Analytics in a Big Data World: The Essential Guide to Data Science and Its Applications - Book by Bart Baesens, Wiley Publications, 2004

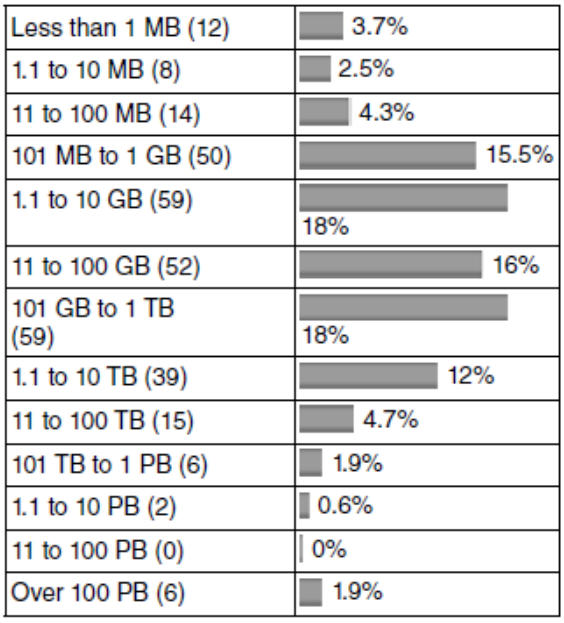
Module 1 - Big Data and Analytics

Example applications, Basic Nomenclature, Analytics Process Model, Analytical Model Requirements, Types of Data Sources, Sampling, Types of Data Elements, Data Exploration and Exploratory Statistical Analysis, Missing Values, Outlier Detection and Treatment, Standardizing Data, Categorization.

Name	Equal To	Size(In Bytes)
Bit	1 Bit	1/8
Nibble	4 Bits	1/2
Byte	8 Bits	1
Kilobyte	1024 Bytes	10^3
Megabyte	1, 024 Kilobytes	10^6
Gigabyte	1, 024 Megabytes	10^9
Terrabyte	1, 024 Gigabytes	10^{12}
Petabyte	1, 024 Terabytes	10^{15}
Exabyte	1, 024 Petabytes	10^{18}
Zettabyte	1, 024 Exabytes	10^{21}
Yottabyte	1, 024 Zettabytes	10^{24}

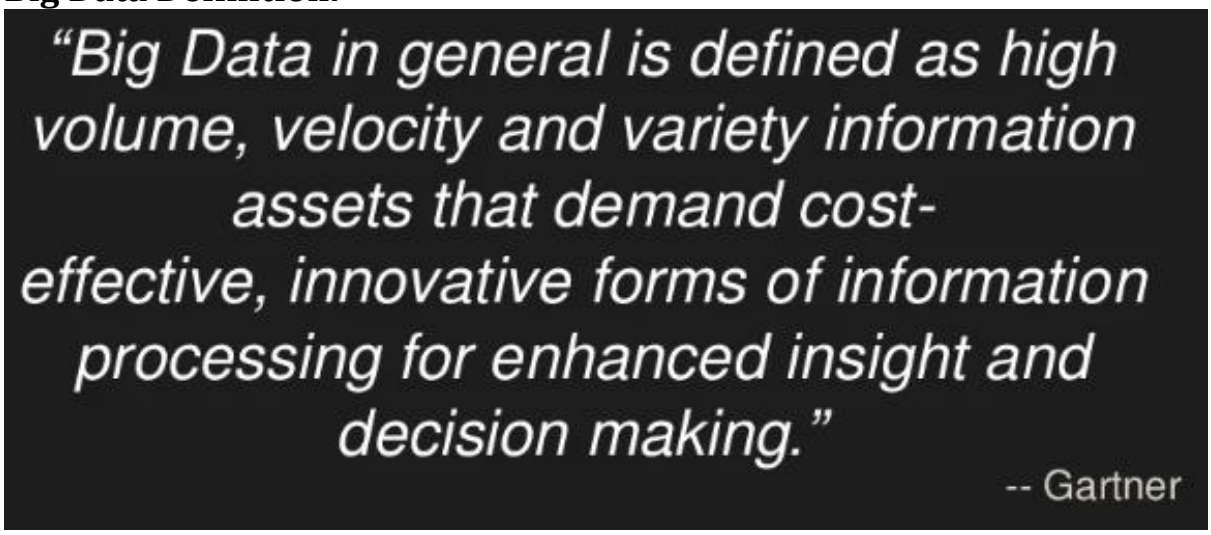
Introduction to Big Data and Analytics:

- Data are everywhere.
- IBM projects that every day we generate 2.5 quintillion ($2.5x10^{18}$) bytes of data.
- Gartner projects that, 85% of Fortune 500 organizations will be unable to exploit big data for competitive advantage and about 4.4 million jobs will be created around big data.
- This strongly indicates the ubiquity of big data and the strong need for analytical skills and resources because, as the data piles up, managing and analysing these data resources in the most optimal way becomes critical in creating competitive advantage and strategic leverage.
- As per the surveys, the average size of the data sets analysed during the recent past is in the range 40 to 50 Giga-Byte (GB). This clearly shows the quick increase in size of data that analysts are working on.
- A main obstacle to fully harnessing the power of big data using analytics is the lack of skilled resources and "data scientist" talent required to exploit big data.
- In another survey it was projected that the need for analytics/big data/data mining/data science education is emerging.



Details of largest data sets analysed in the recent past

Big Data Definition:



- **Big Data** is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently.
- **Big Data** refers to the large, diverse sets of information that grow at ever-increasing rates. It encompasses the volume of information, the velocity or speed at which it is created and collected, and the variety or scope of the data points being covered.

