly used transformation technique used in these situations.

- (https://www.analyticsvidhya.com/wp-content/uploads/2015/03/Relation.png)**Symmetric distribution is preferred over skewed distribution** as it is easier to interpret and generate inferences. Some modeling techniques requires normal distribution of variables. So, whenever we have a skewed distribution, we can use transformations which reduce skewness. For right skewed distribution, we take square / cube root or logarithm of variable and for left skewed, we take square / cube or exponential of variables.

  (https://www.analyticsvidhya.com/wp-content/uploads/2015/03/Transformation_1.png)

- Variable Transformation is also done from an **implementation point of view** (Human involvement). Let's understand it more clearly. In one of my project on employee performance, I found that age has direct correlation with performance of the employee i.e. higher the age, better the performance. From an implementation stand point, launching age based progamme might present implementation challenge. However, categorizing the sales agents in three age group buckets of <30 years, 30-45 years and >45 and then formulating three different strategies for each group is a judicious approach. This categorization technique is known as Binning of Variables.

What are the common methods of Variable Transformation?

There are various methods used to transform variables. As discussed, some of them include square root, cube root, logarithmic, binning, reciprocal and many others. Let's look at these methods in detail by highlighting the pros and cons of these transformation methods.

- **Logarithm:** Log of a variable is a common transformation method used to change the shape of distribution of the variable on a distribution plot. It is generally used for reducing right skewness of variables. Though, It can't be applied to zero or negative values as well.

- **Square / Cube root:** The square and cube root of a variable has a sound effect on variable distribution. However, it is not as significant as logarithmic transformation. Cube root has its own advantage. It can be applied to negative values including zero. Square root can be applied to positive values including zero.

- **Binning:** It is used to categorize variables. It is performed on original values, percentile or frequency. Decision of categorization technique is based on business understanding. For example, we can categorize income in three categories, namely: High, Average and Low. We can also perform co-variate binning which depends on the value of more than one variables.

What is Feature / Variable Creation & its Benefits?

Feature / Variable creation is a process to generate a new variables / features based on existing variable(s). For example, say, we have date(dd-mm-yy) as an input variable in a data set. We can generate new variables like day, month, year, week, weekday that may have better relationship with target variable. This step is used to highlight the hidden relationship in a variable:

There are various techniques to create new features. Let's look at the some of the commonly used methods:

- **Creating derived variables:** This refers to creating new variables from existing variable(s) using set of functions or different methods. Let's look at it through "**Titanic – Kaggle competition** (https://www.kaggle.com/c/titanic-gettingStarted/data)". In this data set, variable age has missing values. To predict missing values, we used the salutation (Master, Mr, Miss, Mrs) of name as a new variable. How do we decide which variable to create? Honestly, this depends on business understanding of the analyst, his curiosity and the set of hypothesis he might have about the problem. Methods such as taking log of variables, binning variables and other methods of variable transformation can also be used to create new variables.

- **Creating dummy variables:** One of the most common application of dummy variable is to convert categorical variable into numerical variables. Dummy variables are also called Indicator Variables. It is useful to take categorical variable as a predictor in statistical models. Categorical variable can take values 0 and 1. Let's take a variable 'gender'. We can produce

two variables, namely, "**Var_Male**" with values 1 (Male) and 0 (No male) and "**Var_Female**" with values 1 (Female) and 0 (No Female). We can also create dummy variables for more than two classes of a categorical variables with n or n-1 dummy variables.

Emp_Code	Gender	Var_Male	Var_Female
A001	Male	1	0
A002	Female	0	1
A003	Female	0	1
A004	Male	1	0
A005	Female	0	1
A006	Male	1	0
A007	Male	1	0

(https://www.analyticsvidhya.com/wp-content/uploads/2015/03/Dummy.png)

For further read, here is a list of transformation / creation ideas (https://www.analyticsvidhya.com/blog/2013/11/simple-manipulations-extract-data/) which can be applied to your data.

End Notes

As mentioned in the beginning, quality and efforts invested in data exploration differentiates a good model from a bad model.

This ends our guide on data exploration and preparation. In this comprehensive guide, we looked at the seven steps of data exploration in detail. The aim of this series was to provide an in depth and step by step guide to an extremely important process in data science.

Personally, I enjoyed writing this guide and would love to learn from your feedback. Did you find this guide useful? I would appreciate your suggestions/feedback. Please feel free to ask your questions through comments below.

