×

-

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

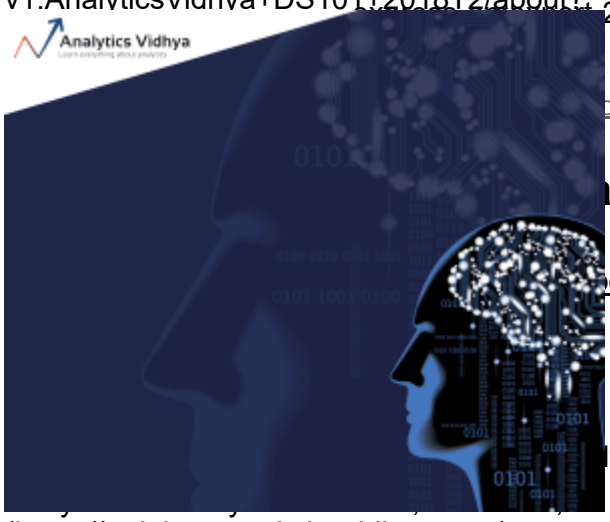
Analytics Vidhya
Learn everything about analytics

www.analyticsvidhya.com/blog/)

(https://datahack.analyticsvidhya.com/contest/american-

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2018T2/about?

2018/?utm_source=AVtopblogBanner)



OM/BLOG/CATEGORY/BUSINESS-ANALYTICS/)

Reply

ta Exploration

G/AUTHOR/SUNIL-RAY/), JANUARY 10, 2016

If you are in a state of mind, that machine learning can sail you 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.

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)

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

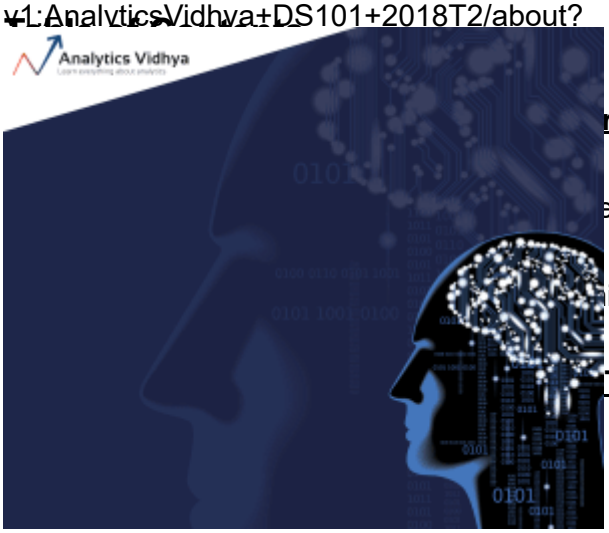
I have been a Business Analytics professional for close to three years now. In my initial days, one of my mentor suggested me to spend significant time on exploration and analyzing data. Following his advice has served me well.

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.

Subscribe!

/uploads/2016/01/de.jpg)

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2018T2/about?



ration

equired ?

issing value ?

Treatment

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)

- What is the impact of outliers on dataset ?
- 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.

Subscribe!

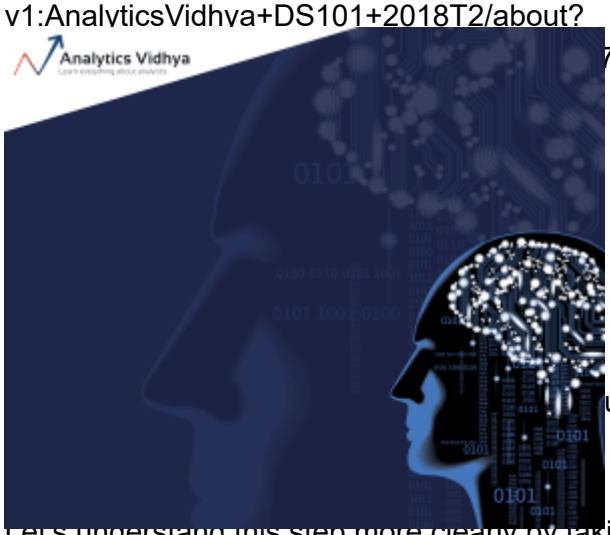
1. Steps of Data Exploration and Preparation

Remember the quality of your inputs decide the quality of your output. So, once you have got your business

lot of time and efforts here. With my personal estimate, data e up to 70% of your total project time.

lean and prepare your data for building your predictive model:

7. Variable creation (https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2018T2/about?



7 multiple times before we come up with our refined model.

utput) variables. Next, identify the data type and category of the

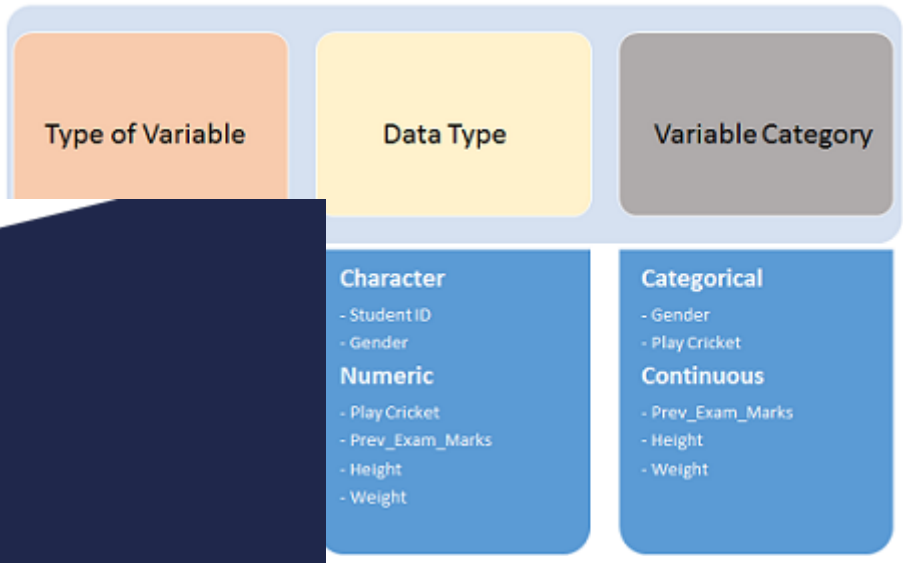
Let's understand this step more clearly by taking an example. (https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)

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.

Student_ID	Gender	Prev_Exam_Marks	Height (cm)	Weight Caregory (kgs)	Play Cricket
S001	M	65	178	61	1
S002	F	75	174	56	0
S003	M	45	163	62	1
S004	M	57	175	70	0
S005	F	59	162	67	0

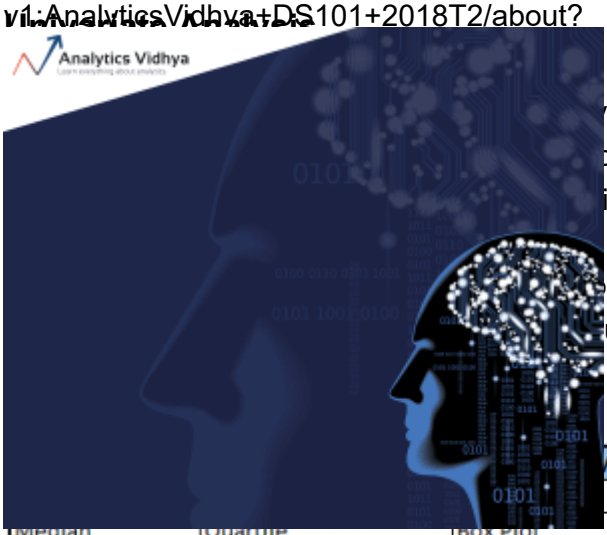
(https://www.analyticsvidhya.com/wp-content/uploads/2015/02/Data_exploration_11.png)Below, the variables have been defined in different category:

Subscribe!



/uploads/2015/02/Data_exploration_2.png)

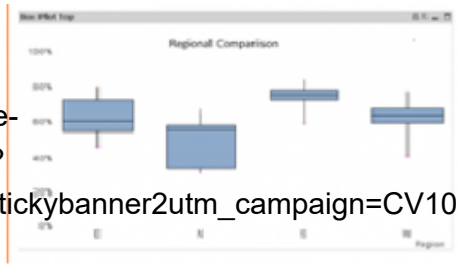
(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2018T2/about?

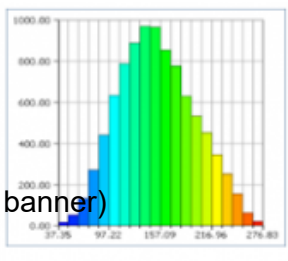


one. Method to perform uni-variate analysis will depend on ontinuous. Let's look at these methods and statistical measures idually:

us variables, we need to understand the central tendency and using various statistical metrics visualization methods as shown

Methods		
Median	Quartile	Box Plot
Mode	IQR	
Min	Variance	
Max	Standard Deviation	
	Skewness and Kurtosis	





(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+GVDL101+CVDL101_T1/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)

(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.