Big Data - Basic Nomenclature

- In order to start doing analytics, some basic vocabulary needs to be defined.
- A first important concept here concerns the ***basic unit of analysis***.
- ***Customers***, ***roles of customers, account behaviour***, ***transactions*** etc. can be considered from various perspectives, as basic units for analysis purposes.
- ***Customer lifetime value (CLV)*** can be measured for either individual customers or at the household level.
- Another alternative is to look at ***account behaviour***. For example, consider a credit scoring exercise for which the aim is to predict whether the applicant will default on a particular mortgage loan account.
- The analysis can also be done at the ***transaction level***. For example, in insurance fraud detection, one usually performs the analysis at insurance claim level. Also, in web analytics, the basic unit of analysis is usually a web visit or session.
- It is also important to note that ***customers can play different roles***. For example, parents can buy goods for their kids, such that there is a clear distinction between the payer and the end user. In a banking setting, a customer can be primary account owner, secondary account owner, main debtor of the credit, co-debtor, guarantor, and so on. It is very important to clearly distinguish between those different roles when defining and/or aggregating data for the analytics exercise.

Big Data – Example Applications

- Analytics is everywhere and strongly embedded into our daily lives.
- The list includes
  1. Behavioural scoring model
  2. Telephone service
  3. Social Media Ads
  4. Twitter posts
  5. Supermarkets
  6. Fraud Detection
- As a result of a **response modelling analytical exercise**, given one's characteristics and previous **purchase behaviour**, it can be predicted what likely going to be purchased by oneself.
- As a result of **behavioural scoring model**, (model that will look at, among other things, checking account balance from the past 12 months and credit payments of a customer during that period, together with other kinds of information available in the bank) it is easy to predict whether a customer will default on his/her loan during the next year. Bank needs to know this for provisioning purposes.
- Also, telephone services provider analyses calling behaviour and account information of users to predict whether he/she will change plan/provider during the next three months.
- As we login to Facebook page, the social ads appearing are based on analysing all information (posts, pictures, my friends and their behaviour, etc.) available to Facebook.
- Our Twitter posts are analysed by **social media analytics** to understand both the subject of our tweets and the sentiment of them.
- As we enter into the **supermarket**, our loyalty card is scanned first, followed by all our purchases. This will be used by supermarket to analyse our market basket, which will help it decide on product bundling, next best offer, improving shelf organization, and so forth.
- Our credit card provider, uses a **fraud detection model** to see whether it was a legitimate transaction. When we receive our credit card statement later, it will be accompanied by various vouchers that are the result of an analytical customer segmentation exercise to better understand my expense behaviour.

To summarize, the relevance, importance, and impact of analytics are now bigger than ever before and, given that more and more data are being collected and that there is strategic value in knowing what is hidden in data, analytics will continue to grow.

Some examples of how analytics is applied in various settings is detailed below.
(Descriptive, Diagnostic, Predictive, Prescriptive Analytics).

Marketing:

- Response Modelling - The task is to classify the customers who will respond to the next marketing campaign on the basis of information collected about the customers.
- Net Lift Modelling - Directly models the incremental impact of a treatment (such as a direct marketing action) on an individual's behaviour. Net lift modelling tries to optimize the difference in response of the customers who have received the campaign with the customers who didn't receive the campaign.
- Retention Modelling - Utilize historical first-party data to predict future behaviour.
- Market Basket Analysis - To increase sales by better understanding customer purchasing patterns.
- Recommender Systems - It is an algorithm that suggests relevant items to users based on rating or preference.
- Customer Segmentation - The process by which you divide your customers into segments up based on common characteristics – such as demographics or behaviours, so you can market to those customers more effectively.

Risk management:

- Credit Risk Modelling - Used by a bank to estimate a credit portfolio
- Market Risk Modelling - Estimation of losses on a portfolio arising from the movement of market prices.
- Operational Risk Modelling - Banks and financial firms use to gauge their risk of loss from operational failings.
- Fraud Detection - Activities undertaken to prevent money or property from being obtained through false evidences.

Government:

- Tax Avoidance – Income Tax, Property Tax, Professional Tax, Disproportionate assets
- Social Security Fraud - Providing false information or evidence for a benefit claim, impersonation, Misuse and trafficking of Social Security numbers and cards by people or businesses
- Money Laundering – Concealing financial assets so they can be used without detection of the illegal activity that produced them.
- Terrorism Detection - Aimed at detecting suspicious users on the Internet by the content of information they access.

Web:

- Web Analytics - Process of analysing the behaviour of visitors to a website. This involves tracking, reviewing and reporting data to measure web activity
- Social Media Analytics - The process of collecting and analysing audience data shared on social networks to improve an organization's strategic business decisions.

- Multivariate Testing - Techniques for testing a hypothesis in which multiple variables are modified. Websites and mobile apps are made of combinations of changeable elements.

