Exploratory Data Analysis (EDA) in Data Science

Cheng Peng

Table of Contents

1. What Exploratory Data Analysis

The US National Institute of Standards and Technology (NIST) defines EDA as:

"An approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to maximize insight into a data set, uncover underlying structure, extract important variables, detect outliers and anomalies, test underlying assumptions, develop parsimonious models and determine optimal factor settings."

The term EDA was coined by John Tukey in 1970s. According to Tukey: "It (EDA) is important to understand what you CAN DO before you learn to measure how WELL you seem to have DONE it… Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone –as the first step."

Since the methods of EDA are purely data driven, it does NOT rely on the statistical assumptions about the process of data generation and mathematical formulations.

In classical statistical modeling, EDA is used to discover new patterns in the data and use it to select appropriate statistical methods to make inference about the population. It is also used to identify potential transformation of some variables to meet the assumptions of statistical models in order to make valid inference. The latter is also a typical feature engineering tool in the fields of data science and machine learning.

If the data set to be explored is a population, whatever patterns we found from the data reflect the population. However, if the data is only a sample of a population, whether the patterns revealed in the process of EDA reflect the population characteristics is dependent on whether the data represents the population. In other words, if one does not have good knowledge of the the data generating process or has failed to perform data validation, then EDA is doomed to fail.

1.1. Visual Data Exploration (VDE)

Visual Data Exploration (VDE) uses visualizations to better understand data, and find clues about the tendencies of the data, its quality and to formulate assumptions and the hypothesis of our analysis. VDE is NOT about making fancy visualizations or even aesthetically pleasing ones; the goal is to try and answer questions with data quickly.

The interactive VEA is useful since we can make initial assumptions based on interactive exploratory visualizations, then build some models. We then make visualizations of the model results and tune our models.

Some of the benefits of VDE are

- Visual data exploration can easily deal with highly nonhomogeneous and noisy data.
- Visual data exploration is intuitive and requires no understanding of complex mathematical or statistical algorithms or parameters.
- Visualization can provide a qualitative overview of the data, allowing data phenomena to be isolated for further quantitative analysis.

1.2. The Role Transformations in EDA

EDA begins by understanding the distribution of a variable and how it could be transformed in order to describe a more meaningful source variation. Transformations lie at the heart of EDA.

However, the primary focus in the classical statistical modeling is association analysis. A transformation of variables will make the model interpretation difficult.

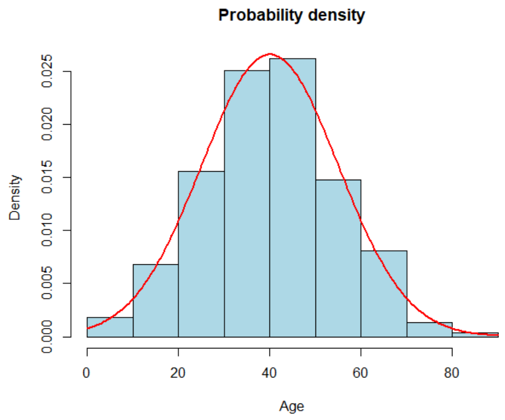
Since most of the data science problems are related to prediction, the interpretation of model parameters are not less important than the accuracy of prediction. The transformation as an important feature engineering tool is widely in data science and machine learning projects.

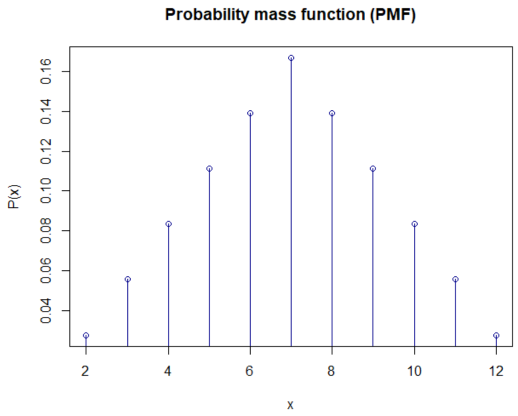
2. Visual Methods and Tools for EDA

Here are some statistical and data science tools that can be considered for various types EDA.