If you like what you just read & want to continue your analytics learning, subscribe to our emails (http://feedburner.google.com/fb/a/mailverify?uri=analyticsvidhya), follow us on twitter (http://twitter.com/analyticsvidhya) or like our facebook page (http://facebook.com/analyticsvidhya).

You can also read this article on Analytics Vidhya's Android APP 

(//play.google.com/store/apps/details?id=com.analyticsvidhya.android&utm_source=blog_article&utm_campaign=blog&pcampaignid=MKT-Other-global-all-co-prtnr-py-PartBadge-Mar2515-1)

Share this:

(https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/?share=linkedin&nb=1&nb=1)

(https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/?share=facebook&nb=1&nb=1)

(https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/?share=twitter&nb=1&nb=1)

(https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/?share=pocket&nb=1&nb=1)

(https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/?share=reddit&nb=1&nb=1)

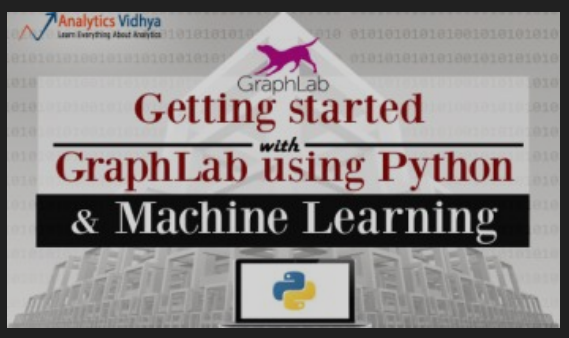
Like this:

★ Like

Be the first to like this.

Related Articles

(https://www.analyticsvidhya.com/blog/2019/03/11-steps-data-science-role-reporting-mis-bi-professionals/)
11 Steps to Transition into Data Science (for Reporting / MIS / BI Professionals)
(https://www.analyticsvidhya.com/blog/2019/03/11-steps-data-science-role-reporting-mis-bi-professionals/)
March 6, 2019
In "Career"



(http://www.analyticsvidhya.com/blog/2015/12/started-graphlab-python/)
Tutorial - Getting Started with GraphLab For Machine Learning in Python
(https://www.analyticsvidhya.com/blog/2015/12/started-graphlab-python/)
December 3, 2015
In "Machine Learning"

(http://www.analyticsvidhya.com/blog/2015/12/kaggle-solution-cooking-text-mining-competition/)
Kaggle Solution: What's Cooking ? (Text Mining Competition)
(https://www.analyticsvidhya.com/blog/2015/12/kaggle-solution-cooking-text-mining-competition/)
December 23, 2015
In "Machine Learning"

TAGS : BIVARIATE ANALYSIS (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/BIVARIATE-ANALYSIS/), DATA EXPLORATION (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/DATA-EXPLORATION/), DUMMY VARIABLES (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/DUMMY-VARIABLES/), FEATURE ENGINEERING (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/FEATURE-ENGINEERING/), KNN IMPUTATION (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/KNN-IMPUTATION/), MEDIAN IMPUTATION (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/MEDIAN-IMPUTATION/), MISSING VALUE (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/MISSING-VALUE/), MISSING VALUE IMPUTATION (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/MISSING-VALUE-IMPUTATION/), ONE HOT ENCODING (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/ONE-HOT-ENCODING/), OUTLIER REMOVAL (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/OUTLIER-REMOVAL/), UNIVARIATE ANALYSIS (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/UNIVARIATE-ANALYSIS/)

NEXT ARTICLE

20 Powerful Images which perfectly captures the growth of Data Science

(https://www.analyticsvidhya.com/blog/2016/01/20-powerful-images-perfectly-captures-growth-data-science/)

•••

PREVIOUS ARTICLE

The Ultimate Plan to Become a Data Scientist in 2016

(https://www.analyticsvidhya.com/blog/2016/01/ultimate-plan-data-scientist-2016/)

(https://www.analyticsvidhya.com/blog/author/sunil-ray/)

Sunil Ray (Https://Www.analyticsvidhya.com/Blog/Author/Sunil-Ray/)

I am a Business Analytics and Intelligence professional with deep experience in the Indian Insurance industry. I have worked for various multi-national Insurance companies in last 7 years.