Logistics:

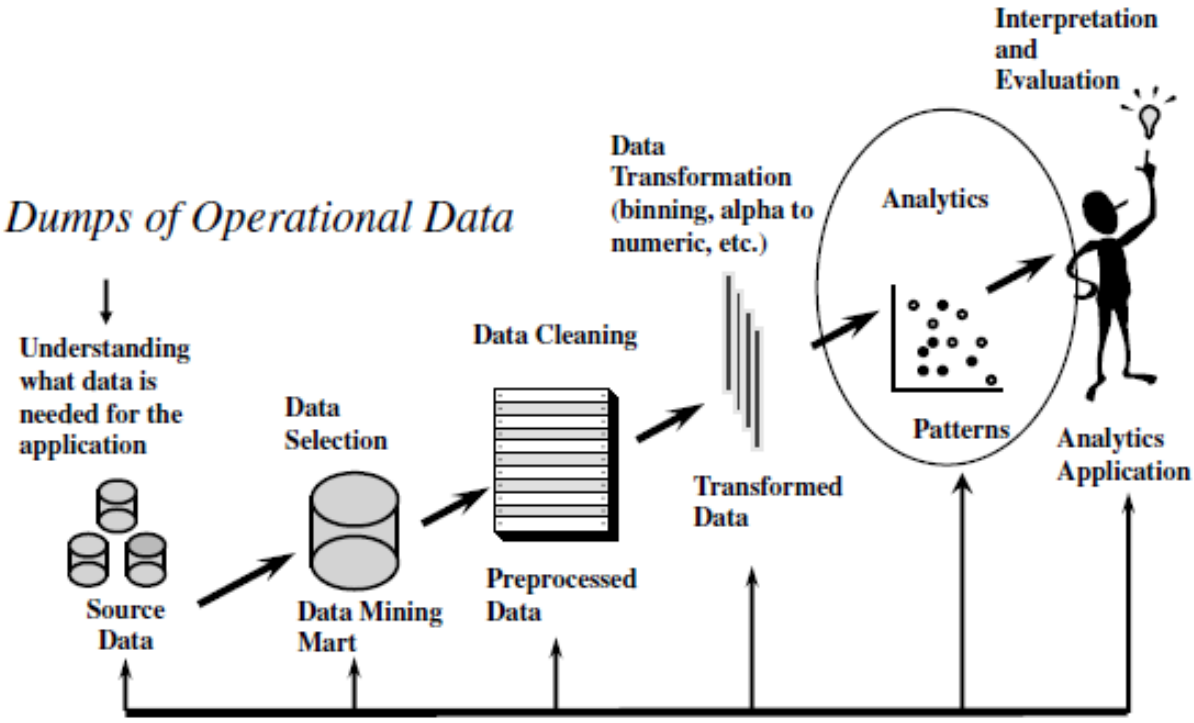
- Demand Forecasting - The process of making future estimations in relation to customer demand over a specific period.
- Supply Chain Analytics - The processes organizations use to gain insight and extract value from the large amounts of data associated with the procurement, processing and distribution of goods.

Others:

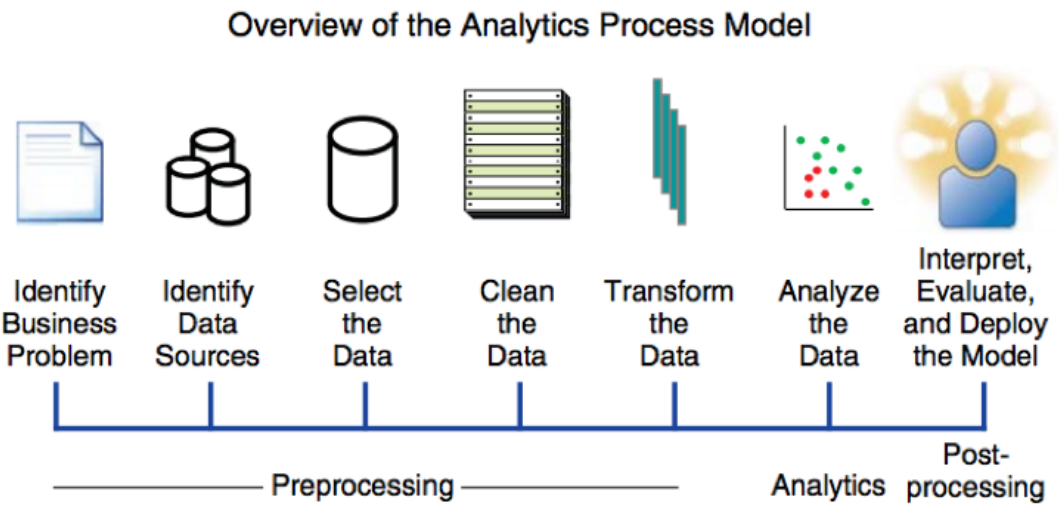
- Text Analytics Business Process Analytics Sentiment Analytics

Big Data – Analytics Process Model

- Following figure gives a high-level overview of the analytics process model.



OR



Steps involved in Analytics Process Model:

1. Define the business problems to be solved
2. All source-data need to be identified
3. All data to be gathered as a pool
4. Data cleaning
5. Analytical model estimation
6. Model interpretation and evaluation

Step 1: A **thorough definition of the business problem** to be solved with analytics is needed. The objective of applying analytics needs to be unambiguously defined.

Step 2: Next, all **source data** that could be of potential interest need to be identified. This is a very important step, as data is the key ingredient to any analytical exercise and the selection of data will have a deterministic impact on the analytical models that will be built. The golden rule here is: **the more data, the better**! The analytical model itself will later decide which data are relevant and which are not for the task at hand.

Step 3: All data will then be gathered in a staging area, which could be, for example, a data mart or **data warehouse**. Some basic exploratory analysis can be considered here using, for example, online analytical processing (OLAP) facilities for multi-dimensional data analysis (e.g., roll-up, drill down, slicing and dicing).

Step 4: The collected **data is cleaned** and **transformed** to get rid of any inconsistencies, such as missing values, outliers and duplicate data. Additional transformations may also be considered, such as alphanumeric to numeric coding, geographical aggregation, and so forth.

Step 5: In the analytics step, an **analytical model** will be estimated on the pre-processed and transformed data. Different types of analytics can be considered here (e.g., to do churn prediction, fraud detection, customer segmentation, market basket analysis).

