# EDA Using dlookr

## The Value of Exporatory Data Analysis (EDA)

The first step of any type of statistical analysis or machine learning on a dataset should always be to complete a thorough analysis of the data you are using. (Assuming you are familiar and comfortable with the study design and execution.)

EDA is the practice of using visual and quantitative methods to understand and summarize a dataset without making any assumptions about its contents.

EDA provides the context needed to develop an appropriate model for the problem at hand and to correctly interpret its results. Through a thorough exploration of the data, data scientists will determine whether the data will allow them to produce valid and applicable results.

## Key Steps in EDA

EDA typically relies heavily on visualizing data to assess patterns and identify characteristics that the analyst would not otherwise know to look for. It also takes advantage of a number of quantitative methods to describe the data.

EDA usually involves a combination of the following methods:

1. Univariate visualization and summary statistics for each field in the dataset to understand distribution of each variable.

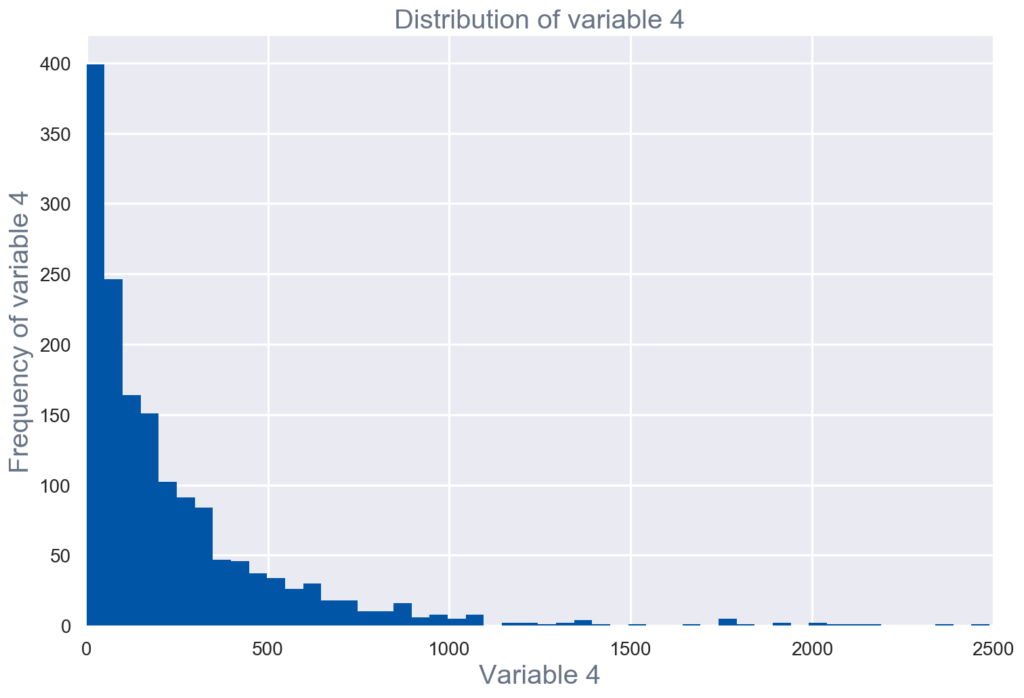


Figure : Univariate visualization of and summary statistics for each field in the raw dataset.

1. Bivariate visualization and summary statistics for assessing the relationship between each variable in the dataset and the target variable of interest.