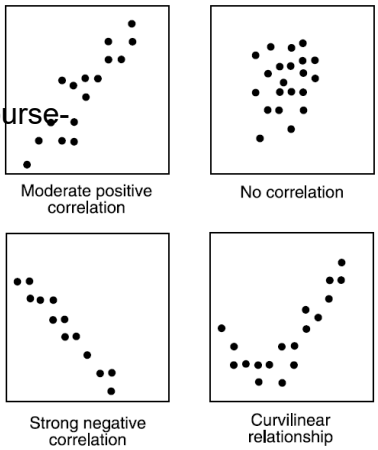
Subscribe!

Bi-variate Analysis

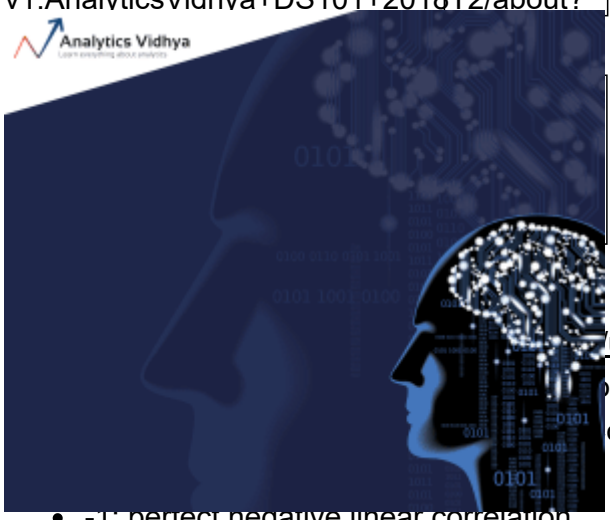
Bi-variate Analysis finds out the relationship between two variables. Here, we look for association and disassociation between variables at a pre-defined significance level. We can perform bi-variate analysis for any combination of categorical and continuous variables. The combination can be: Categorical & Categorical, & Continuous. Different methods are used to tackle these

detail:

-variate analysis between two continuous variables, we should t the relationship between two variables. The pattern of scatter les. The relationship can be linear or non-linear.



(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2018_T2/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2&utm_campaign=CV101banner)



/uploads/2015/02/Data_exploration_4.png)Scatter plot shows es not indicates the strength of relationship amongst them. To orrelation. Correlation varies between -1 and +1.

- -1: perfect negative linear correlation
- +1:perfect positive linear correlation and
- 0: No correlation

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2&utm_campaign=CV101banner)
(https://www.analyticsvidhya.com/wp-content/uploads/2015/02/Data_exploration_4.png)

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:

Subscribe!

X	65	72	78	65	72	70	65	68
Y	72	69	79	69	84	75	60	73

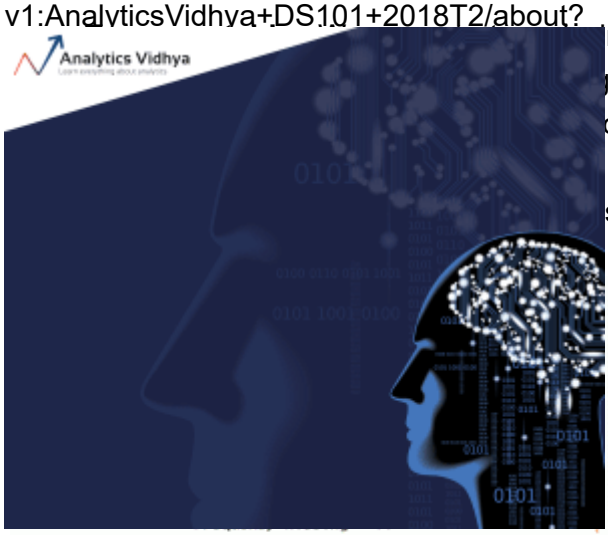
Metrics	Formula	Value
Co-Variance (X,Y)	=COVAR(F6:L6,F7:L7)	18.77
		18.48
		45.23
		0.65

/uploads/2015/02/Data_exploration_51.png)

ationship(0.65) between two variables X and Y.

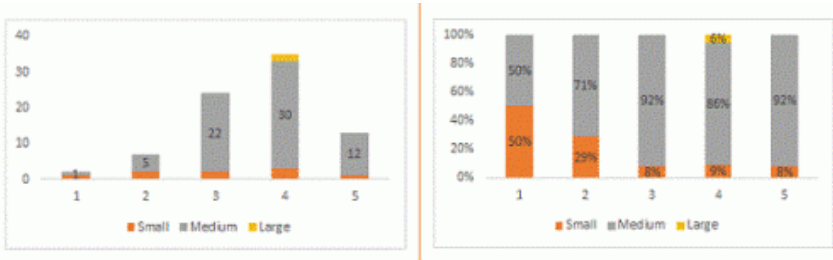
onship between two categorical variables, we can use following methods:

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2018T2/about?



ng the relationship by creating a two-way table of count and gory of one variable and the columns represent the categories of ount% of observations available in each combination of row and

s more of a visual form of Two-way table.



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

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)

- **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.

Subscribe!

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:

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

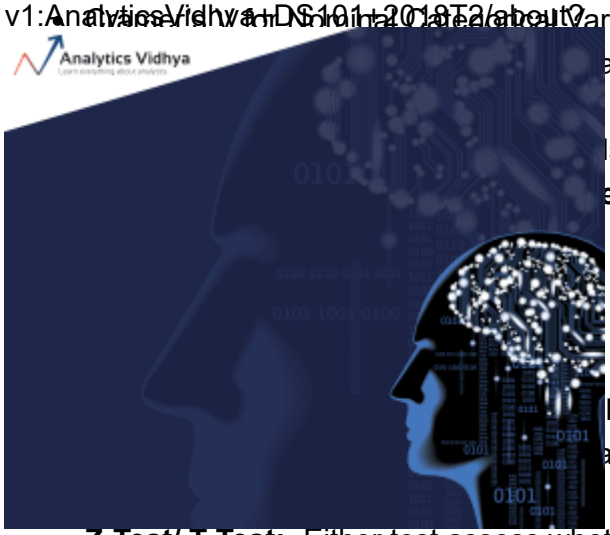
tent/uploads/2015/02/Data_exploration_7.png)where *O*
the expected frequency under the null hypothesis and computed

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

tent/uploads/2015/02/Data_exploration_8.png)
ted count for product category 1 to be of small size is 0.22. It is
) times the column total for Product category (2) then dividing by
s conducted for each cell. Statistical Measures used to analyze

(https://trainings.analyticsvidhya.com/courses/course-
v1:AnalyticsVidhya+DS101+2018T2/about?
utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)

- **Chi-Square:** It is a measure of statistical significance between two categorical variable



l categorical variable.

ls have specific methods to perform chi-square test. In SAS, we
eq to perform this test.

relation between categorical and continuous variables, we can
l variables. If levels are small in number, it will not show the
l 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

(https://trainings.analyticsvidhya.com/courses/course-
v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?
utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)

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

(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.

Subscribe!

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{S^2\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

where:

- $\overline{X}_1, \overline{X}_2$: *Averages*
- S_1^2, S_2^2 : *Variances*
- N_1, N_2 : *Counts*
- t : has t distribution with $N_1 + N_2 - 2$ degree of freedom

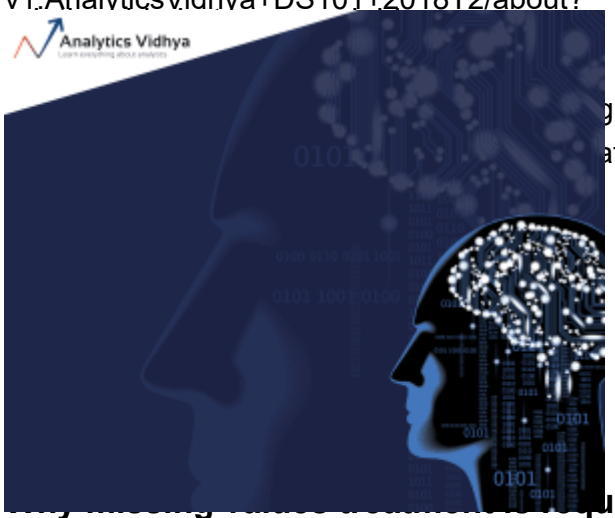
ontent/uploads/2015/02/ttest.png)

age of more than two groups is statistically different.

t of five different exercises. For this, we recruit 20 men and
os). Their weights are recorded after a few weeks. We need to
on them is significantly different or not. This can be done by
en each.

ee stages of Data Exploration, Variable Identification, Uni-Variate and Bi-Variate analysis. We also looked at various statistical and visual methods to identify the

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2018T2/about?



g values Treatment. More importantly, we will also look at why
ating them is necessary.

uired?

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)

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.

Subscribe!