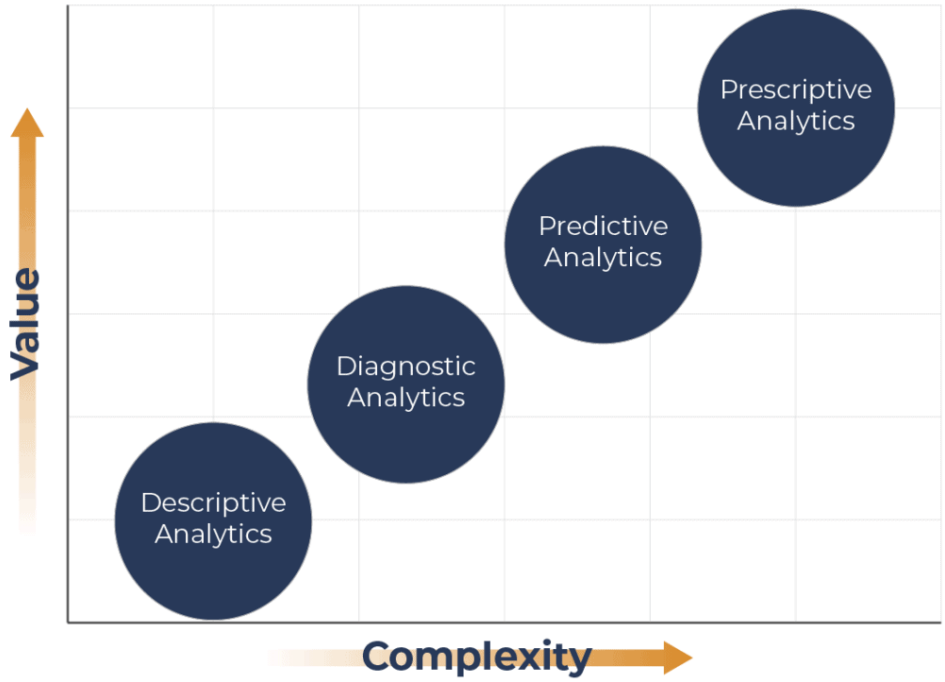
Step 6: Finally, once the model has been built, it will be **interpreted and evaluated** by the business experts.

- It is important to note that the process model outlined in the above figure is iterative in nature, in the sense that one may have to go back to previous steps during the exercise.
- For example, during the analytics step, the need for additional data may be identified, which may necessitate additional cleaning, transformation, and so forth. Also, the most time-consuming step is the data selection and pre-processing step; this usually takes around 80% of the total efforts needed to build an analytical model.
- Usually, many trivial patterns will be detected by the model. For example, in a market basket analysis setting, one may find that Coke and Biscuits are often purchased together.
- These patterns are interesting because they provide some validation of the model. But of course, the key issue here is to find the unexpected yet interesting and actionable patterns (sometimes also referred to as *knowledge diamonds*) that can provide added value in the business setting.
- Once the analytical model has been appropriately validated and approved, it can be put into production as an analytics application (e.g., decision support system, scoring engine).
- It is important to consider here how to represent the model output in a user-friendly way, how to integrate it with other applications (e.g., campaign management tools, risk engines), and how to make sure the analytical model can be appropriately monitored and back-tested on an ongoing basis.

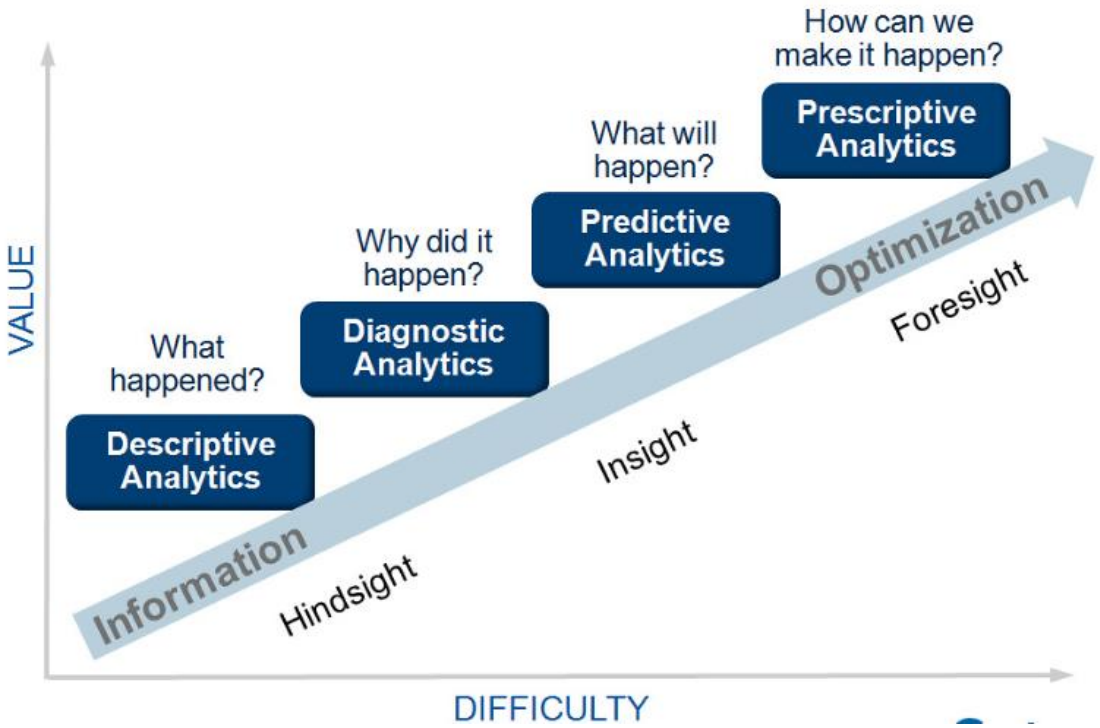
Big Data – Analytical Model Requirements

Analytical models are mathematical models that have a closed form solution, i.e. the solution to the equations used to describe changes in a system can be expressed as a mathematical analytic function.