2.1. Univariate visualization

Univariate visualization of each field in the raw dataset, with summary statistics.



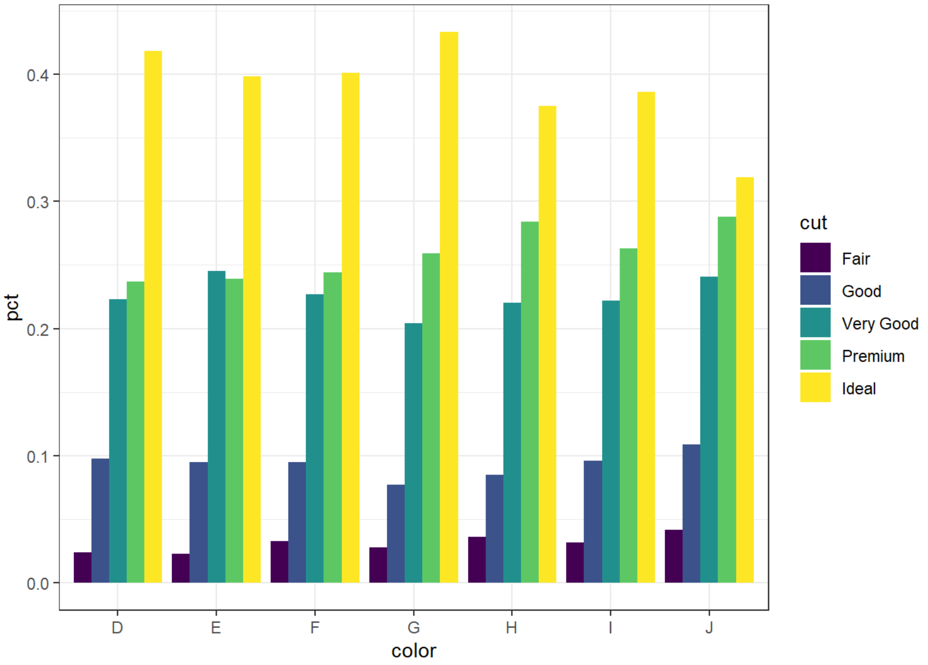


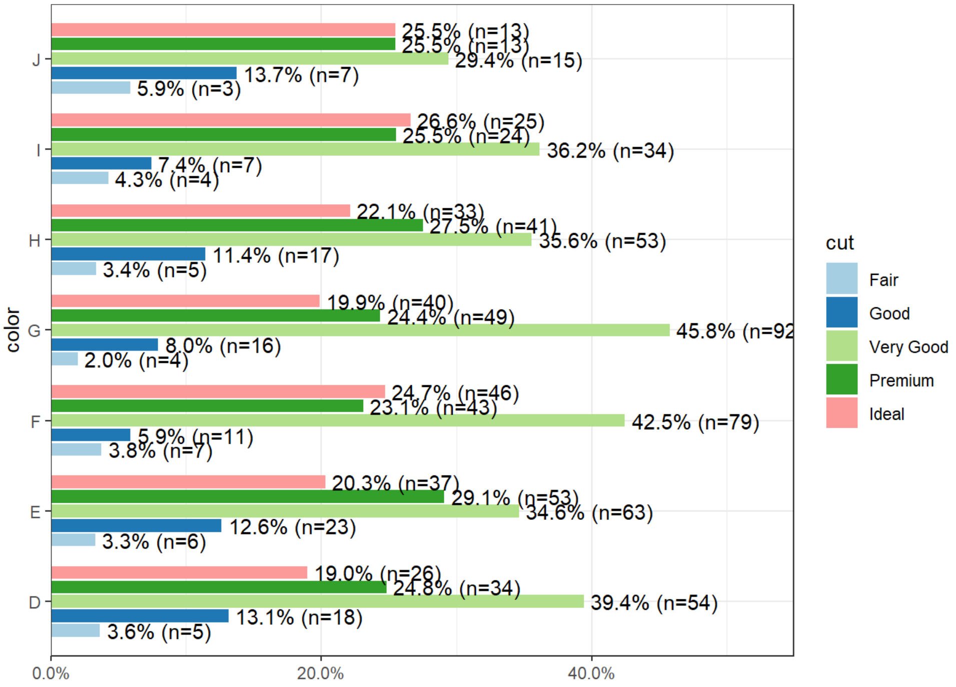
The distributional information of a continuous numerical variable can be visualized using a histogram (or a density curve). For discrete and categorical variables, probability distribution charts such as bar plots and pie charts can be used to display the distributional information.