Name	Weight	Gender	Play Cricket/ Not
Mr. Amit	58	M	Y
Mr. Anil	61	M	Y
Miss Swati	58	F	N
Miss Richa	55		Y
Mr. Steve	55	M	N

Name	Weight	Gender	Play Cricket/ Not
Mr. Amit	58	M	Y
Mr. Anil	61	M	Y
Miss Swati	58	F	N
Miss Richa	55	F	Y
Mr. Steve	55	M	N
Miss Reena	64	F	Y
Miss Rashmi	57	F	Y
Mr. Kunal	57	M	N

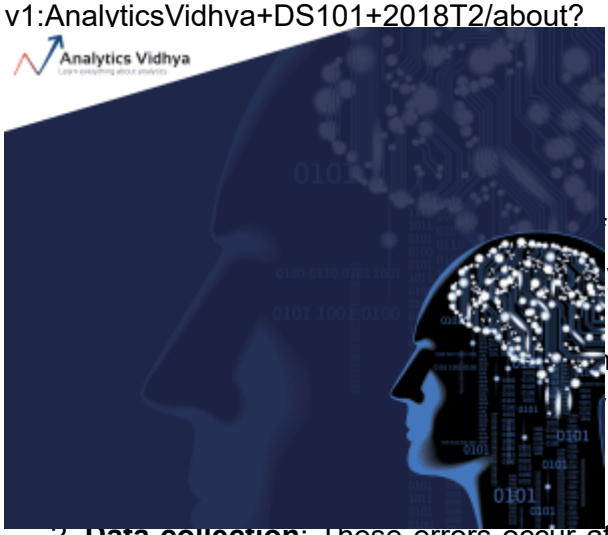
y Cricket
50%
50%
00%

Gender	#Students	#Play Cricket	%Play Cricket
F	4	3	75%
M	4	2	50%

/uploads/2015/02/Data_Exploration_2_11.png)

above: In the left scenario, we have not treated missing values. ances of playing cricket by males is higher than females. On the which shows data after treatment of missing values (based on), where females have higher chances of playing cricket compared to males.

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2018T2/about?



missing values in a dataset. Now, let's identify the reasons for y occur at two stages:

here are problems with extraction process. In such cases, we ith data guardians. Some hashing procedures can also be used Errors at data extraction stage are typically easy to find and can

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)

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.

Subscribe!

- **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.

...ng values ?

...Deletion and Pair Wise Deletion.
...servations where any of the variable is missing. Simplicity is one ...ethod, but this method reduces the power of model because it ...
...n analysis with all cases in which the variables of interest are ...od is, it keeps as many cases available for analysis. One of the ...ses different sample size for different variables.

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

List wise deletion

Sales
343
~~280~~
332
~~272~~
~~.~~
326
~~259~~
297

Pair wise deletion

Gender	Manpower	Sales
M	25	343
F	~~.~~	280
M	33	332
M	~~.~~	272
F	25	~~.~~
M	29	326
~~ ~~	26	259
M	32	297

...(https://www.analyticsvidhya.com/wp-content/uploads/2015/02/Data_Exploration_2_2.png) ... 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
e based on other attributes of the training data set and populate
can use regression, ANOVA, Logistic regression and various
ere are 2 drawbacks for this approach:
usually more well-behaved than the true values
attributes in the data set and the attribute with missing values,
for estimating missing values.

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2018_T2/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)

utation, the missing values of an attribute are imputed using the
t similar to the attribute whose values are missing. The similarity
distance function. It is also known to have certain advantage &

Advantages:

- k-nearest neighbour can predict both qualitative & quantitative attributes
- el for each attribute with missing data is not required
- sing values can be easily treated
- data is taken into consideration

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)

-consuming in analyzing large database. It searches through all
most similar instances.
critical. Higher value of k would include attributes which are
what we need whereas lower value of k implies missing out of

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.

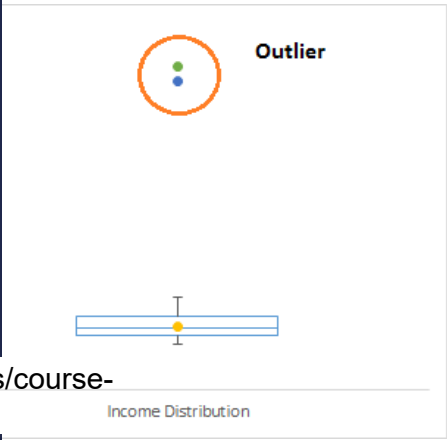
3. Techniques of Outlier Detection and Treatment

What is an Outlier?

Subscribe!

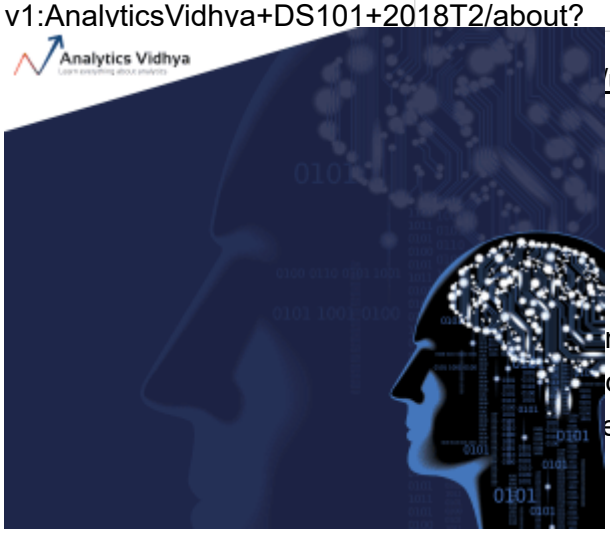
Outlier is a commonly used terminology by analysts and data scientists as it needs close attention else it can result in wildly wrong estimations. Simply speaking, Outlier is an observation that appears far away and diverges from an overall pattern in a sample.

Let's take an example, we do customer profiling and find out that the average annual income of customers is ving annual income of $4 and $4.2 million. These two customers e population. These two observations will be seen as Outliers.



(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2018T2/about?

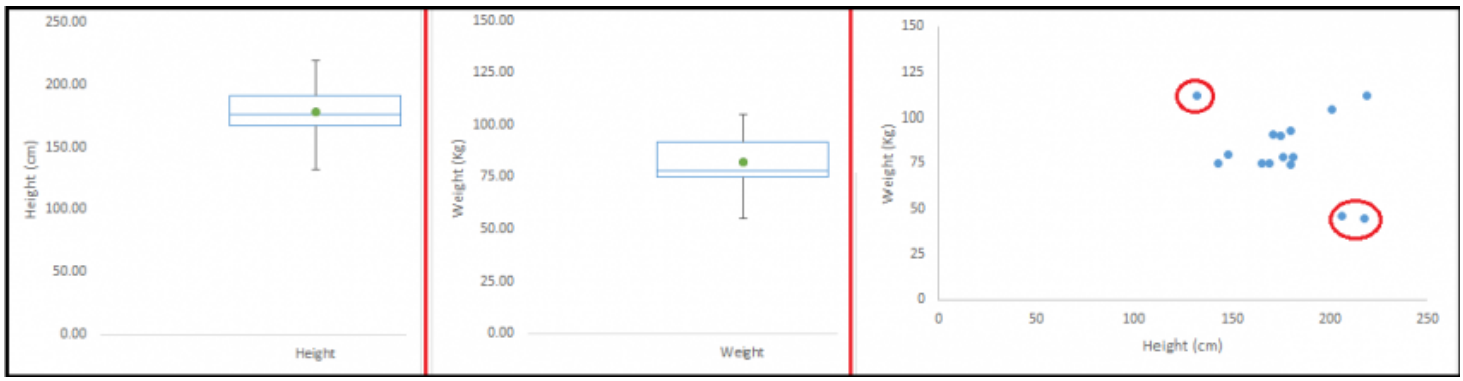
/uploads/2015/02/Outlier.png)



nd **Multivariate**. Above, we have discussed the example of d when we look at distribution of a single variable. Multi-variate e. In order to find them, you have to look at distributions in multi-

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)

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)

Subscribe!

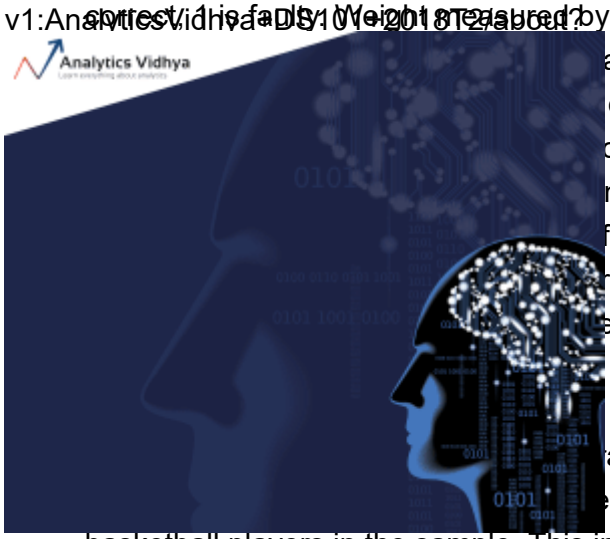
What causes Outliers?