A Good analytical model must be able to explain some facet of the business problem. Purpose of descriptive models is to extract the patterns in the data that are non-trivial, unknown, potentially useful and actionable.



As you move from descriptive to prescriptive analytics, each model offers increasing value to an organisation. But, at the same time, they increase in complexity.



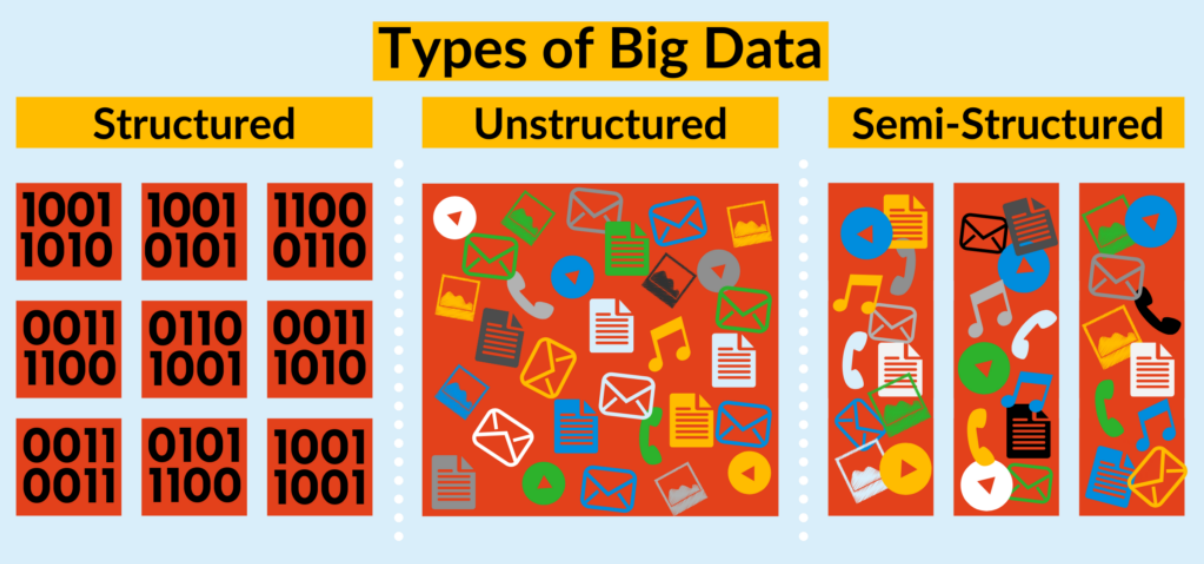
Business relevance
Statistical performance
Interpretability
Justifiability
Operationally efficient
Economical
Local and international regulation and legislation.

- A good analytical model should satisfy several requirements, depending on the application area.
- A first critical success factor is ***business relevance***. The analytical model should actually solve the business problem for which it was developed. It makes no sense to have a working analytical model that got side-tracked from the original problem statement. In order to achieve business relevance, it is of key importance that the business problem to be solved is appropriately defined, qualified, and agreed upon by all parties involved at the outset of the analysis.
- A second criterion is ***statistical performance***. The model should have statistical significance and predictive power. How this can be measured will depend upon the type of analytics considered. For example, in a classification setting (churn, fraud), the model should have good discrimination power. In a clustering setting, the clusters should be as homogenous as possible.
- Depending on the application, analytical models should also be interpretable and justifiable. ***Interpretability*** refers to understanding the patterns that the analytical model captures. This aspect has a certain degree of subjectivism, since interpretability may depend on the business user's knowledge. In many settings, however, it is considered to be a key requirement. For example, in credit risk modelling or medical diagnosis, interpretable models are absolutely needed to get good insight into the underlying data patterns. In other settings, such as response modelling and fraud detection, having interpretable models may be less of an issue.
- ***Justifiability*** refers to the degree to which a model corresponds to prior business knowledge and intuition. For example, a model stating that a higher debt ratio results in more creditworthy clients may be interpretable, but is not justifiable because it contradicts basic financial intuition.
- Note that both interpretability and justifiability often need to be balanced against statistical performance. Often one will observe that high performing analytical models are incomprehensible and black box in nature. A popular example of this is neural networks, which are universal approximators and are high performing, but offer no insight into the underlying patterns in the data. On the contrary, linear regression models are very transparent and comprehensible, but offer only limited modelling power.
- Analytical models should also be ***operationally efficient***. This refers to the efforts needed to collect the data, pre-process it, evaluate the model, and feed its outputs to the business application (e.g., campaign management, capital calculation). Especially in a real-time online scoring environment (e.g., fraud detection) this may be a crucial characteristic. Operational efficiency also entails the efforts needed to monitor and back-test the model, and re-estimate it when necessary.
- Another key attention point is the ***economic cost*** needed to set up the analytical model. This includes the costs to gather and pre-process the data, the costs to analyse the data, and the costs to put the resulting analytical models into production. In addition, the software costs and human and computing resources

should be taken into account here. It is important to do a thorough cost–benefit analysis at the start of the project.