This article is quite old and you might not get a prompt response from the author. We request you to post this comment on Analytics Vidhya's **Discussion portal** (https://discuss.analyticsvidhya.com/) to get your queries resolved

102 COMMENTS

DR.D.K.SAMUEL Reply

January 11, 2016 at 4:26 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-103484)

Really useful and comprehensive, thanks

BAGUINEBIE BAZONGO Reply

January 11, 2016 at 5:52 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-103489)

Hi Ray,
I would like to thank you very much for this useful post
I took more than 30 statistical courses but your post has summarized them for me
Now all things are clear about EDA
I'm member of the John Hopkins University Data Scientists (Coursera) Group
Best,

NANDU KULKARNI Reply

January 11, 2016 at 6:45 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-103490)

Excellent series of blog posts. Thanks and keep up the good work!

ABHISHEK ANAND Reply

January 11, 2016 at 7:05 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-103491)

Superb writing, crisp and comprehensive. Certainly a good refresher. Keep writing!

SHARON Reply

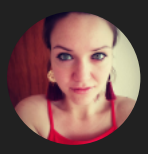
January 11, 2016 at 8:00 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-103494)

Very comprehensive. Thanks

KARTHIKEYAN SANKARAN (HTTP://WWW.TWITTER.COM/KARTHIKONBI) Reply

January 11, 2016 at 9:20 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-103510)

Excellent article on the most important aspects of Machine Learning. The points are explained in a simple and concise manner. Thank you.



ANABELLE Reply

January 11, 2016 at 9:41 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-103511)

Thank you very much for this tutorial!

SATISH Reply

January 11, 2016 at 2:15 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-103525)

I haven't come across any other article as detailed as this one. Anyone who is keen about data exploration and Predictive Analytics in general has to go through this. Wondering if you have any data set where in I can work on it.
Bookmarked!

KHALID RIAZ Reply

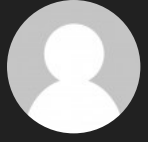
January 11, 2016 at 3:09 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-103528)

Hi Ray,
This is a great post. You have treated a fairly vast topic with just the right amount of detail. This makes it very useful, and also very intresting. Thank you for the good work. Keep it up.

ATTILA SCHMID Reply

January 11, 2016 at 3:26 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-103529)

Thank you so much for this very valuable post. I like your blogs, Please continue your good work !



NANDEESH Reply

January 11, 2016 at 6:01 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-103539)

I would like to thank Mr. Sunil Ray for such comprehensive information. Also, I would request some to write a blog on ETL, SAS BI and how SAS BI is better than other BI tools like Tableau, Qlikview….gaining more popularity in market.
Thanks again for sharing helpful information!!

MOHAMED Reply

January 11, 2016 at 10:31 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-103548)

Well defined process of data exploration Sunil. I appreciated if you continue this wonderful work and post an example of data analysis step by step using Python.

Thanks

JOHNPAULINEPINEDA Reply

January 12, 2016 at 1:11 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-103553)

Thank you Mr. Ray for the very comprehensive discussion on data exploration. I specially liked how you emphasized on the importance of EDA with this statement “quality and efforts invested in data exploration differentiates a good model from a bad model”. Great work Sir! I wish you can tackle dimensionality reduction techniques, principal components analysis, discriminant analysis and the likes in the future. Thanks again Mr. Ray.

SANDRA (HTTP://VWFXPAOAXM.COM) Reply

February 9, 2016 at 3:08 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-105436)

I found myself nodding my noggin all the way thguorh.

DEBASHIS ROUT Reply

January 12, 2016 at 3:56 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-103557)

Its really worth to read. Very comprehensive and easy to understand . I will be happy to read your article using R on data exploration & Data preparation.

SUNIL RAY Reply

January 12, 2016 at 5:03 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-103559)

Thank you all for exciting comments and I'm glad it helped.

Regards,
Sunil

DARIO ROMERO Reply

January 12, 2016 at 5:04 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-103560)

Hi Sunil: Thanks for your article on such an important topic. BTW, there is a missing graph on the paragraph Continuous & Continuous under Bi-Variate Analysis. Could you please edit it and add the missing graph. I think is pointing to a wrong place looking for the ping file. Thanks.

DARIO (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/01/GUIDE-DATA-EXPLORATION/) Reply

January 12, 2016 at 6:01 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-103561)

The missing picture/draw might be located here:
http://www.analyticsvidhya.com/wp-content/uploads/2015/02/Data_exploration_4.png (http://www.analyticsvidhya.com/wp-content/uploads/2015/02/Data_exploration_4.png)
This picture is the missing one below the paragraph:
"Continuous & Continuous: While doing bi-variate analysis between two continuous variables, we should look at scatter plot. It is a nifty way to find out the relationship between two variables. The pattern of scatter plot indicates the relationship between variables. The relationship can be linear or non-linear."