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:

nore detail:

n as errors caused during data collection, recording, or entry can nnual income of a customer is $100,000. Accidentally, the data in the figure. Now the income becomes $1,000,000 which is 10 utlier value when compared with rest of the population.

mmon 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 asured on faulty machine can lead to outliers.



outliers is experimental error. For example: In a 100m sprint of 7 centrating on the ‘Go’ call which caused him to start late. Hence, more than other runners. His total run time can be an outlier.

found in self-reported measures that involves sensitive data. For report the amount of alcohol that they consume. Only a fraction e actual values might look like outliers because rest of the teens

perform data mining, we extract data from multiple sources. It is raction errors may lead to outliers in the dataset.

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.

n example to check what happens to a data set with and without

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2018T2/about?

	With Outlier
4	4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7,300
	Mean = 30.00
Median = 5.00	Median = 5.50
	Mode = 5.00
	Standard Deviation = 85.03

/uploads/2015/02/Outlier_31.png)

significantly different mean and standard deviation. In the first ut with the outlier, average soars to 30. This would change the

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?utm_source=CV101AvBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)

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.

Subscribe!

- 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.

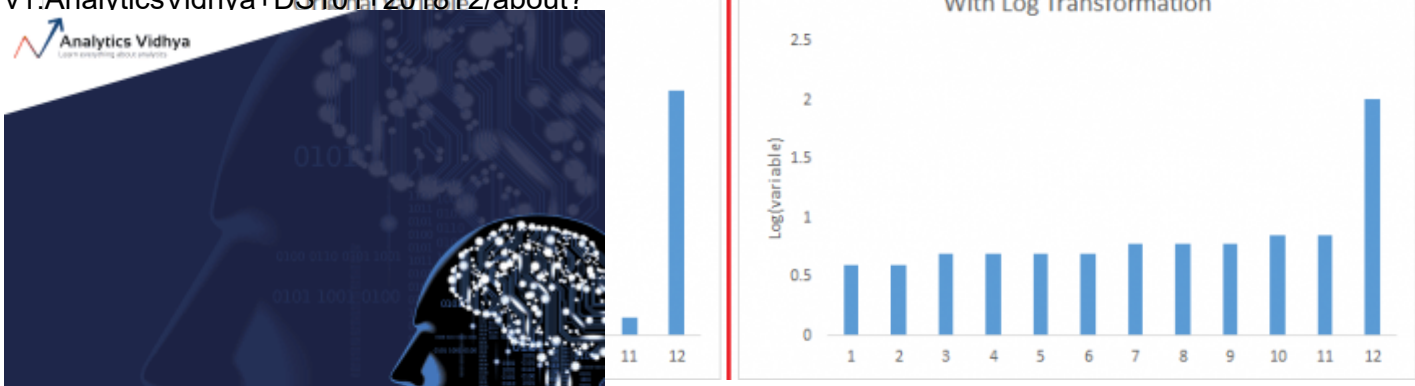
How to remove Outliers?

ilar to the methods of missing values like deleting observations, m as a separate group, imputing values and other statistical echniques used to deal with outliers:

values if it is due to data entry error, data processing error or s. We can also use trimming at both ends to remove outliers.

ming variables can also eliminate outliers. Natural log of a value ues. Binning is also a form of variable transformation. Decision ell due to binning of variable. We can also use the process of

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2018T2/about?



/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.

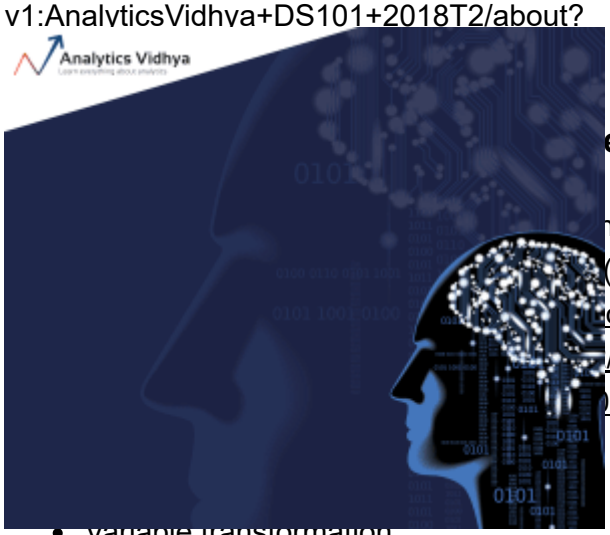
Subscribe!

4. The Art of Feature Engineering

of extracting more information from existing data. You are not
ly making the data you already have more useful.

ct foot fall in a shopping mall based on dates. If you try and use
tract meaningful insights from the data. This is because the foot
nan it is by the day of the week. Now this information about day
ring it out to make your model better.

This exercising of bringing out information from data in known as feature engineering.
(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2018T2/about?



eering ?

nave completed the first 5 steps in data exploration – Variable
(https://www.analyticsvidhya.com/blog/2015/02/data-exploration-
on (https://www.analyticsvidhya.com/blog/2015/02/7-steps-data-
/) and Outliers Treatment
02/outliers-detection-treatment-dataset/). Feature engineering

- variable transformation.
- Variable / Feature creation.

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)
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.

Subscribe!

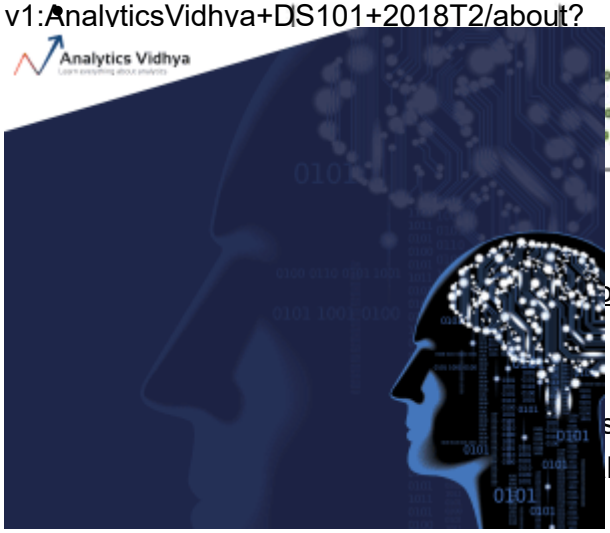
When should we use Variable Transformation?

Below are the situations where variable transformation is a requisite:

of a variable or standardize the values of a variable for better ation is a must if you have data in different scales, this ape of the variable distribution

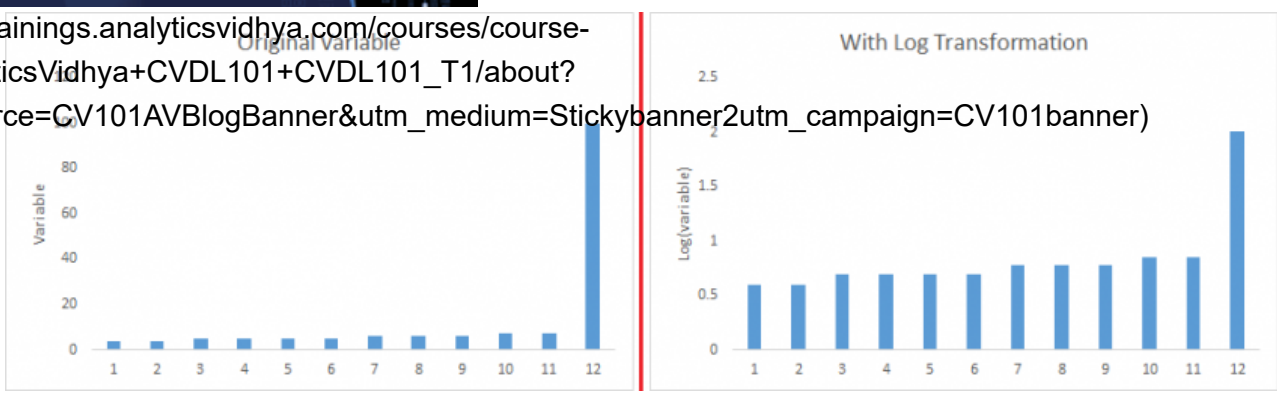
n-linear relationships into linear relationships. Existence of a is easier to comprehend compared to a non-linear or curved nvert a non-linear relation into linear relation. Scatter plot can be two continuous variables. These transformations also improve e of the commonly used transformation technique used in these

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2018T2/about?



ontent/uploads/2015/03/Relation.png)**Symmetric distribution is** as it is easier to interpret and generate inferences. Some distribution of variables. So, whenever we have a skewed s which reduce skewness. For right skewed distribution, we take ble and for left skewed, we take square / cube or exponential of

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)



(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.

riable Transformation?

m variables. As discussed, some of them include square root, cal and many others. Let's look at these methods in se transformation methods.

ommon transformation method used to change the shape of bution plot. It is generally used for reducing right skewness of zero or negative values as well.

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2018_T2/about?)

- **Square / Cube root:** The square and cube root of a variable has a sound effect on variable distribution. arithmic transformation. Cube root has its own advantage. It can g zero. Square root can be applied to positive values including

bles. It is performed on original values, percentile or frequency. is based on business understanding. For example, we can namely: High, Average and Low. We can also perform co-variate more than one variables.

What is Feature / Variable Creation & its Benefits?