- Finally, analytical models should also comply with both ***local and international regulation and legislation***. For example, in a credit risk setting, the Basel II and Basel III Capital Accords have been introduced to appropriately identify the types of data that can or cannot be used to build credit risk models. In an insurance setting, the Solvency II Accord plays a similar role. Given the importance of analytics now-a-days, more and more regulation is being introduced relating to the development and use of the analytical models. In addition, in the context of privacy, many new regulatory developments are taking place at various levels. A popular example here concerns the use of cookies in a web analytics context.

Types of Data



Big data is classified in three ways:

- Structured Data - RDBMS
- Unstructured Data – Images, Videos
- Semi-Structured Data – XML, JSON

Types of Big Data

There are primarily three types of data in big data:

1. Structured

Structured data refers to the data that you can process, store, and retrieve in a fixed format. It is highly organized information that you can readily and seamlessly store and access from a database by using simple algorithms. This is the easiest type of data to manage as you know what data format you are working with in advance. For example, the data that a company stores in its databases in the form of tables and spreadsheets is structured data. Ex: RDBMS

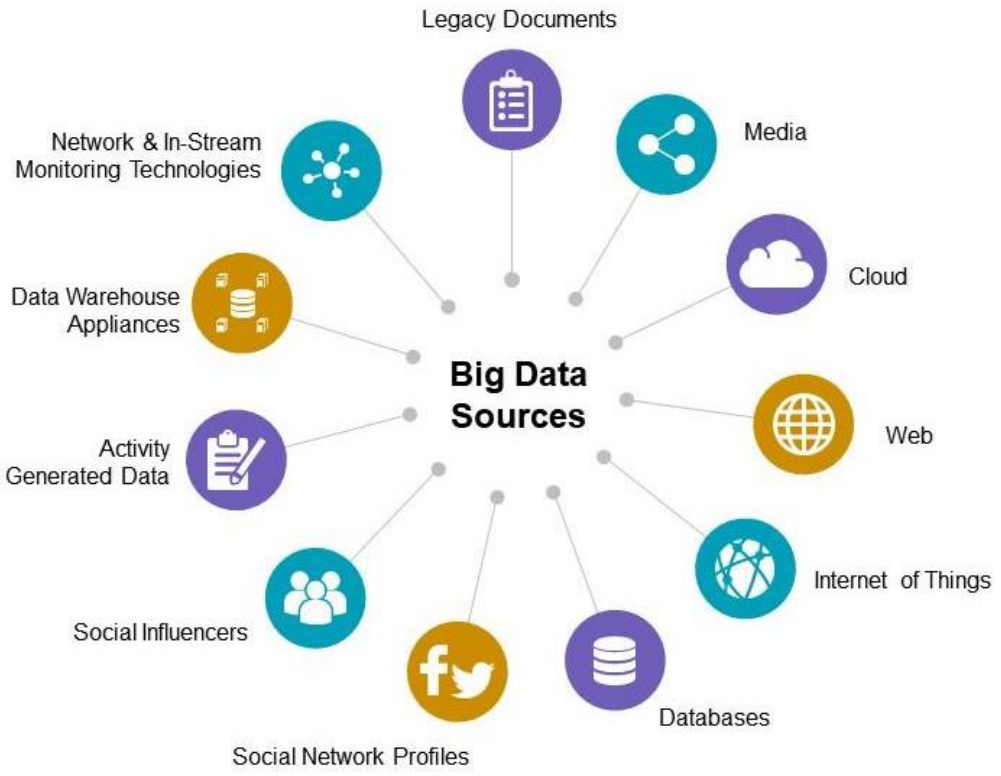
2. Unstructured

Data with an unknown structure is termed unstructured data. Its size is substantially bigger than structured data and is heterogeneous in nature. A great example of unstructured data includes the results you get when you perform a Google search. You get webpages, videos, images, text, and other data formats of varying sizes. Ex: Images, Audio Files, Video Files

3. Semi-structured

As the name suggests, semi-structured data contains a combination of structured and unstructured data. It is data that hasn't been classified into a specific database but contains vital tags that separate individual elements within the same. For example, a table definition in relational DBMS has semi-structured data. Ex: XML, JSON

Analytics – Types of Data Sources



- Data can originate from a variety of different sources. The bulk of big data generated comes from three primary sources: **social data, machine data and transactional data.**