AKSHAY KHER (HTTPS://AKSHAYKHER.WORDPRESS.COM/) Reply

January 12, 2016 at 9:44 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-103573)

Hi Sunil,

An intriguing article, I can see the amount of hard work you must have put into it. Its a must read.

Thanks,
Akshay Kher

HIGHSPIRITS Reply

January 12, 2016 at 1:35 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-103588)

Thanks a lot for the comprehensive material Sunil. I had All these points scattered across but you got all of them together, along with few new pointers. Bookmarked this page and this would now be my first page to refer for any data analysis project.

SOMNATH Reply

January 13, 2016 at 6:11 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-103633)

Clear explanation with example and graph.
Thanks.

ARUN Reply

January 16, 2016 at 5:18 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-103871)

This common techniques are core of any data analytics project. Good work keep up.

TAPASVI SONI Reply

January 21, 2016 at 11:24 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-104271)

Very well explained and interesting article..It helped me a lot….Thanks a lot

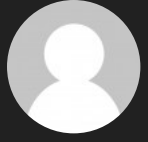
---

AZIM Reply

January 26, 2016 at 12:06 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-104638)

when we create new variable like var_male and var_female we assign 0,1 to them? how is this 0,1 is used in our model? can we assign 200 instead of 0 and 2000 instead of 1?

Please help .



BRAJENDRA GOUDA Reply

February 4, 2016 at 6:43 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-105186)

clear, Concise and Very well explained. !!

SUHEL Reply

February 13, 2016 at 6:57 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-105638)

Great article.

One quick suggestion regarding log transform for zero or negative values.

For all values, convert to absolute value, add one to all values (if data has lots if zeros), take log, then finally reapply the negative sign where original was negative.

E.g. log(-2) = -1×(log(abs(-2)+1))

Hope that helps.

FRANK SAUVAGE Reply

February 15, 2016 at 10:02 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-105714)

Excellent guide! Thank you very much! Very pedagogic and comprehensive. Two thumbs up!
An excellent place to come back when starting a new data project…

HIRENDRASINGH CHAUHAN Reply

February 17, 2016 at 2:47 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-105802)

Very well explained article.. Person having basic math /statistics understanding can also understand subject well..

BIDHAN Reply

February 21, 2016 at 11:07 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-106000)

Concise and comprehensive. Great article.

WHY STATISTICS Reply

February 25, 2016 at 6:55 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-106218)

Very well written.

MATHU Reply

March 6, 2016 at 8:48 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-106727)

One of the best blogs I have ever read till date!

PANKAJ NEGI Reply

March 16, 2016 at 4:56 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-107505)

This is really great. Thank you so much!!

ABHISHEK Reply

March 28, 2016 at 5:15 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-108501)

Can we use Weight of Evidence to impute outliers and Missing Values??

BATOOL HAIDER Reply

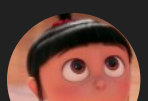
March 30, 2016 at 8:51 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-108686)

Great! Thanks!

ANUSHRI Reply

March 30, 2016 at 9:49 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-108689)

Very Useful. Thank you.. 🙂



AMY Reply

April 12, 2016 at 6:22 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-109365)

Amazing guide.. very structured and simplistic. enjoyed and learnt a lot reading this article.

ANDRII Reply

May 31, 2016 at 6:54 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-111650)

Many thanks for the guide, very useful. Would you advise R packages that help with data exploration?
Thanks

GUSTAVO Reply

June 1, 2016 at 2:26 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-111695)

THANK YOU FOR SHARING THIS CONCEPTS AND METHOD.

ARIJIT Reply

June 17, 2016 at 4:19 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-112328)

If a variable is very skewed at 0 but valid. How should we treat them in a logistic regression framework?

BHUVANESWARI Reply

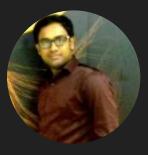
June 18, 2016 at 7:23 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-112364)

Very useful.
precise and clear
Thank you.

YASSER ABBASS Reply

June 19, 2016 at 8:31 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-112395)

Excellent article
Thank you very much



SWARNENDU **Reply**

June 26, 2016 at 5:37 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-112716)

really awesome..crisp and concise

BERKER **Reply**

June 28, 2016 at 2:10 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-112794)

I open a file in google drive to keep this page alone as a cheatsheet…Thank you so much..