2.2. Bivariate visualizations

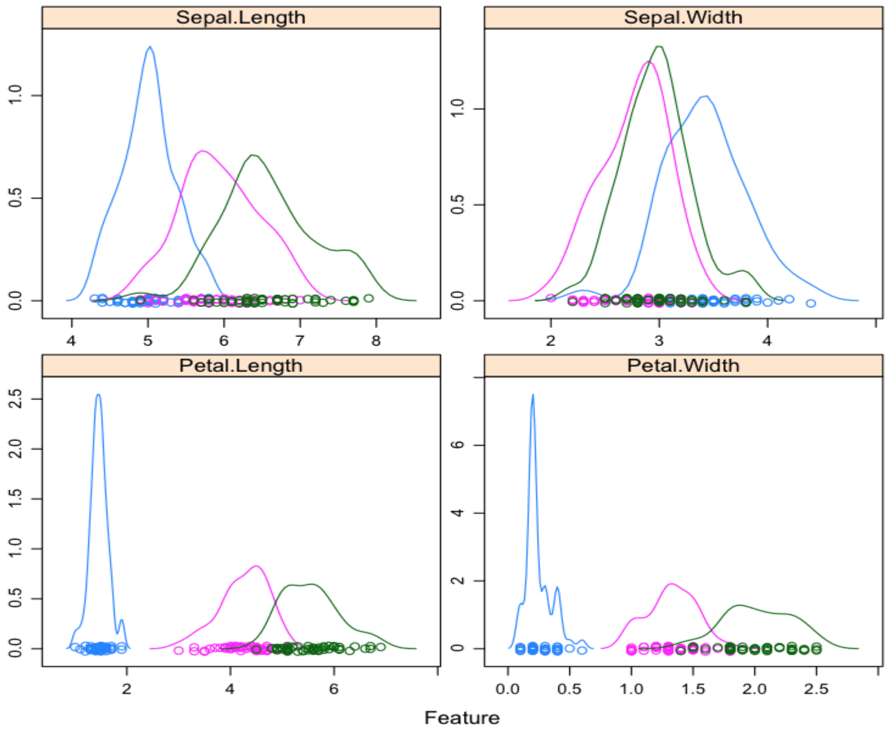
Bivariate visualizations and summary statistics that allow you to assess the relationship between each variable in the dataset and the target variable we're looking at.