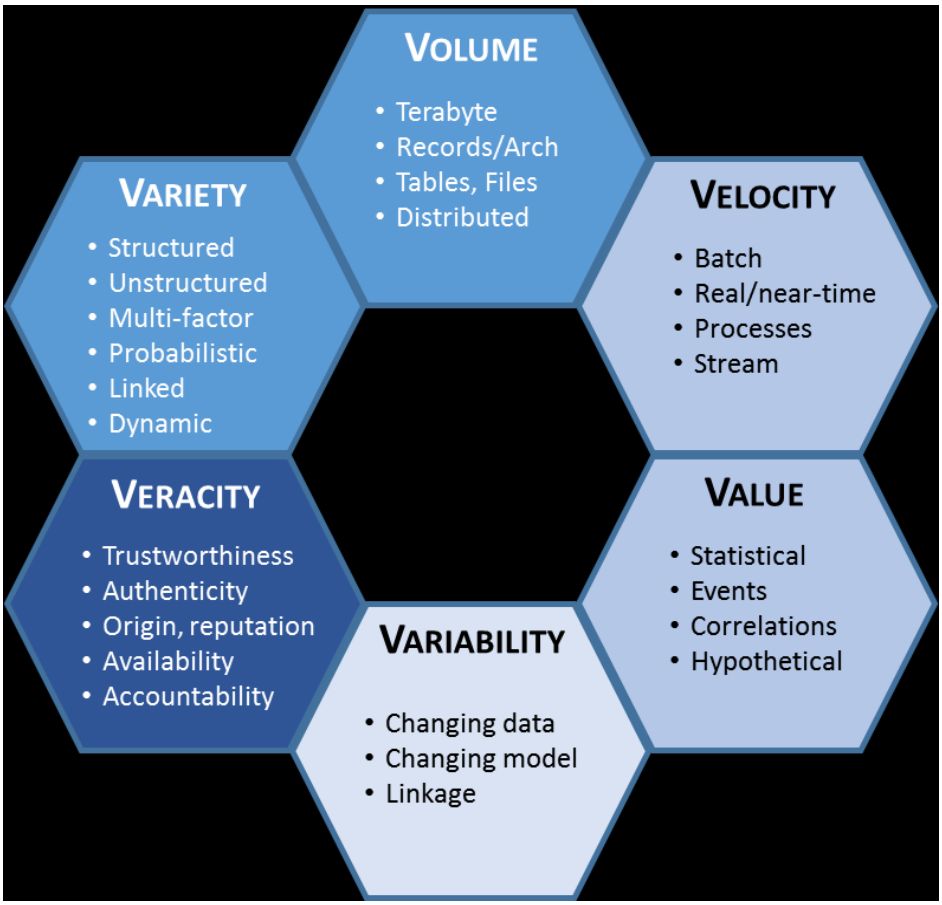
- ***Social data*** comes from the Likes, Tweets & Retweets, Comments, Video Uploads, and general media that are uploaded and shared via the world's favourite social media platforms. This kind of data provides invaluable insights into consumer behaviour and sentiment and can be enormously influential in marketing analytics. The public web is another good source of social data, and tools like Google Trends can be used to good effect to increase the volume of big data.
- ***Machine data*** is defined as information which is generated by industrial equipment, sensors that are installed in machinery, and even web logs which track user behaviour. This type of data is expected to grow exponentially as the internet of things grows ever more pervasive and expands around the world. Sensors such as medical devices, smart meters, road cameras, satellites, games and the rapidly growing Internet Of Things will deliver high velocity, value, volume and variety of data in the very near future.
- ***Transactional data*** is generated from all the daily transactions that take place both online and offline. Invoices, payment orders, storage records, delivery receipts – all are characterized as transactional data yet data alone is almost meaningless, and most organizations struggle to make sense of the data that they are generating and how it can be put to good use.

- Data are key ingredients for any analytical exercise.
- Hence, it is important to thoroughly consider and list all data sources that are of potential interest before starting the analysis.
- The rule here is the more data, the better. However, real life data can be dirty because of inconsistencies, incompleteness, duplication, and merging problems.
- Throughout the analytical modelling steps, various data filtering mechanisms will be applied to clean up and reduce the data to a manageable and relevant size.
- Worth mentioning here is the garbage in, garbage out (GIGO) principle, which essentially states that messy data will yield messy analytical models.
- It is of the utmost importance that every data pre-processing step is carefully justified, carried out, validated, and documented before proceeding with further analysis.
- Even the slightest mistake can make the data totally unusable for further analysis.
- In what follows, we will elaborate on the most important data pre-processing steps that should be considered during an analytical modelling exercise.

- ***Transactions*** are the first important source of data. Transactional data consist of structured, low level, detailed information capturing the key characteristics of a customer transaction (e.g., purchase, claim, cash transfer, credit card payment).
- This type of data is usually stored in massive online transaction processing (OLTP) relational databases. It can also be summarized over longer time horizons by aggregating it into averages, absolute/relative trends, maximum/minimum values, and so on.
- ***Unstructured data*** embedded in text documents (e.g., emails, web pages, claim forms) or multimedia content can also be interesting to analyse. However, these sources typically require extensive pre-processing before they can be successfully included in an analytical exercise.
- Another important source of data is ***qualitative, expert based data***. An expert is a person with a substantial amount of subject matter expertise within a particular setting (e.g., credit portfolio manager, brand manager). The expertise stems from both common sense and business experience, and it is important to elicit expertise as much as possible before the analytics is run. This will steer the modelling in the right direction and allow you to interpret the analytical results from the right perspective. A popular example of applying expert based validation is checking the univariate signs of a regression model. For example, one would expect *a priori* that higher debt has an adverse impact on credit risk, such that it should have a negative sign in the final scorecard. If this turns out not to be the case (e.g., due to bad data quality, multi collinearity), the expert/business user will not be tempted to use the analytical model at all, since it contradicts prior expectations.
- Now-a-days, ***data poolers*** are becoming more and more important in the industry. Popular examples are Dun & Bradstreet, Thomson Reuters. The core business of these companies is to gather data in a particular setting (e.g., credit risk, marketing), build models with it, and sell the output of these models (e.g., scores), possibly together with the underlying raw data, to interested customers. A popular example of this in the United States is the FICO score, which is a credit score ranging between 300 and 850 that is provided by the three most important credit bureaus: ***Experian, Equifax, and Transunion***. Many financial institutions use these FICO scores either as their final internal model, or as a benchmark against an internally developed credit scorecard to better understand the weaknesses of the latter.

- Finally, plenty of ***publicly available data*** can be included in the analytical exercise. A first important example is macroeconomic data about gross domestic product (GDP), inflation, unemployment, and so on.
- By including this type of data in an analytical model, it will become possible to see how the model varies with the state of the economy. This is especially relevant in a credit risk setting, where typically all models need to be thoroughly stress tested.
- In addition, ***social media data*** from Facebook, Twitter, and others can be an important source of information. However, one needs to be careful here and make sure that all data gathering respects both local and international privacy regulations.