MARKETING ANALYST (HTTP://WWW.DATANANALYTICS.COM) **Reply**

July 6, 2016 at 11:55 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-113148)

This is very useful summary, thank you for that!
I particularly liked the before-after comparisons to demonstrate the importance of the process steps.
Thanks,
Chill

NIRAV **Reply**

July 16, 2016 at 6:52 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-113543)

Great article! Few questions:
1) Do you run your data exploration on sample or full data set? If sample then what percentage and any article on how to take samples for unstructured text based dataset.
2) How to explore fields which are unstructured text, images etc. Do we need to run feature extraction before we explore. how do we explore them anyway?
I understand there's no single answer but in your opinion what's the best way to explore unstructured dataset.

GANESH **Reply**

July 19, 2016 at 10:58 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-113671)

The blog articles from AV are just awesome! Thanks to all the blog writers for sharing their knowledge.

YASH CHOUDHARY Reply

July 24, 2016 at 12:53 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-113920)

Definitely going to Bookmark this blog ! Thank you .

KRISHNA CHAITANYA Reply

August 9, 2016 at 4:50 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-114585)

you nailed the process. I thoroughly enjoyed reading your blog and learned a lot!!!! Thanks a lot for investing time and sharing your experience.

ASHISH SINGH (HTTP://ANALYTICSVIDHYA) Reply

August 14, 2016 at 8:22 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-114806)

Well Written. it really shows how to tackle the data

RAJESH SRINIVASAN Reply

August 24, 2016 at 8:33 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-115083)

Excellent read on EDA simple and to the point. Great Help to newbie like me.

MANGESH PANCHWAGH Reply

August 29, 2016 at 10:35 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-115281)

Thank you for sharing knowledge. It helps a lot.

AARON Reply

August 29, 2016 at 2:52 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-115291)

great article! Very useful!

R Reply

September 6, 2016 at 6:26 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-115690)

You Sir are amazing…

CAMILLE Reply

September 9, 2016 at 11:43 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-115852)

Great article!
I would like to add or comment on the imputation of missing values. I once had a dataset with missing values in one of the categorical variables. Instead of replacing missing values with the most frequent value of that variable, I looked at the distribution of unique values and found that they were all uniformly distributed. With this information, I would replace a missing value by randomly choosing a value among the set of unique values. It worked quite well but I would love to hear if this was statistically the right thing to do?

RAJIV Reply

September 11, 2016 at 1:02 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-115929)

The Best. Period.

NEERAJA Reply

September 12, 2016 at 1:33 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-115953)

Hi Sunil
Thank you very much for really useful and clear structure.

GAURAV Reply

September 15, 2016 at 7:48 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-116078)

Great explanation, would be better. If you could give us some sample data and then explain step by step on that.

ANUJ JAIN Reply

September 22, 2016 at 4:58 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-116356)

Loved reading it.
Thanks for sum it up in the best explanatory manner. 🙂
Best,

ASHISH KUMAR NAYAK Reply

October 6, 2016 at 8:19 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-116883)

Well summarised explanations covering each topic of data exploration with enough details to understand. Thanks a lot for this post.

YVETTE Reply

October 10, 2016 at 7:26 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-117051)

This is such an amazing resource. Thank you very much for sharing

---

ABHAY KUMAR (HTTP://ABHAYMISE.GITHUB.IO) Reply

November 4, 2016 at 5:25 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-117927)

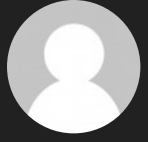
It was a crisp and clear and more importantly step by step explanation of EDA process.
I read all these things here and there but first time as an organized flow.

Keep up the good work sir.You understood the pain points of novice data scientist.

CAUI Reply

January 11, 2017 at 12:13 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-120843)

I’ve started to study Data Science fewmonths ago, this tutorial was one of the most clarifying for me, the step by step guide introduced the theory that can easily be used at practice. Thanks for the advices.



POONAM LATA Reply

January 25, 2017 at 8:03 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-121584)

Great! Very crisp, yet comprehensive.

BILL Reply

January 30, 2017 at 1:17 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-121813)

“Though, It can’t be applied to zero or negative values as well”. Did you mean “can” and not “can’t”

PARAM Reply

March 10, 2017 at 12:50 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-124381)

Excellent article. thanx

SANJAY Reply

April 3, 2017 at 10:39 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-126073)

Simple excellent post… keep writing.