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)

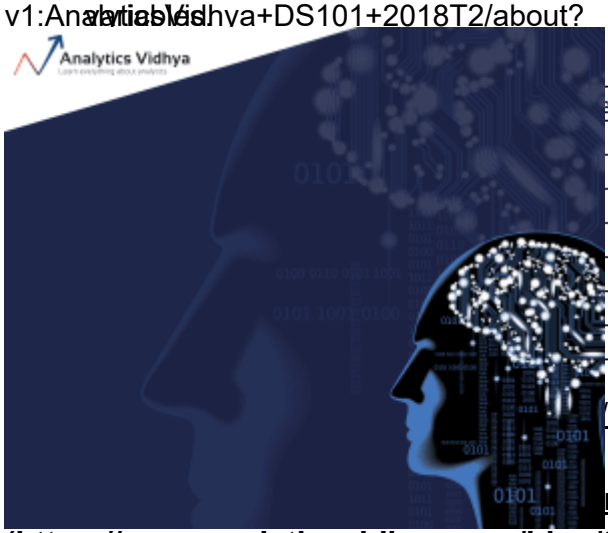
Feature / Variable creation is a process to generate a new variables / features based on existing variable(s). For example, say, we have date(dd-mm-yy) as an input variable in a data set. We can generate new variables like day, month, year, week, weekday that may have better relationship with target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

Emp_Code	Gender	Date	New_Day	New_Month	New_Year
A001	Male	21-Sep-11	21	9	2011
A002	Female	27-Feb-13	27	2	2013
A003	Female	14-Nov-12	14	11	2012
A004	Male	07-Apr-13	7	4	2013
A005	Female	21-Jan-11	21	1	2011
A006	Male	26-Apr-13	26	4	2013
A007	Male	15-Mar-12	15	3	2012

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 used the salutation (Master, Mr, Miss, Mrs) of name as a new variable to create? Honestly, this depends on business riosity and the set of hypothesis he might have about the of variables, binning variables and other methods of variable te new variables.

he most common application of dummy variable is to convert ables. Dummy variables are also called Indicator Variables. It is a predictor in statistical models. Categorical variable can take ender'. We can produce two variables, namely, "**Var_Male**" with **ar_Female**" with values 1 (Female) and 0 (No Female). We can (https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2018T2/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner) use n classes of a categorical variables with n or n-1 dummy



Gender	Var_Male	Var_Female
Male	1	0
Female	0	1
Female	0	1
Male	1	0
Female	0	1
Male	1	0
Male	1	0

/uploads/2015/03/Dummy.png)

rmation / creation ideas (https://www.analyticsvidhya.com/blog/2013/11/simple-manipulations-extract-data/) which can be applied to your data.

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)

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.

Subscribe!

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.

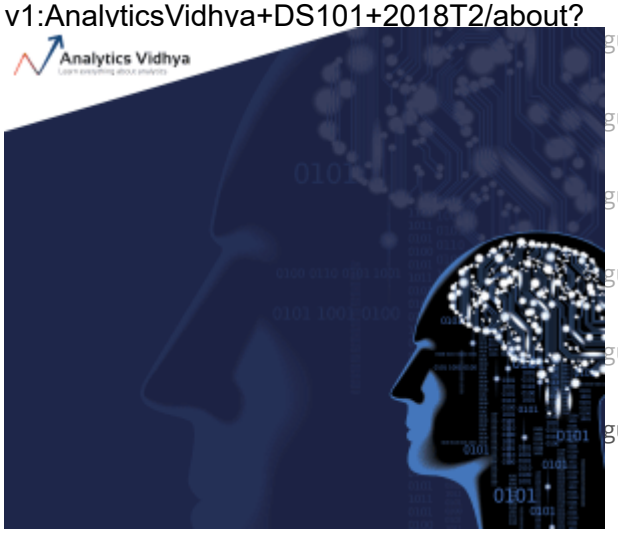
If you like what you just read & want to continue your analytics learning, subscribe to our [n/fb/a/mailverify?uri=analyticsvidhya), follow us on svidhya) or like our facebook page).

hya's Android APP

GET IT ON Google Play

=blog_article&utm_campaign=blog&pcampaignid=MKT-Other-

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2018T2/about?



(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)

Share this:

guide-data-exploration/?share=linkedin&nb=1)

guide-data-exploration/?share=facebook&nb=1)

568

guide-data-exploration/?share=google-plus-1&nb=1)

guide-data-exploration/?share=twitter&nb=1)

guide-data-exploration/?share=pocket&nb=1)

guide-data-exploration/?share=reddit&nb=1)

Like this:

Loading...

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

Subscribe!

(HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/OUTLIER-REMOVAL/), UNIVARIATE ANALYSIS (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/UNIVARIATE-ANALYSIS/)

NEXT ARTICLE

erfectly captures the growth of Data Science

016/01/20-powerful-images-perfectly-captures-growth-data-science/)

•••

PREVIOUS ARTICLE

to Become a Data Scientist in 2016

m/blog/2016/01/ultimate-plan-data-scientist-2016/)

RELATED ARTICLES

TAVISH SRIVASTAVA (HTTPS://WWW.A...

(https://www.analyticsvidhya.com/

KUNAL JAIN (HTTPS://WWW.ANALYTIC...

(https://www.analyticsvidhya.com/

ANALYTICS VIDHYA CONTENT TEAM (H...

(https://www.analyticsvidhya.com/

time-series-data-r/)

Exploration of Time Series Data in R (https://www.analyticsvidhya.co

review-pg-diploma-in-data-analytics-by-upgrad-iiit-b/)

Course Review – PG Diploma in Data Analytics by UpGrad & IIIT-

RAY (HTTPS://WWW.ANALYTICS...

entrepreneurs-big-data-analytics-data-science/)

Top Datapreneurs who made data science what it is today

(https://www.analyticsvidhya.co

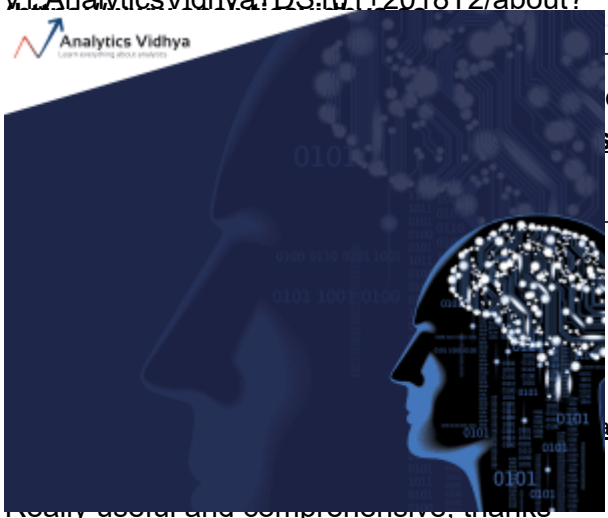
based way to learn Machine Learning
(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2018T2/about?

//www.analyticsvidhya.com/sas-macro/)

uction to SAS Macros
://www.analyticsvidhya.co
m/blog/2014/10/introduction-sas-macro/)

(https://www.analyticsvidhya.com/base-sas-interview-questions-part-ii/)

Tricky Base SAS interview questions : Part-II (https://www.analyticsvidhya.com/blog/2014/04/tricky-base-sas-interview-questions-part-ii/)



et a prompt response from the author. We request you to post
ssion portal (https://discuss.analyticsvidhya.com/) to get your

Reply

analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-103484)

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)

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**

Subscribe!

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!

Reply

analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-103491)

rtainly a good refresher. Keep writing!

Reply

analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-103494)

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)

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

Reply

analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-103510)

ts of Machine Learning. The points are explained in a simple

Reply

analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-103511)

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)

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)

Subscribe!

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.

Reply

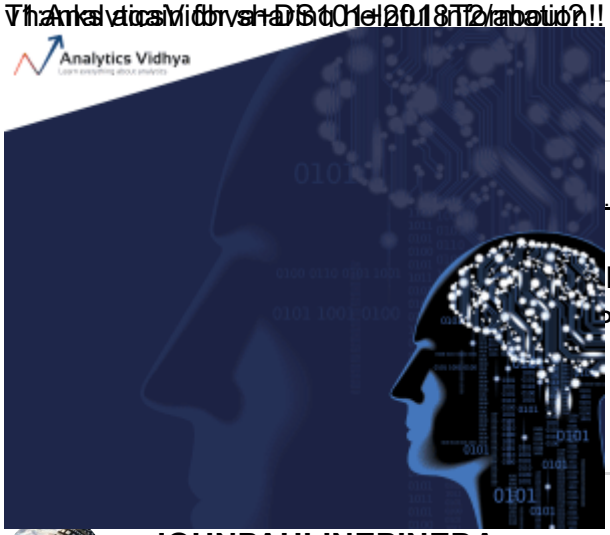
analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-103529)

t. I like your blogs, Please continue your good work !

Reply

analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-103539)

mprehensive information. Also, I would request some to write a
r than other BI tools like Tableau, Qlikview….gaining more
analyticsvidhya.com/courses/course-
Thanks Vidhya!! DB101-2018T2/about?!!



