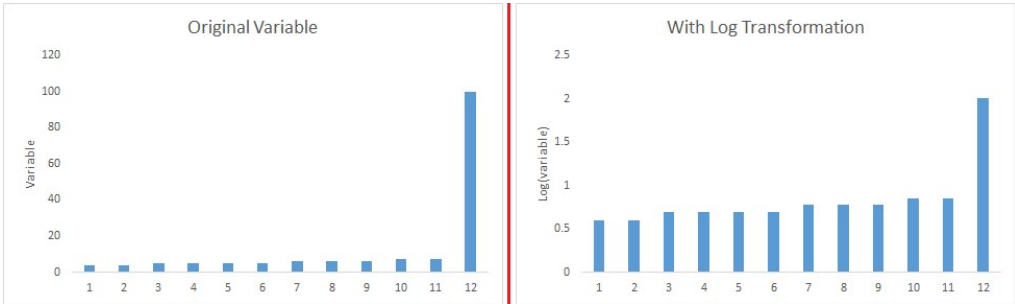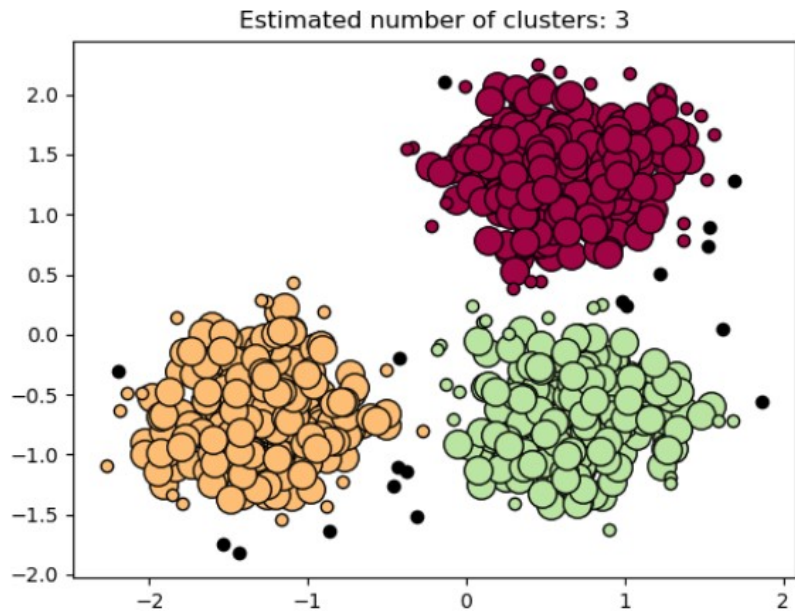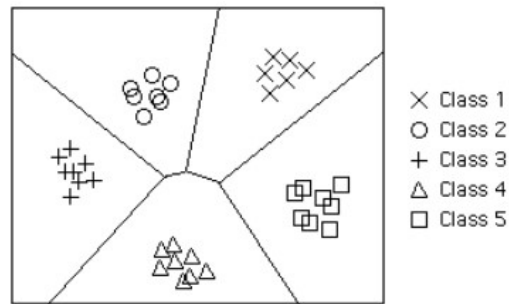| Word	Description
Data Mining	Data mining is a study of extracting useful information from structured/unstructured data taken from various sources. This is done usually for 1. Mining for frequent patterns 2. Mining for associations 3. Mining for correlations 4. Mining for clusters 5. Mining for predictive analysis Data Mining is done for purposes like Market Analysis, determining customer purchase pattern, financial planning, fraud detection, etc
Data Science	Data science is a combination of data analysis, algorithmic development and technology in order to solve analytical problems. The main goal is a use of data to generate business value.
Data Transformation	Data transformation is the process to convert data from one form to the other. This is usually done at a preprocessing step. For instance, replacing a variable x by the square root of x _(see table below)_
Database	Database (abbreviated as DB) is an structured collection of data. The collected information is organised in a way such that it is easily accessible by the computer. Databases are built and managed by using database programming languages. The most common database language is SQL.
Dataframe	DataFrame is a 2-dimensional labeled data structure with columns of potentially different types. You can think of it like a spreadsheet or SQL table, or a dict of Series objects. DataFrame accepts many different kinds of input: • Dict of 1D ndarrays, lists, dicts, or Series • 2-D numpy.ndarray • Structured or record ndarray • A series • Another DataFrame

Table within Data Transformation:

X	SQUARE_ROOT(X)
1	1
4	2
9	3

Dataset	A dataset (or data set) is a collection of data. A dataset is organized into some type of data structure. In a database, for example, a dataset might contain a collection of business data (names, salaries, contact information, sales figures, and so forth). Several characteristics define a dataset's structure and properties. These include the number and types of the attributes or variables, and various statistical measures applicable to them, such as standard deviation and kurtosis.
Dashboard	Dashboard is an information management tool which is used to visually track, analyze and display key performance indicators, metrics and key data points. Dashboards can be customised to fulfil the requirements of a project. It can be used to connect files, attachments, services and APIs which is displayed in the form of tables, line charts, bar charts and gauges. Popular tools for building dashboards include Excel and Tableau.
DBScan	DBSCAN is the acronym for **Density-Based Spatial Clustering of Applications with Noise**. It is a clustering algorithm that isolates different density regions by forming clusters. For a given set of points, it groups the points which are closely packed. The algorithm has two important features: • distance • the minimum number of points required to form a dense region The steps involved in this algorithm are: • Beginning with an arbitrary starting point it extracts the neighborhood of this point using the distance • If there are sufficient neighboring points around this point then a cluster is formed • This point is then marked as visited • A new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise • This process continues until all points are marked as visited The below image is an example of DBScan on a set of normalized data points: Estimated number of clusters: 3

Decision Boundary	In a statistical-classification problem with two or more classes, a decision boundary or decision surface is a hypersurface that partitions the underlying vector space into two or more sets, one for each class. How well the classifier works depends upon how closely the input patterns to be classified resemble the decision boundary. In the example sketched below, the correspondence is very close, and one can anticipate excellent performance. Here the lines separating each class are decision boundaries.
Decision Tree	Decision tree is a type of supervised learning algorithm (having a pre-defined target variable) that is mostly used in classification problems. It works for both categorical and continuous input & output variables. In this technique, we split the population (or sample) into two or more homogeneous sets (or sub-populations) based on most significant splitter / differentiator in input variables.