- **A COMPLETE TUTORIAL WHICH TEACHES DATA EXPLORATION IN DETAIL | SHUJIAN'S BLOG (HTTP://WWW.SHUJIANLIU.COM/BLOGS/A-COMPLETE-TUTORIAL-WHICH-TEACHES-DATA-EXPLORATION-IN-DETAIL-2/)**

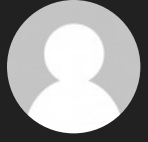
  April 3, 2017 at 2:10 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-126084)

[…] Source: http://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/?utm_content=buffer087f0 (http://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/?utm_content=buffer087f0) […]

JACK Reply

April 3, 2017 at 8:20 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-126101)

great one.
Could you please also add python sample code for these examples? Thank you.



JOSEPH MACHADO Reply

August 19, 2017 at 10:55 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-134701)

Hi Jack,
I am working on a prediction problem for which I am using this post as a guide for EDA. If you want some code examples please check out https://github.com/JosephKevin/sales_prediction (https://github.com/JosephKevin/sales_prediction)

Regards,
Joseph

HIRENDRASINGH CHAUHAN Reply

April 18, 2017 at 8:07 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-127075)

Very well written article. One suggestion for next Enhanced version of the Article

It would have been good of sample data set along with example from same data set is provided.

ABHIJIT DHADA (HTTP://WWW.THECRAZYANALYST.COM) Reply

April 18, 2017 at 9:29 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-127082)

Thank you Sunil for explaining the Data Exploration process very lucidly. Kudos !

VENUGOPAL Reply

June 12, 2017 at 7:29 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-130364)

Hi Sunil that was a nice article. Thank U

DN Reply

June 15, 2017 at 6:04 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-130519)

Good and nice flow of explanation. Really useful for base understanding.

STIF Reply

June 19, 2017 at 7:42 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-130719)

Thank you for the article, It is super helpful!
Do you mind providing the download of the dataset as well? Thanks! As a beginner, I'd like to follow your tutorial step by step!

KISHORE Reply

June 22, 2017 at 11:48 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-130915)

Hello Sunil,
Really an amazing stuff . Appreciate you for sharing your hard work..

AKASH GOYAL Reply

June 26, 2017 at 3:19 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-131134)

please tell me ,which course are better for statistical and exploratory analysis in sense of industry.

LAUTARO Reply

June 30, 2017 at 1:17 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-131332)

Best guide ever!

MIKE Reply

July 8, 2017 at 5:12 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-131690)

Hi, I was trying to research into covariate binning through Google, unfortunately I couldn't find anything.

Is there another term I could use that's more popular?

Thx.

AR Reply

July 16, 2017 at 1:01 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-

132191)

amazing guide, thanks so much for posting this. would love to hear more from you and dive deeper into this topic.

PANKAJ DNYANESHWAR GAIKWAD Reply

July 20, 2017 at 11:22 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-132531)

Really great help for beginners in data exploration and feature engineering!

BHUVANA NARAYANAN Reply

July 21, 2017 at 12:27 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-132558)

Very clear and concise as well as informative . Well done.

RAFAEL Reply

July 23, 2017 at 11:54 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-132696)

Very good article! Comprehensive and very easy to understand. Do you guys have any ebooks with all of this content?

ANU (HTTP://WWW.ANALYTICSVIDHYA.COM) Reply

July 31, 2017 at 8:52 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-133264)

great article. precisely written. Thanks for the clarity in the explanation given. keep up the good work.

WILLIAM Reply

August 1, 2017 at 11:11 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-133402)

well done. very helpful.

ROGER J. BROWN, PHD (HTTP://WWW.MATHESTATE.COM) Reply

August 7, 2017 at 2:53 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-

133796)

I agree with everyone else that this is a very good article. There are, however, some caveats. I am not a statistician so here is an incomplete list

1. Be sure whatever you do to data makes common sense, which should guide all your actions;
2. Be sure your data set is large enough such that the modifications you make have a small impact.
3. Beware of “messing with the randomness.” Remember that the reason the Monty Hall problem works the way it does is that the randomness of the first draw (3 doors) is disturbed midstream
4. Know about what effect your change can have on small samples. Two good examples to Google are Abscombe’s Quartet and Simpson’s Paradox. There are others.
5. Know the difference between mistakes and extreme values even though both are sometimes referred to as “outliers.”. The effect of extreme values may be valid and eliminating them can be very misleading (There is a huge literature on Extreme Value Theory. See http://www.mathestate.com (http://www.mathestate.com) for an in depth look at heavy tail phenomena).
6. Run a test for normality such as Jacque-Berta. If your model (like comparison of difference of means) requires normality and you use non-normal data you produce gibberish.