- Two categorical variables

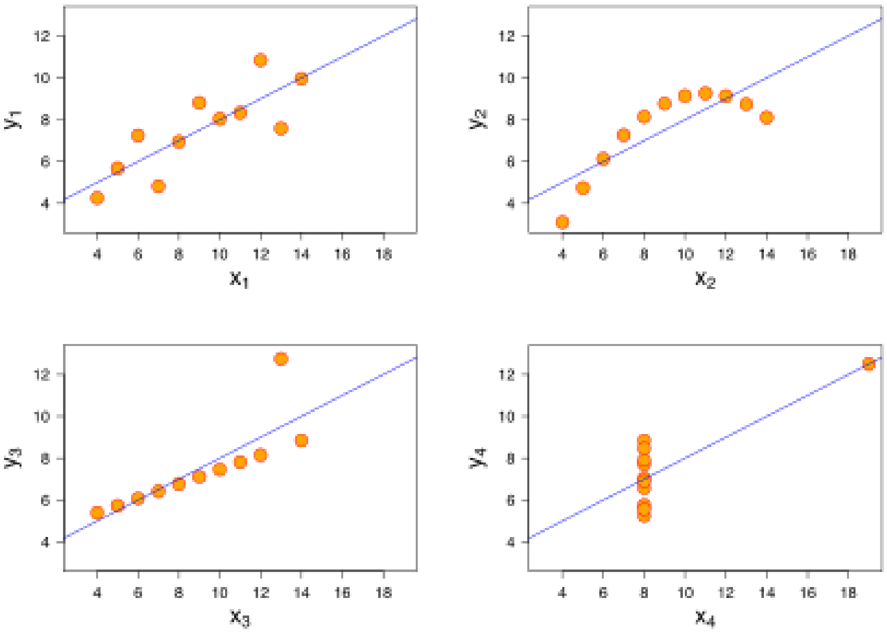




- One continuous variable and one categorical variable



- Two continuous variables

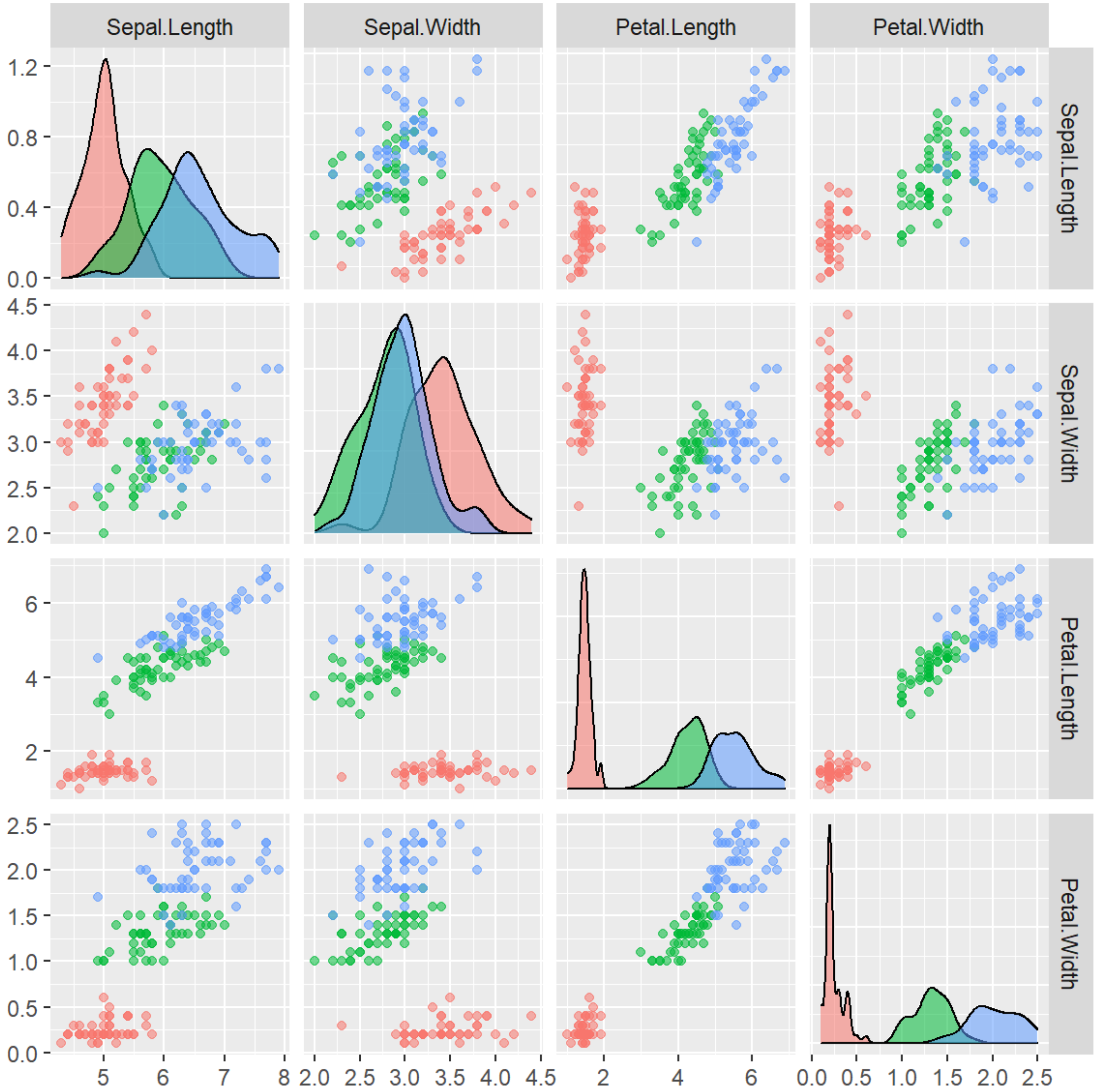


All four sets are identical when examined using