Reply

.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-103548)

l. I appreciated if you continue this wonderful work and post an
Python.

JOHNPAUL INERINEDA **Reply**

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

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+CV101+CV101_T1/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)
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.

Subscribe!

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.

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2018T2/about?utm_source=CV101AvBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)

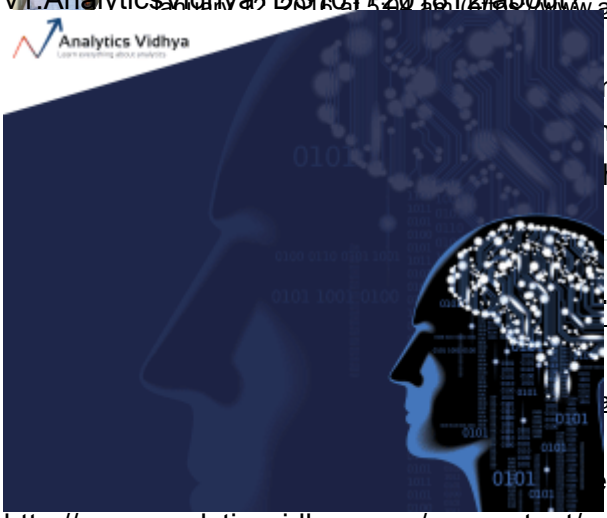
Reply

analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-103559)

glad it helped.

DARIO ROMERO Reply

January 12, 2016 at 7:26 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-103560)



portant topic. BTW, there is a missing graph on the paragraph
nalysis. Could you please edit it and add the missing graph. I
he ping file. Thanks.

ICSVIDHYA.COM/BLOG/2016/01/GUIDE-DATA- Reply

analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-103561)

ere:
http://www.analyticsvidhya.com/wp-content/uploads/2015/02/Data_exploration_4.png
(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?utm_source=CV101AvBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)
(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.

Subscribe!

Thanks,
Akshay Kher

Reply

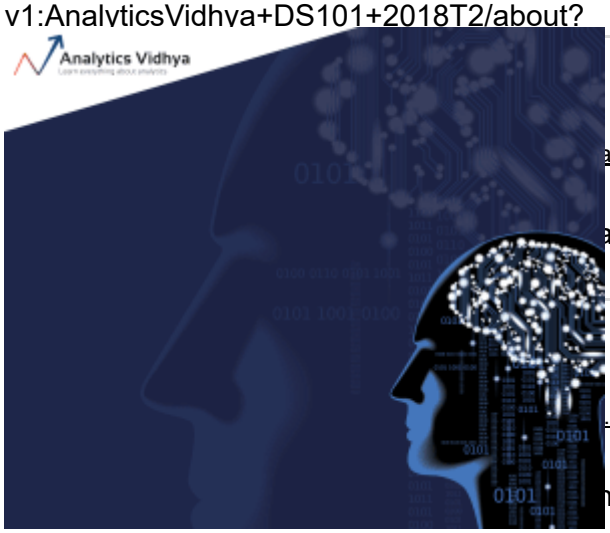
analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-103588)

Sunil. I had All these points scattered across but you got all of
Bookmarked this page and this would now be my first page to

Reply

analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-103633)

Thanks
(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2018T2/about?



Reply

analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-103871)

a analytics project. Good work keep up.

Reply

.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-104271)

helped me a lot….Thanks a lot
(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)

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. !!

Subscribe!

SUHEL Reply

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

n for zero or negative values.

one to all values (if data has lots if zeros), take log, then finally
negative.

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2018T2/about?

FRANK SAUVAGE Reply

w.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-105714)

pedagogic and comprehensive. Two thumbs up!
g a new data project…

Reply

analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-105802)

sic math /statistics understanding can also understand subject

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)

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

Subscribe!

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!

Reply

alyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-107505)

Reply

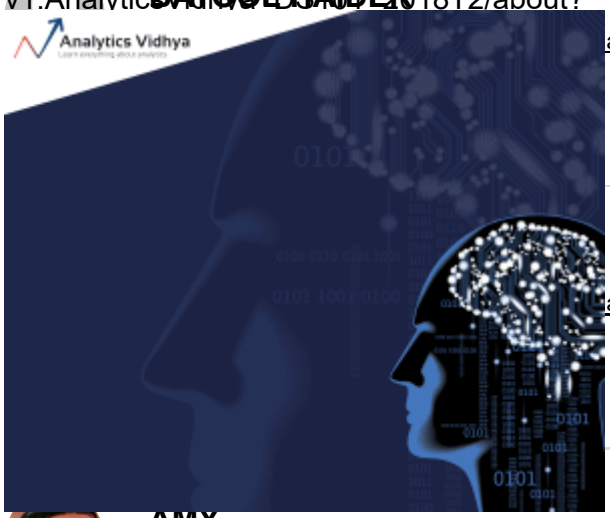
alyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-108501)

tliers and Missing Values??

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS401+2018T2/about?

BATOOL HAIDER

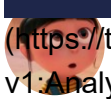
Reply



alyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-108686)

Reply

alyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-108689)



AMY

Reply

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

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)
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)

Subscribe!

THANK YOU FOR SHARING THIS CONCEPTS AND METHOD.

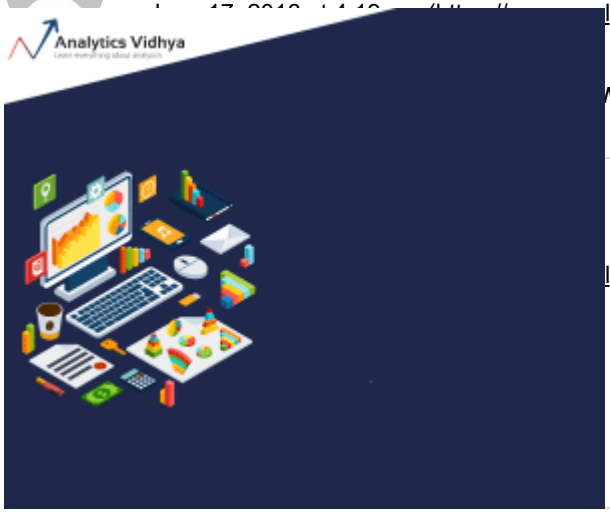
ARIJIT Reply

lyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-112328)

w should we treat them in a logistic regression framework?

Reply

lyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-112364)



(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2018T2/about?

VISHWAR ARYAN Reply

lyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-112395)

Reply

lyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-112716)

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?utm_source=CV101AVBlogBanner&utm_medium=StickyBanner2&utm_campaign=CV101Banner)

BERKER Reply

June 28, 2016 at 12:18 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

Subscribe!

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

ed text based dataset.

ed text, images etc. Do we need to run feature extraction before

?

our opinion what's the best way to explore unstructured dataset.

Reply

lyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-113671)

Thanks to all the blog writers for sharing their knowledge.

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2018_T2/about?

SWETA KUSHWAHA Reply

lyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-113920)

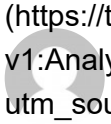
k you .

Reply

alyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-114585)

reading your blog and learned a lot!!!! Thanks a lot for investing

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2018_T2/about?utm_source=av&utm_medium=Sticky_banner_2018_01_campaign=CV101b#comment-114806)



ASHISH SINGH (HTTP://ANALYTICSVIDHYA) Reply

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

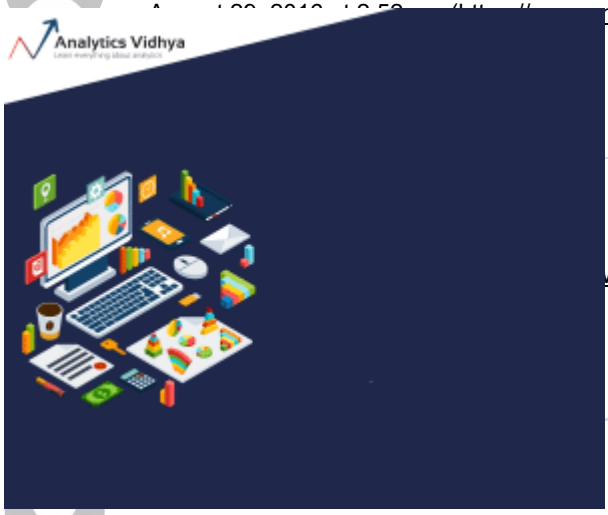
August 29, 2016 at 10:35 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-115281)

Subscribe!

Thank you for sharing knowledge. It helps a lot.