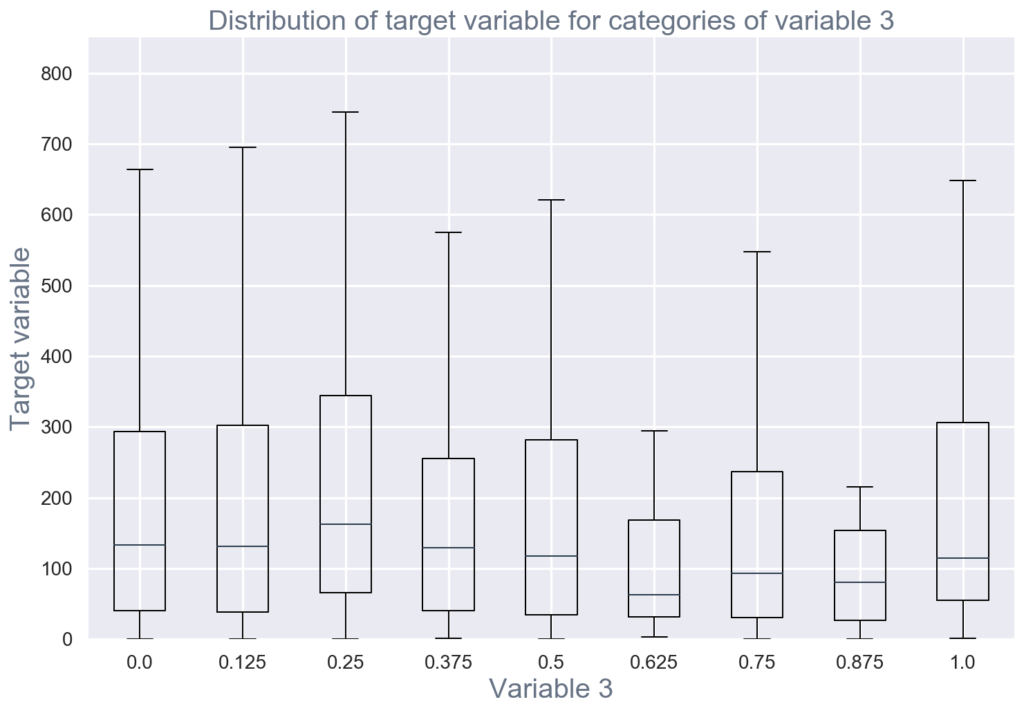


Figure : This Boxplot shows the median and distribution of each variable.

