**CRISP-DM** framework**- Cross Industry Standard Process for Data Mining.**

Analytics problem solving involves multiple steps like

* Business understanding
* Data understanding
* Data Preparation
* Data Modeling
* Model Evaluation
* Model Deployment

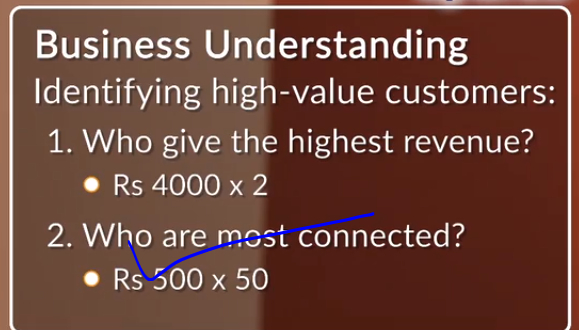
Completing a typical analytics project may take several months, and thus it is important to have a structure for it.

For a data analyst, **understanding the business** and its specific problems is of utmost importance

1. **Business Understanding**

* **Determine the business objectives clearly**
* **Determine the goals of data analysis**

**Telecom *C*hurn, identify most valued customers**



1. **Data Understanding**

* **Collect relevant data**

First, you need to identify and collect the right set of data sets that can be used for the analysis. They can be available within the firm, or you may have to collect the data from other sources such as open source repositories or government data sets. For example, to understand the investments across sectors globally, you can used the structured CrunchBase data set. You could also collect unstructured data, like news articles or twitter feeds about acquisitions, or semi-structured data such as annual reports, if it is relevant

* **Describe datasets**

for explicit information: Once you have identified the data set, you need to describe its contents and explore insights to better understand the data and its business implications. To describe the CrunchBase data, we can create a data dictionary that lists down the types of variables (e.g. sectors, company names, etc.), the number of records, and the types of analysis.

* **Explore data by plotting graphs**

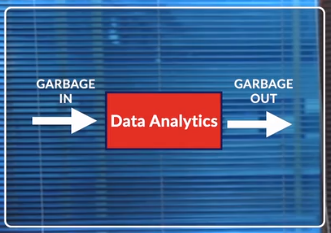
– for implicit insights: To explore data, you can plot simple graphs on Excel/R, e.g. to understand the range of funding received by companies

* **Check data quality**

– to remove errors: Once you have understood the data structure, you can next examine the quality of data and address various factors: • Is the data complete, does it cover all the cases and records? • Is the data correct, or does it contain errors and, if there are errors, how common are they? • Are there missing values in the data? If so, how are they represented? • Where do the missing values occur, and how common are they?

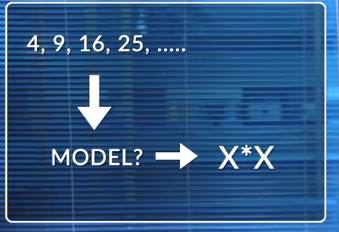
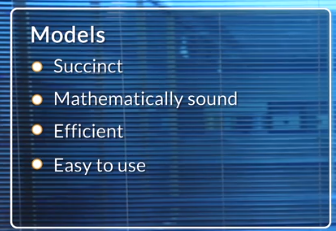
1. **Data Preparation (50% -80%)**
2. **Select relevant data**
3. **Integrate data**
4. **Clean data**
5. **Construct Data: Derive new features**
6. **Format Data**

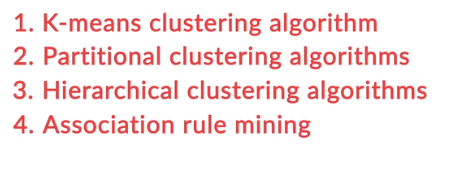
* Across projects, data analysts spend around **50-80% of the time on data cleaning and preparation**
* It is very crucial for data analysis

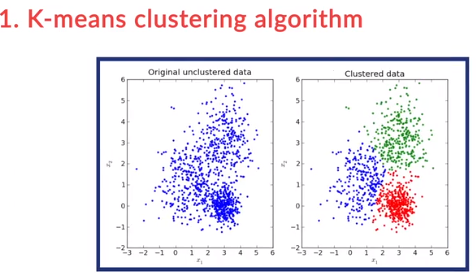


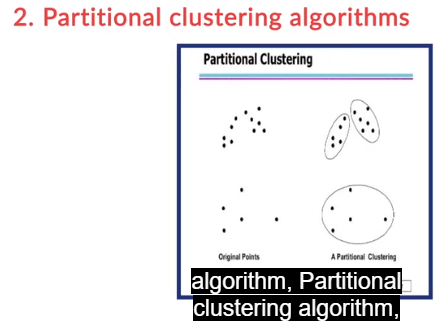
* Data which is of interest is selected. It may be spread across different files or sources.
* Then, you integrate data from these multiple sources. For example, in the CrunchBase data, information is spread across different files and we integrate them together to solve a particular business problem.
* After data integration comes the data preparation stage. Missing value treatment, outlier treatment, and removing the erroneous values are a few major components of the data cleaning process.
* Constructing the data is the next step, which involves the creation of new features originally not present in the data set. The purpose of this step is to increase the information we get from the data and to reduce the number of variables originally present in the data set.
* Finally, format the data. In this step, no changes are made to the data in the data set, but to its structure. The variables present in the data set are set to the correct format as required by the analysis tool.

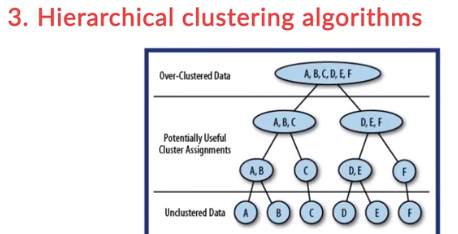
1. **Data Modeling (Heart of Data Analysis)**

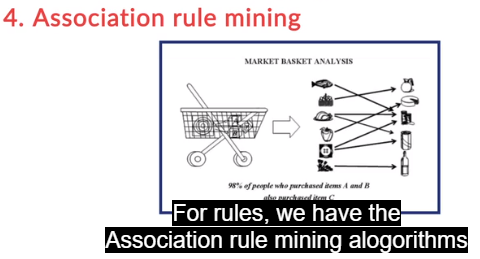
* Modeling is the heart of data analytics. One can think of a model as a black box which takes relevant data as input and gives an output you are interested in
* 
* 
* **Clustering Algorithms**

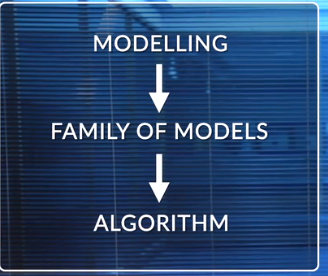












1. **Model Evaluation**

* The predictive models can be tested to assess their effectiveness in solving the problem. This is the fifth stage of the framework – model evaluation. Modelling and evaluation together is an iterative process in which the models are tweaked until satisfactory evaluation results are obtained.

1. **Model Deployment**

* This is the last stage of the framework, where the model is translated into a business strategy. Business data is fed into the model and the model results are used to inform business decisions on an on-going basis. The CRISP-DM framework does not end at the last stage of model deployment. The important thing to note is that CRISP-DM is an iterative process. For example, your data understanding can enhance your business understanding. Similarly, after model evaluation, if the model does not perform great, you will have to go back to the data preparation stage, and then develop the model again.