RJB

VIVEK Reply

August 11, 2017 at 6:31 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-134102)

The guide is super. IF you can take a sample dataset and apply all the steps to make dataset more informative then it would be very helpful.

JOSEPH MACHADO Reply

August 19, 2017 at 10:43 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-134700)

Hi Sunil,
Thank you for the amazing article, very organized and clear. I have a question
In the ‘Categorical & Continuous’ bivariate analysis part, if ANOVA shows a statistically significant difference between various groups in one variable, how do we incorporate this knowledge into the prediction process ?

Regards,
Joseph

PRABUDDH Reply

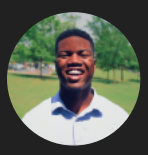
August 24, 2017 at 11:33 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-135043)

Thanks so much bro. Really useful stuff

ZUBAIR Reply

August 25, 2017 at 10:19 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-135117)

Thanks bro..for such an awesome article.



GARRETT RANSOM (HTTP://RULEUR.COM) Reply

August 30, 2017 at 5:26 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-135526)

Extremely helpful. Does a great job at breaking down each individual concept. Adding some actual code to the examples would also be helpful from a practical standpoint.

PROF DEEPALI N PANDE Reply

September 1, 2017 at 9:11 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-135996)

Wonderful and Descriptive but can I get some Working Codes which can highlight the procedure “what if the data is heterogeneous..?” (I mean to say multi-valued data and mixture of numeric and text form). Does Python, R or Matlab provide any help in this regard..?

- **FEATURE ENGINEERING 特徵工程中常見的方法 – I FAILED THE TURING TEST (HTTPS://VINTA.WS/CODE/FEATURE-ENGINEERING.HTML)**

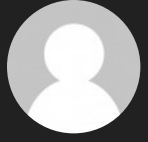
  September 18, 2017 at 4:32 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-137404)

[…] ref: https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/ (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/) […]

SRINI Reply

October 25, 2017 at 11:07 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-140750)

Thanks alot. Great article.



JOHAN (HTTP://WWW.MEDISENTIO.COM) Reply

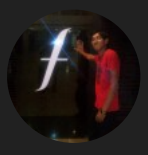
March 10, 2018 at 2:14 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-151818)

Help is make information from our data !
Thanks !

NWUT Reply

March 26, 2018 at 1:47 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-152181)

i think data that is scatter plot.
is Discrete variable, not continuous variable.



FAIZAN SHAIKH Reply

March 27, 2018 at 4:26 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-152208)

Hey – Can you clarify what your doubt is?

MOHAMMED ABDUL RAOOF Reply

April 13, 2018 at 4:14 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-152555)

Hi Ray,
It is good post as i am fresher it is very useful to me

DR DOR RAYAMAJHI Reply

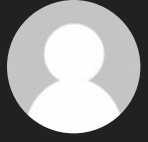
April 29, 2018 at 9:49 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-152915)

It is very useful. Thank you for your efforts Sunil.

NADA B Reply

May 14, 2018 at 7:41 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-153252)

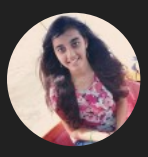
Very complete and useful ! Thank you !



BHAGWAT Reply

May 20, 2018 at 11:50 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-153413)

Extremely useful article, can someone guide me to a link or any resource where all steps mentioned above are applied on real dataset.



AISHWARYA SINGH Reply

May 21, 2018 at 8:43 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-153442)

Hi Bhagwat,

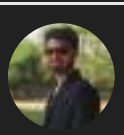
Here (https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+BigMS01+2018_1/about) is a training course on R for big mart sales dataset. A similar course will be made available soon.

NICOLÒ Reply

May 22, 2018 at 7:06 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-153471)

Great article, thanks!