correlation coefficient, but they vary considerably when graphed.

2.3. Multivariate visualizations

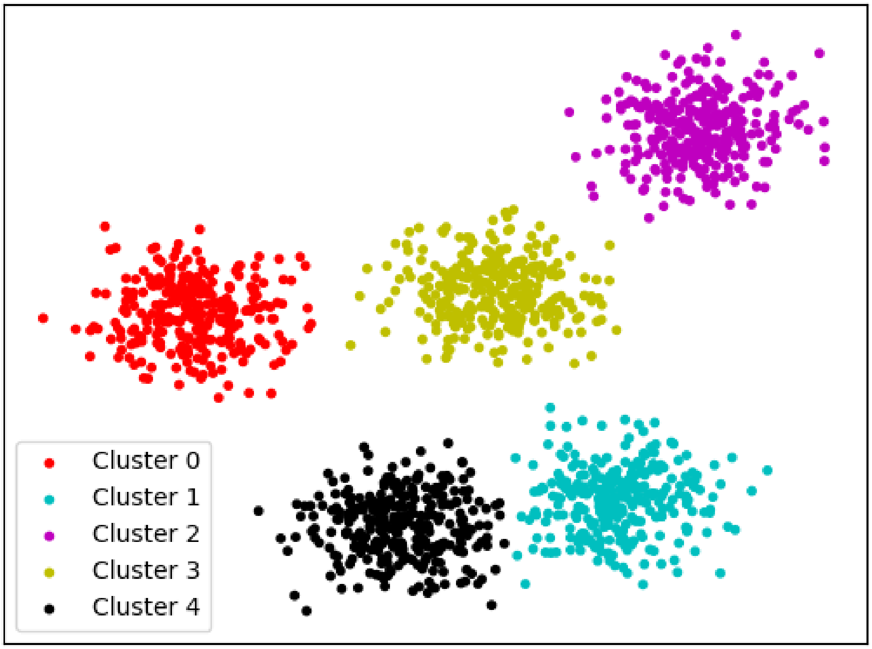
Multivariate visualizations, for mapping and understanding interactions between different variables in the data.

