

[Open in app](#)**uday sai**

4 Followers   About

# How to perform EDA which complements our Data Science story- Part 1


 **uday sai** Apr 13, 2020 · 5 min read



Image Source : <https://d1png.com/png/515073>

---

*EDA is critical aspect of the data science projects. The main purpose of EDA is to analyze the data to get acquainted with the data and gain some first-hand experience. Besides EDA forms the basis of our quest to establish a compelling story with our dataset.*

---

In most of the real-world projects we often collect data from multiple data sources and perform some extensive data cleansing operations to structurize our data which is starting point of our data science story. Following this step we will start to explore our data using some basic operations. Let us explore the main purposes of EDA and go through them one at a time:

---

### *Univariate Analysis*

*Detection of Missing Values and Outliers*

*Checking Assumptions*

*Determining relationships between explanatory variables*

*Assessing the relationship with target variable*

*Preliminary selection of appropriate models*

---

## **Univariate Analysis**

The first step in our EDA is to explore the variables by segregating them into four different sections. This involves categorizing our explanatory variables in to Miscellaneous, Numerical, Categorical and Temporal variables.

The miscellaneous variables are the kind of Id, Customer Name details. These columns does not add any value to our analysis in predicting the target variable and can be treated as a column that can be potentially dropped. However the final decision can be taken only after a discussion with domain expert.

Following the miscellaneous columns we will divide our feature space as numerical variables which are in turn classified as discrete and continuous variables. The univariate analysis is done through the barplots and histograms respectively, from which the spread of the variables are analyzed and documented. Likewise, the categorical variables are divided into ordinal and nominal variables and checked for different levels of cardinality.

A special note in this step is of temporal variables. It is quite common to have some date/date-time/time columns in our dataset and it is not appropriate to segregate them as either of numerical or categorical variable as they often require special attention. To gain these insights we need to identify them and visualize some trends of how target variable alters with respect to these columns.

## **Detection of Missing Values and Outliers**

The foremost step in our EDA is to detect the features which contain null/blank/NaN values. After identifying NaN values it is important to understand why these values are getting generated. There are several reasons for this NaN's like organisation failed to capture these details, user is not interested in filling these details, true NaN's where this feature is not applicable for that particular record, NaN's because of the influence of either target variable or other explanatory variables. I will be discussing how to deal with these NaN's in detail in my other article. For now , it is very important to comprehend data at hand and discuss with the domain experts, business analysts/product analysts why these NaN's are being generated and document the insights.

On the other hand we also need to watch out for outliers. To reinforce the concept of outliers these are the values which a variable take that often falls in the tail-ends of the distribution of the variable or have immense variance with respect to standard deviation. In addition to numerical variables categorical variables also contain outliers in terms of the rare category that these variables encompass in their level of cardinality. However there is no standard way to designate rare categories as it depends on data we

are analyzing. After figuring out these outliers we iterate over the steps of meeting with domain experts and documenting insights.

## Checking Assumptions

When we start exploring a dataset intuitively we will have some assumptions of how the data is being generated based on the business case which we are dealing. To check for these assumptions we need to form some hypothesis and test for these assumptions to eliminate our biases. Please note that this is not equivalent to standard hypothesis test in statistics, rather these are tests based on our intuition and modifying our perception in line with the data stand point. This is a powerful mechanism as it compliments the business insight derivation and predictive analytics.

## Determining relationships between explanatory variables

In most of the real-world data science projects the independent variables has a collective effect on the dependent variables and in order to comprehend these trends we need to understand how the independent variables vary with respect to each other. For illustration let us consider a scenario where the target variable is click or no-click in buying a product on a e-commerce website. This variable might be influenced by multiple target variables like surf-time, activetimeonwebsite, coupons. The surf-time and activetimeonwebsite might be worth analyzing and check if they have a linear relation or exponential relation. The insights generated in this step might forms the basis of which variables might influence the target variable strongly and identify the correlations and co variance between the explanatory variables.

## Assessing the relationship with target variable

This is perhaps the most intriguing part of EDA. In this step we will analyze the relationship between the explanatory variables and target variable. This often involves checking the correlation of the target variable with explanatory variables as it gives us a picture of the variance of the target variable explained by the explanatory variable given one at a time. The refinement of this step is to eliminate the variables those are highly correlated (remember we identified some high correlations in the previous step :) ). This step will give us an opportunity to identify variables that needs to be underscored.

## Preliminary selection of appropriate models

The insights generated from the above steps will form the basis for the model selection. If we are dealing with a regression problem then we will come up with some interesting

insights of the relationships between explanatory variables and target variables, distributions of univariate variables. This might ***suggest the modifications we need to do in implementing OLS regression as it is a rigid model given the number of assumptions.*** On the other hand if we are dealing with a classification model these relationships might give insights whether to choose a ***linear vs. non-linear classification model, models which are robust to outliers etc...***

## Conclusion

This article suggests the theoretical measures of how EDA is performed and how it forms basis for Feature Engineering and Model Selection. ***In a nutshell, EDA is the process of getting acquainted with data from Data Scientist standpoint being benefited by a thorough investigation of data and series of meetings with domain experts.***

In the subsequent articles of this EDA series I will discuss about the tools available (python numpy, pandas, matplotlib, seaborn libraries) to perform EDA and how to perform EDA on near real-world Advance House Price Prediction dataset (borrowed from kaggle website).

## References:

<https://www.stat.cmu.edu/~hseltman/309/Book/chapter4.pdf>

Eda   Machine Learning   Data Science   Python3   Data Visualization

[About](#) [Help](#) [Legal](#)

Get the Medium app