TOP ANALYTICS VIDHYA USERS

Rank	Name	Points
1	 SRK (https://datahack.analyticsvidhya.com/user/profile/SRK)	3814
2	 mark12 (https://datahack.analyticsvidhya.com/user/profile/mark12)	3411
3	 nilabha (https://datahack.analyticsvidhya.com/user/profile/nilabha)	3169
4	 Tezdhar (https://datahack.analyticsvidhya.com/user/profile/Tezdhar)	3166
5	 nitish007 (https://datahack.analyticsvidhya.com/user/profile/nitish007)	3159

More User Rankings (https://datahack.analyticsvidhya.com/top-competitor/?utm_source=blog-navbar&utm_medium=web)

POPULAR POSTS

- 24 Ultimate Data Science Projects To Boost Your Knowledge and Skills (& can be accessed freely) (https://www.analyticsvidhya.com/blog/2018/05/24-ultimate-data-science-projects-to-boost-your-knowledge-and-skills/)
- 8 Useful R Packages for Data Science You Aren't Using (But Should!) (https://www.analyticsvidhya.com/blog/2019/04/8-useful-r-packages-data-science/)
- Understanding Support Vector Machine algorithm from examples (along with code) (https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/)
- A Complete Tutorial to Learn Data Science with Python from Scratch (https://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-learn-data-science-python-scratch-2/)
- Essentials of Machine Learning Algorithms (with Python and R Codes) (https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/)
- 7 Types of Regression Techniques you should know! (https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/)
- 6 Easy Steps to Learn Naive Bayes Algorithm (with codes in Python and R) (https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/)
- A Simple Introduction to ANOVA (with applications in Excel) (https://www.analyticsvidhya.com/blog/2018/01/anova-analysis-of-variance/)

RECENT POSTS

A Hands-On Introduction to Deep Q-Learning using OpenAI Gym in Python (https://www.analyticsvidhya.com/blog/2019/04/introduction-deep-q-learning-python/)

APRIL 18, 2019

8 Useful R Packages for Data Science You Aren't Using (But Should!) (https://www.analyticsvidhya.com/blog/2019/04/8-useful-r-packages-data-science/)

APRIL 17, 2019

Build your First Multi-Label Image Classification Model in Python (https://www.analyticsvidhya.com/blog/2019/04/build-first-multi-label-image-classification-model-python/)

APRIL 15, 2019

DataHack Radio #21: Detecting Fake News using Machine Learning with Mike Tamir, Ph.D. (https://www.analyticsvidhya.com/blog/2019/04/datahack-radio-machine-learning-identify-fake-news-mike-tamir/)

APRIL 11, 2019

(http://www.edvancer.in/certified-data-scientist-with-python-course?utm_source=AV&utm_medium=AVads&utm_campaign=AVadsnonfc&utm_content=pythonavad)

(https://courses.analyticsvidhya.com/bundles/ai-blackbelt-beginner-to-master?utm_source=Sticky_banner1&utm_medium=display)

(https://datahack.analyticsvidhya.com/contest/club-mahindra-dataolympics/?utm_source=Sticky_banner2&utm_medium=display)

Analytics Vidhya
Learn everything about analytics

ANALYTICS VIDHYA

About Us (http://www.analyticsvidhya.com/about-me/)

Our Team (https://www.analyticsvidhya.com/about-me/team/)

Career (https://www.analyticsvidhya.com/career-analytics-vidhya/)

Contact Us (https://www.analyticsvidhya.com/contact/)

Write for us (https://www.analyticsvidhya.com/about-me/write/)

DATA SCIENTISTS

Blog (https://www.analyticsvidhya.com/blog/)

Hackathon (https://datahack.analyticsvidhya.com/)

Discussions (https://discuss.analyticsvidhya.com/)

Apply Jobs (https://www.analyticsvidhya.com/jobs/)

Leaderboard (https://datahack.analyticsvidhya.com/leaderboard)

COMPANIES

Post Jobs (https://www.analyticsvidhya.com/corporate/)

Trainings (https://trainings.analyticsvidhya.com)

Hiring Hackathons (https://datahack.analyticsvidhya.com/)

Advertising

Reach Us (https://www.analyticsvidhya.com/contact/)

JOIN OUR COMMUNITY :

(https://www.facebook.com/analyticsvidhya)

46350

Followers

(https://www.facebook.com/analyticsvidhya)

(https://twitter.com/analyticsvidhya)

20387

Followers

(https://twitter.com/analyticsvidhya)

(https://plus.google.com/+Analyticsvidhya/posts)

Followers

(https://plus.google.com/+Analyticsvidhya/posts)

(https://in.linkedin.com/company/analytics-vidhya)

Followers

(https://in.linkedin.com/company/analytics-vidhya)

Subscribe to emailer >

Privacy Policy (https://www.analyticsvidhya.com/privacy-policy/)

Terms of Use (https://www.analyticsvidhya.com/terms/)

Refund Policy (https://www.analyticsvidhya.com/refund-policy/)

Don't have an account

cribe!

Sc
i
e
n
t
i
s
t