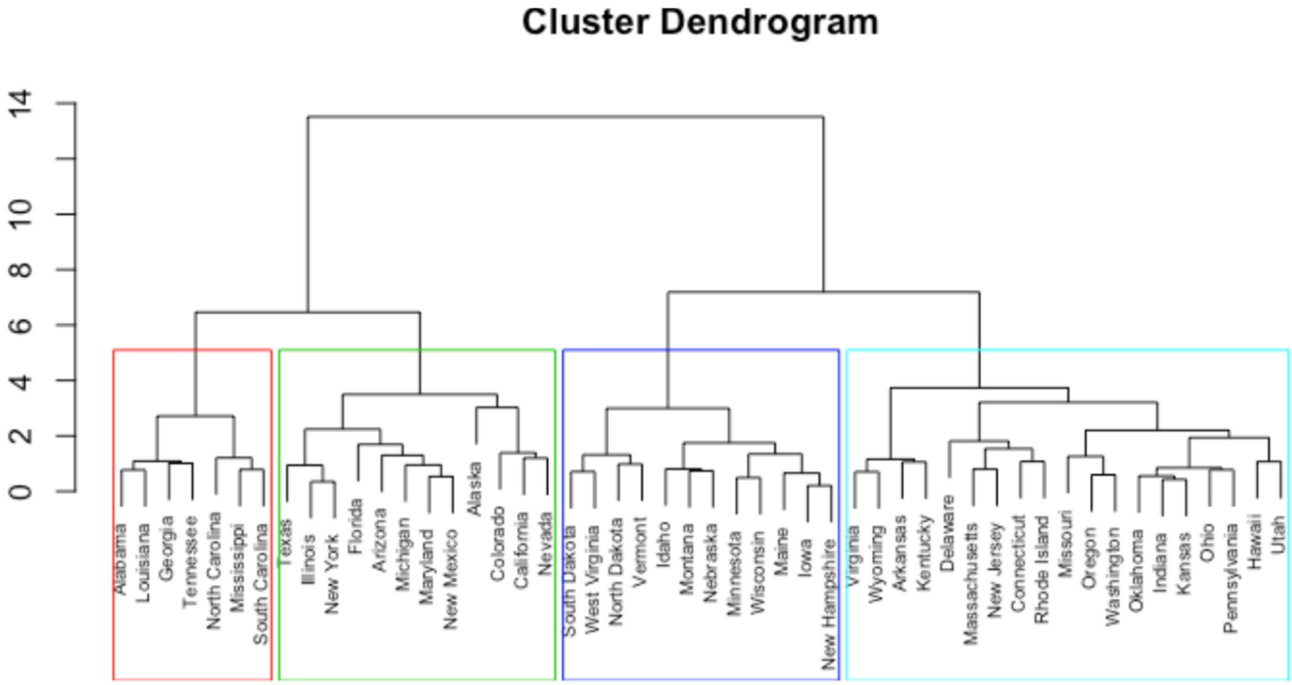
Visualizing relationship between three or more variables can be very hard! The widespread practice is to 2D with various strategies such color schemes, movement, shapes, etc., to display the relationship in high dimensional space.



2.4. Clustering

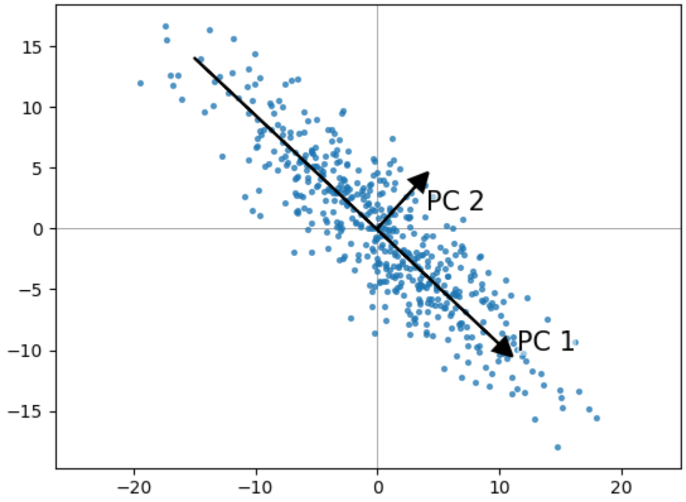
Clustering is an algorithm that groups similar objects into groups called *clusters*. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.





2.5. Dimension reduction

Dimension reduction techniques, which help create graphical displays of high-dimensional data containing many variables.