AARON Reply

[illegible]nalyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-115291)



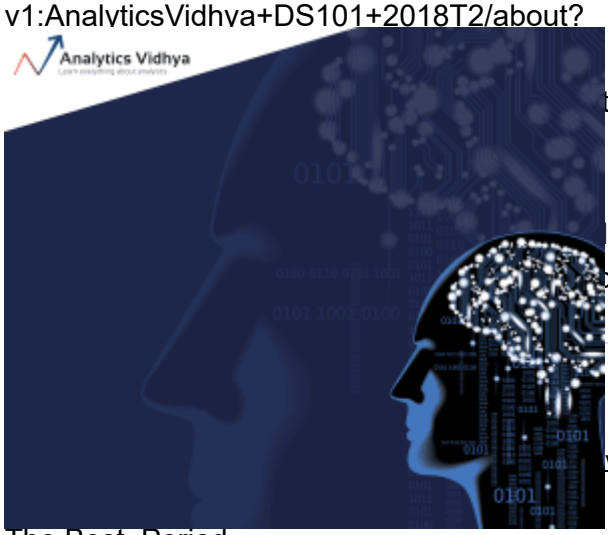
Reply

[illegible].analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-115690)

Reply

September 8, 2016 at 1:43 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-115852)

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2018T2/about?



[illegible]tion of missing values. I once had a dataset with missing values [illegible] replacing missing values with the most frequent value of that [illegible] values and found that they were all uniformly distributed. With [illegible]ue by randomly choosing a value among the set of unique [illegible]o hear if this was statistically the right thing to do?

Reply

[illegible]w.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-115929)

The Best Period

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)

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.

Subscribe!

ANUJ JAIN Reply

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

manner. 🙂

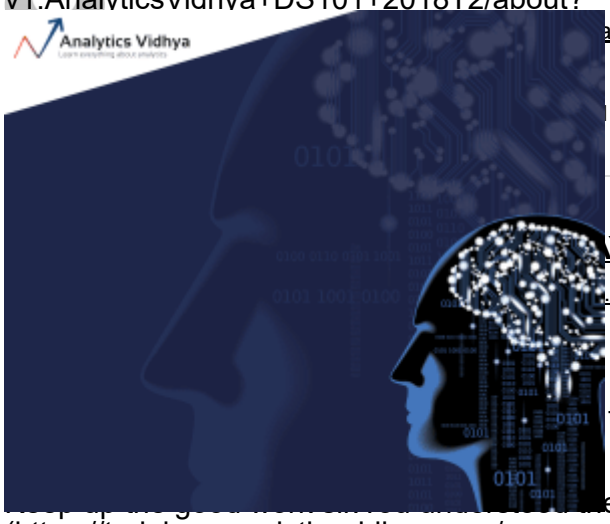
Reply

nalyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-116883)

topic of data exploration with enough details to understand.

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2018T2/about?

YVETTE Reply



analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-117051)

very much for sharing

YMISE.GITHUB.IO) Reply

analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-117927)

step by step explanation of EDA process.

time as an organized flow.

e pain points of novice data scientist.

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)

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.

Subscribe!

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"

Reply

nalyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-124381)

Reply

yticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-126073)

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2018T2/about?

- **A COMPLETE TUTORIAL WHICH TEACHES DATA EXPLORATION IN DETAIL | [illegible]W.SHUJIANLIU.COM/BLOGS/A-COMPLETE-TUTORIAL-[illegible]LORATION-IN-DETAIL-2/)**

ticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-126084)

blog/2016/01/guide-data-exploration/?utm_content=buffer087f0
/guide-data-exploration/?utm_content=buffer087f0) […]

Reply

ticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-126101)

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)

Could you please also add Python 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

Subscribe!

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

long with example from same data set is provided.

V.THECRAZYANALYST.COM) Reply

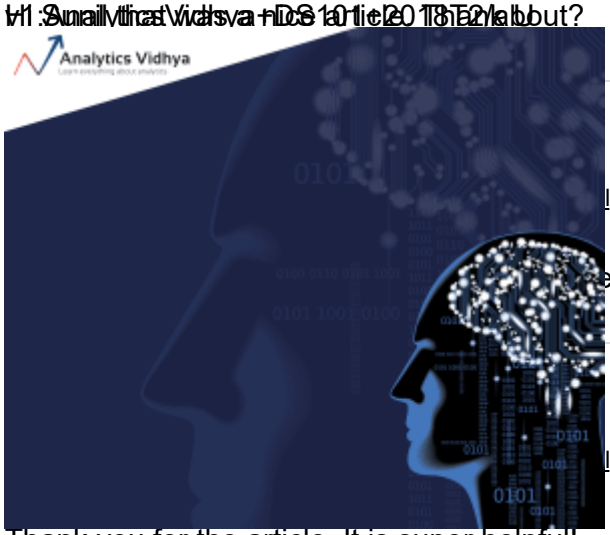
lyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-127082)

oration process very lucidly. Kudos !

Reply

lyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-130364)

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2017_T2/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)



Reply

lyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-130519)

eful for base understanding.

Reply

lyticsvidhya.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!

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)

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

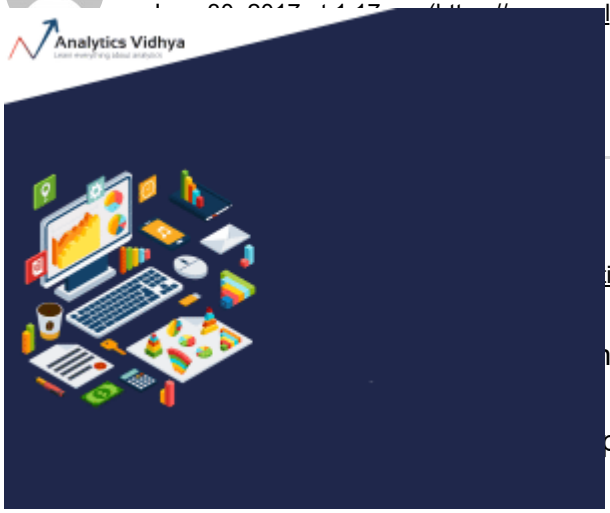
June 26, 2017 at 3:19 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-131134)

Subscribe!

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

LAUTARO Reply

lyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-131332)



Reply

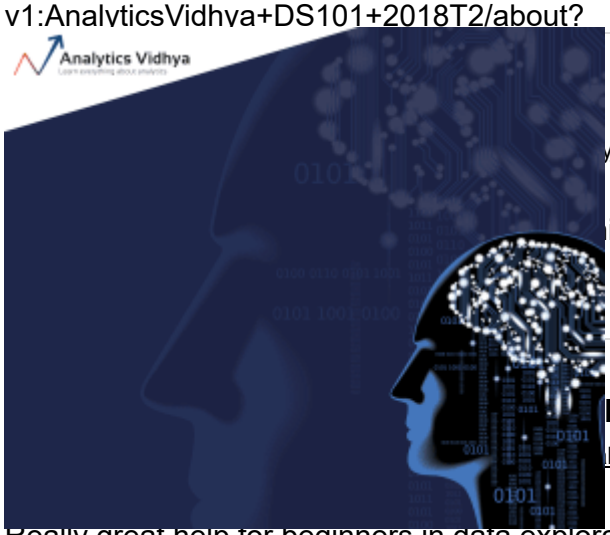
icsvidhya.com/blog/2016/01/guide-data-exploration/#comment-131690)

ning through Google, unfortunately I couldn't find anything.

opular?

Thx

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2018T2/about?



Reply

lyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-132191)

is. would love to hear more from you and dive deeper into this

KWAD Reply

lyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-132531)

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

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)

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?

Subscribe!

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.

Reply

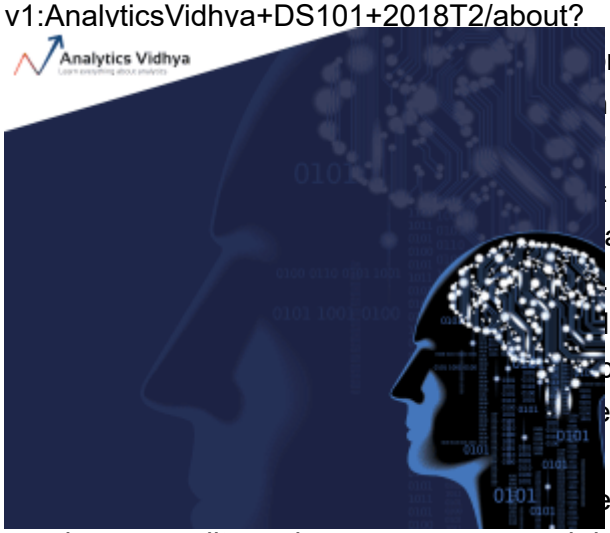
nalyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-133402)

P://WWW.MATHESTATE.COM) **Reply**

alyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-133796)

good article. There are, however, some caveats. I am not a statistician so here is an incomplete list

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2018T2/about?



mmon sense, which should guide all your actions;
that the modifications you make have a small impact.
Remember that the reason the Monty Hall problem works the
draw (3 doors) is disturbed midstream
ave on small samples. Two good examples to Google are
There are others.
extreme values even though both are sometimes referred to as
e valid and eliminating them can be very misleading (There is a
http://www.mathestate.com (http://www.mathestate.com) for an

erta. If your model (like comparison of difference of means) requires normality and you use non-normal data you produce gibberish.

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)

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)

Subscribe!

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 ?