1. Multivariate visualizations to understand interactions between different fields in the data.

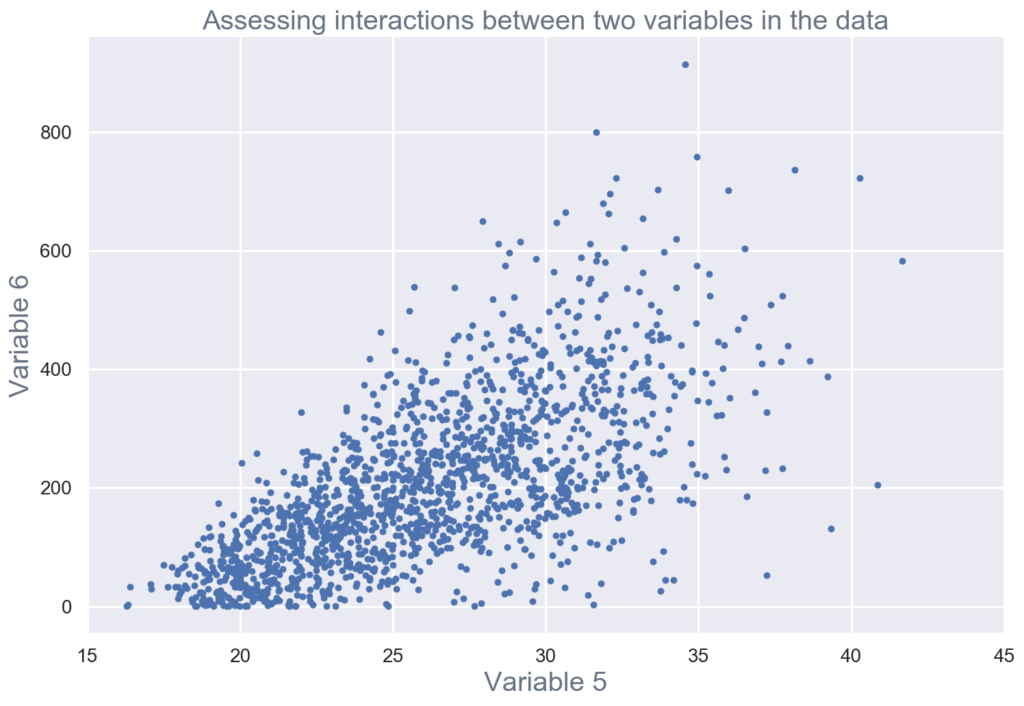


Figure : Scatterplots to visualize correlation between variables.

1. Dimensionality reduction to understand the fields in the data that account for the most variance between observations and allow for the processing of a reduced volume of data.

* Dimension Reduction refers to the process of converting a set of data having vast dimensions into data with lesser dimensions ensuring that it conveys similar information concisely.
* n dimensions of a data set can be reduced to k dimensions (k < n) . These k dimensions can be directly identified (filtered) or can be a combination of dimensions (weighted averages of dimensions) or new dimension(s) that better represent existing multiple dimensions.

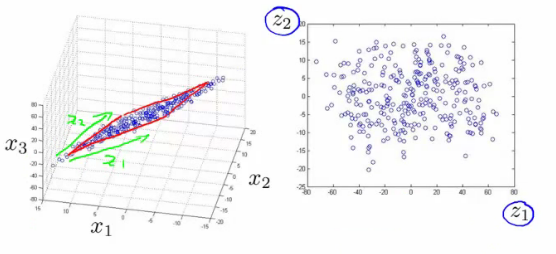


Figure : Dimension Reduction

1. Clustering of similar observations in the dataset into differentiated groupings, which by collapsing the data into a few small data points, patterns of behavior can be more easily identified.

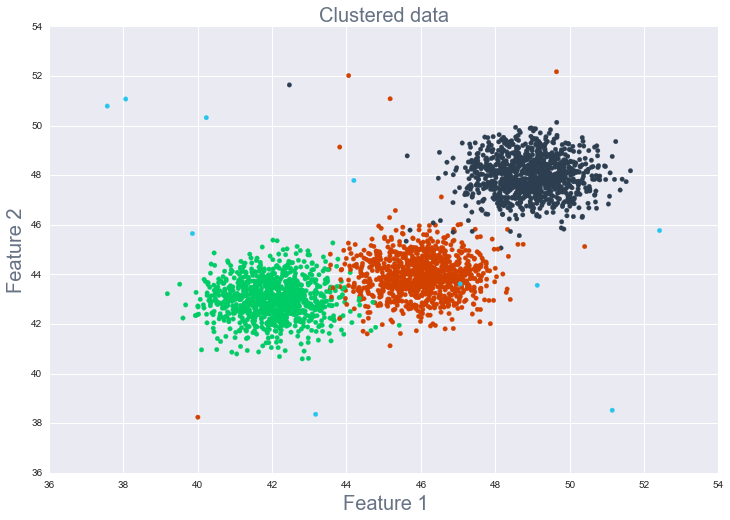


Figure : Clustering

Through these methods, data scientists validate assumptions and identify patterns that will inform the understanding of the problem and model selection, build an intuition for the data to ensure high quality analysis, and validate that the data has been generated in the way it was expected to.

## EDA using R

The ebook [R for Data Science](https://r4ds.had.co.nz/exploratory-data-analysis.html) covers this in great detail and makes use of the dplyr and ggplot2 packages for sorting and visualizing data.

## Using dlookr in R

Download and install the package dlookr from:

* <https://cran.r-project.org/web/packages/dlookr/index.html>
* [User’s Manual](https://cran.r-project.org/web/packages/dlookr/dlookr.pdf)

### Use Cases

* Data Quality Diagnosis
* Exploratory Data Analysis
* Data Transformation

[r demonstration]