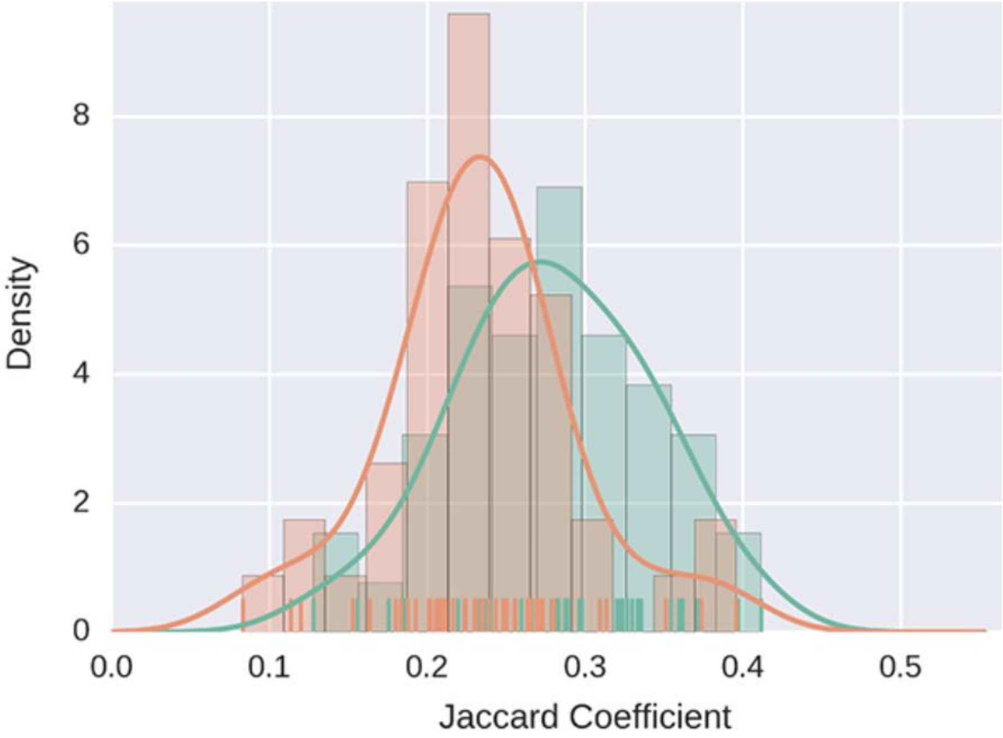
3. Tasks of EDA

Some specific patterns we may want to explore and use for modeling purposes.

3.1. Identify Distributions

The first step in EDA is to develop an understanding on the distributional form of each variable that is under analysis. There are many different methods that can be used for this purpose depending on the types of variables.