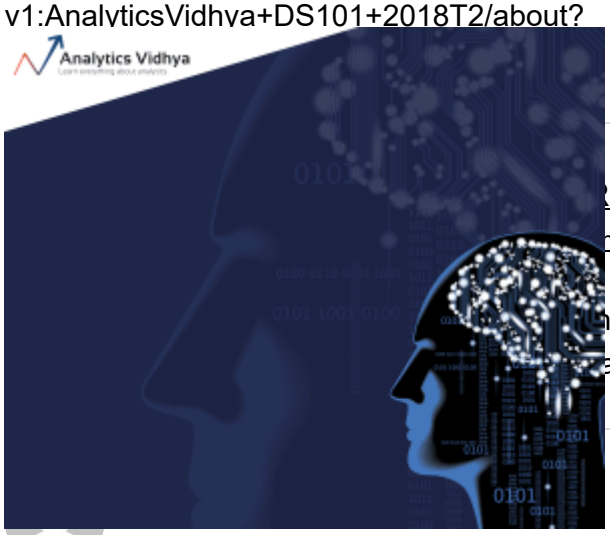
Reply

analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-135043)

Reply

August 20, 2017 at 10:19 pm (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-135117)

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2018T2/about?



RULEUR.COM)

Reply

analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-135526)

ng down each individual concept. Adding some actual code to actical standpoint.

Reply

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

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)
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

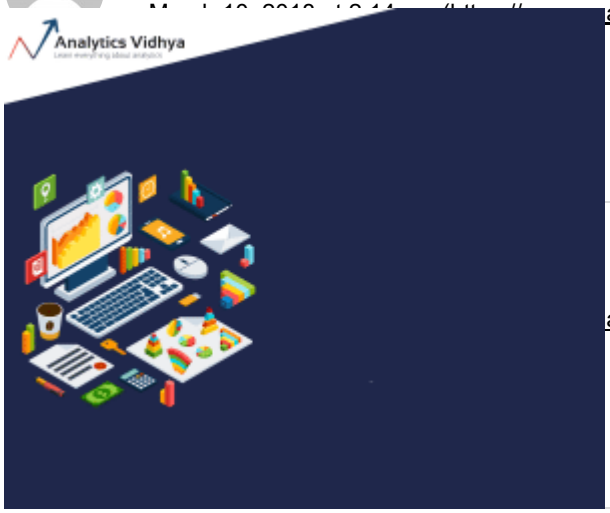
October 25, 2017 at 11:07 am (https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-140750)

Subscribe!

Thanks alot. Great article.

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

alyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-151818)



Reply

alyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-152181)

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2018T2/about?

FABIAN SELIKH Reply

alyticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-152208)

Reply

yticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-152555)

to me

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2&utm_campaign=CV101banner)

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)

Subscribe!

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

Reply

yticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-153442)

n/courses/course-

ut) is a training course on R for big mart sales dataset. A similar

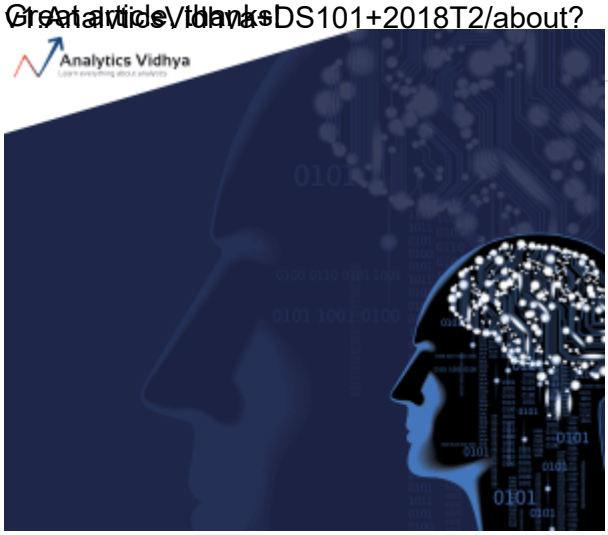
---

Reply

yticsvidhya.com/blog/2016/01/guide-data-exploration/#comment-153471)

(https://trainings.analyticsvidhya.com/courses/course-
v1:AnalyticsVidhya+DS101+2018T2/about?

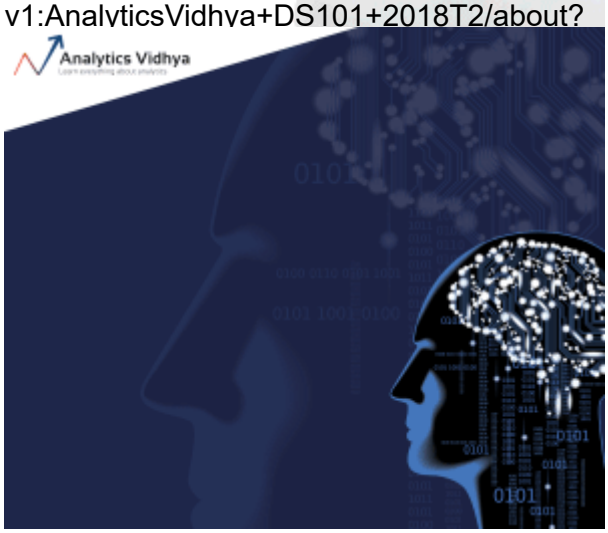
Great article Vidhya. Thanks!



(https://trainings.analyticsvidhya.com/courses/course-
v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?
utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)

Subscribe!

(https://www.analyticsvidhya.com/datahack-summit-2018/?
(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2018T2/about?



(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)

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/)

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/)

Subscribe!

Understanding Support Vector Machine algorithm from examples (along with code)
(https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/)

7 Types of Regression Techniques you should know!
(https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/)

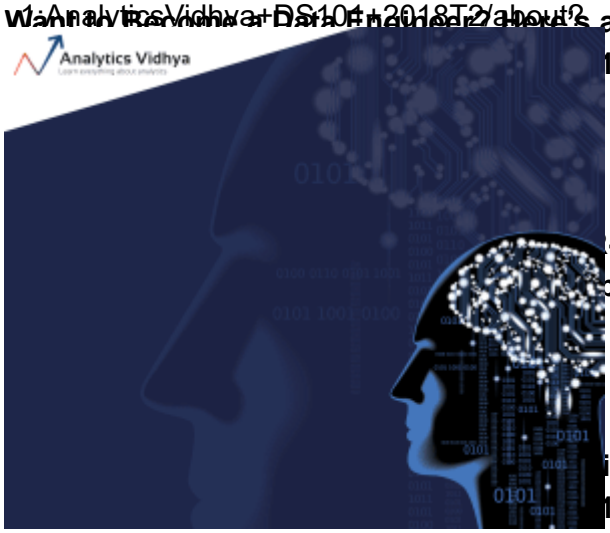
thm (with codes in Python and R)
7/09/naive-bayes-explained/)

te a Time Series Forecast (with Codes in Python)
6/02/time-series-forecasting-codes-python/)

lified (with implementation in Python)
8/03/introduction-k-neighbours-algorithm-clustering/)

(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2018T2/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)

Want to Become a Data Engineer? Here's a Comprehensive List of Resources to get Started
18/11/data-engineer-comprehensive-list-resources-get-



-CNN Algorithm for Object Detection (Part 2 – with Python
olog/2018/11/implementation-faster-r-cnn-python-object-

ies & Reddit Discussions (October 2018)
18/11/best-machine-learning-github-repositories-reddit-
(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+CVDL101+CVDL101_T1/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)

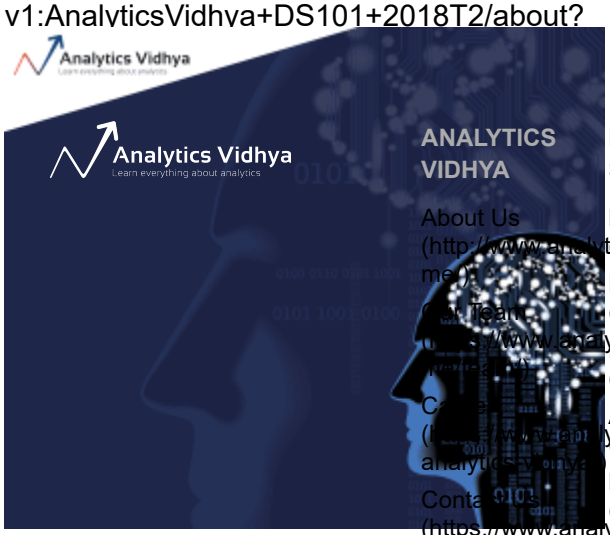
NOVEMBER 1, 2018

An Introduction to Text Summarization using the TextRank Algorithm (with Python implementation)
(https://www.analyticsvidhya.com/blog/2018/11/introduction-text-summarization-textrank-python/)

NOVEMBER 1, 2018

Subscribe!

(http://www.edvancer.in/certified-data-scientist-with-python-s&utm_campaign=AVadsnonfc&utm_content=pythonavad)



(https://trainings.analyticsvidhya.com/courses/course-v1:AnalyticsVidhya+DS101+2018T2/about?utm_source=CV101AVBlogBanner&utm_medium=Stickybanner2utm_campaign=CV101banner)