- Histograms
- Kernel density estimate
- QQ-plot
- Boxplot
- etc

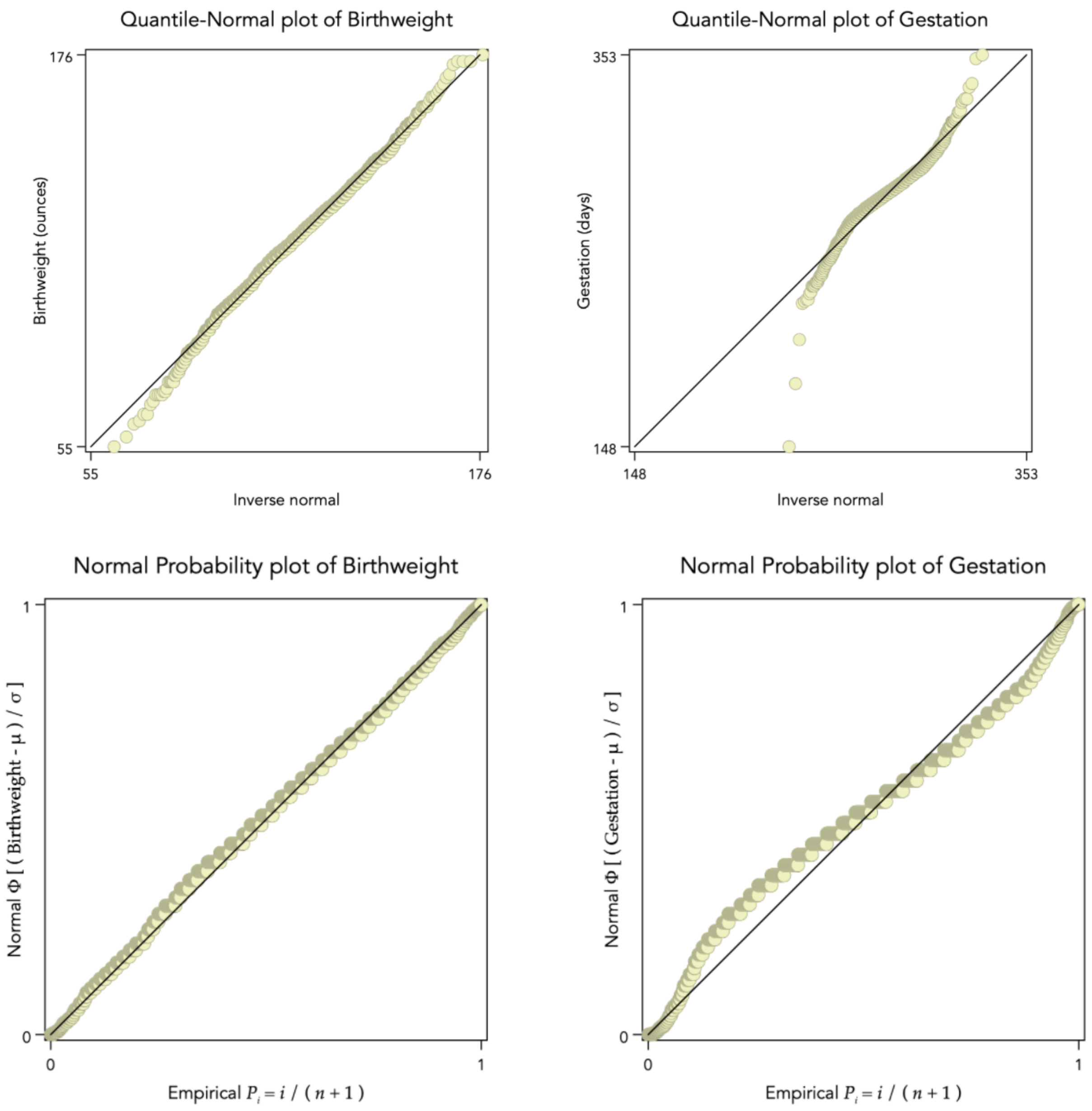


The above kernel density curve estimates the distribution of the population from which the data was taken.

As a cautionary note, the visual density estimators cannot help assessing whether the population is normally distributed.

3.2. Normality Checking

One of the important assumptions in many statistical modeling is the normal distribution. The most relevant EDA methods are the QQ-plot and the normal probability plot.



3.3. Transformations of variables with different domains

If a variable does NOT follow a normal distribution because of its definition, we could perform an appropriate transformation to approximate a normal variable for modeling purpose. The most systematic such transformation is the well-known Box-Cox transformation that takes the following form and is widely used in normal linear regression modeling.

$$y^{(\lambda)} = \begin{cases} \frac{(y+c)^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log_b(y+c) & \text{if } \lambda = 0 \end{cases}$$

Sometimes, we have a variable that takes all real numbers but does follow a normal distribution, John and Draper modified Box-Cox transformation and proposed the following transformation.

$$y^{(\lambda)} = \begin{cases} sign\,\{y\}\,\frac{(|y|+c)^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ sign\,\{y\}\log_b(|y|+c) & \text{if } \lambda = 0 \end{cases}$$

The family of generalized linear models take the advantages of the linear regression model by modeling the probability of the response variable. The probability is in [0, 1]. The links functions are transformations that maps the [0,1] to (-∞, +∞). The well-known logistic regression uses the logit transformation

$$W = \log\left(\frac{P[Y=1]}{1-P[Y=1]}\right).$$