Characteristics of Big Data



1. Volume

Volume refers to the amount of data that you have. We measure the volume of our data in Gigabytes, Zettabytes (ZB), and Yottabytes (YB). According to the industry trends, the volume of data will rise substantially in the coming years.

2. Velocity

Velocity refers to the speed of data processing. High velocity is crucial for the performance of any big data process. It consists of the rate of change, activity bursts, and the linking of incoming data sets.

3. Value

Value refers to the benefits that your organization derives from the data. Does it match your organization's goals? Does it help your organization enhance itself? It's among the most important big data core characteristics.

4. Variability

Variability refers to what extent, and how fast, is the structure of data changing? And how often does the meaning or shape of data change? In purely technical terms this means: if you change variables, your model will also change.

5. Veracity

Veracity refers to the accuracy/authenticity of your data. It is among the most important Big Data characteristics as low veracity can greatly damage the accuracy of your results.

6. Variety

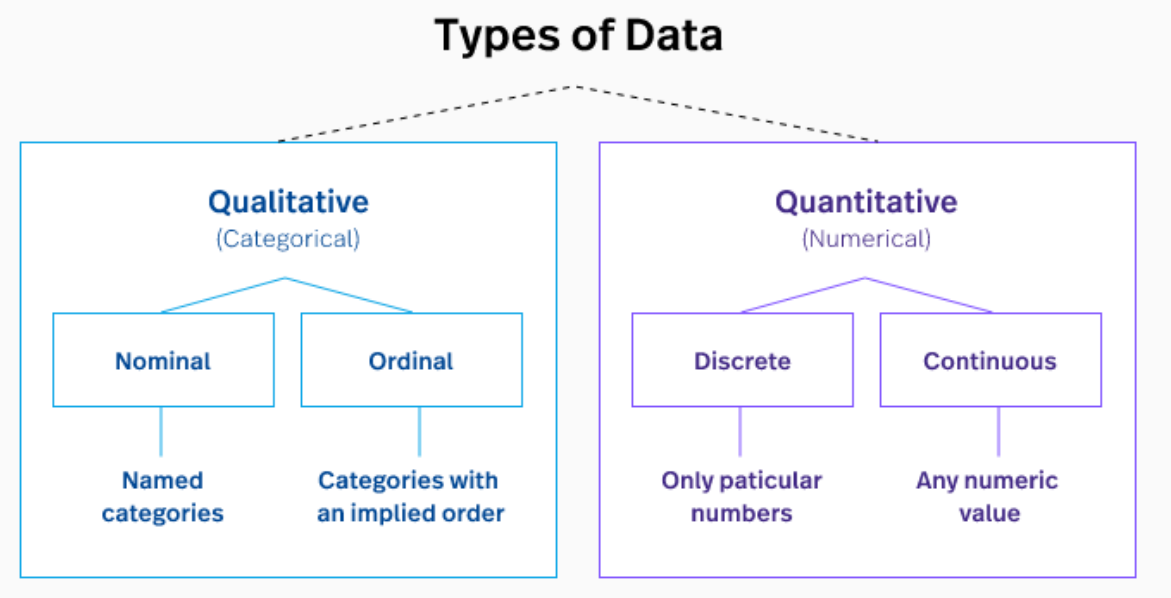
Variety refers to the different types of big data. It is among the biggest issues faced by the big data industry as it affects performance. It's vital to manage the variety of your data properly by organizing it. Variety is the various types of data that you gather from different kinds of sources.

Analytics – Sampling

- The aim of sampling is to take a subset of past customer data and use that to build an analytical model.
- A first obvious question concerns the ***need for sampling***.
- With the availability of high performance computing facilities (e.g., grid/cloud computing), one could also directly analyze the full data set. However, a key requirement for a good sample is that it should be ***representative of the future customers*** on which the analytical model will be run. Hence, the ***timing*** aspect becomes important because customers of today are more similar to customers of tomorrow than customers of yesterday.
- Choosing the ***optimal time window*** for the sample involves a trade-off between lots of data (and hence a more robust analytical model) and recent data (which may be more representative).
- The sample should also be taken from an average business period to get a picture of the target population that is as accurate as possible.
- It speaks for itself that ***sampling bias should be avoided*** as much as possible. However, this is not always straightforward.

Analytics – Types of Data Elements or Classification of Data

- When dealing with datasets, the category of data plays an important role to determine which pre-processing strategy would work for a particular set to get the right results or which type of statistical analysis should be applied for the best results.
- It is important to appropriately consider the different types of data elements at the start of the analysis.
- The following types of data elements can be considered:
  a. **Qualitative or Categorical**
  b. **Quantitative or Numerical**



- **Qualitative or Categorical Data**: This describes the object under consideration using a finite set of classes. It means that this type of data can't be counted or measured easily using numbers and therefore divided into categories. The gender of a person (male, female, or others) is a good example of this data type.
- These are usually extracted from audio, images, or text medium. Another example can be of a smartphone brand that provides information about the current rating, the colour of the phone, category of the phone, and so on. All this information can be categorized as Qualitative data. There are 3 subcategories under this:
  - **Nominal**: These are the set of values that don't possess a natural ordering. Let's understand this with some examples. The colour of a smartphone can be considered as a nominal data type as we can't compare one colour with others.
    It is not possible to state that 'Red' is greater than 'Blue'. The gender of a person is another one where we can't differentiate between male, female, or others. Mobile phone categories whether it is midrange, budget segment, or premium smartphone is also nominal data type.

    Ordinal: These types of values have a natural ordering while maintaining their class of values. If we consider the size of a clothing brand then we can easily sort them according to their name tag in the order of small < medium < large. The grading system while marking candidates in a test can also be considered as an ordinal data type where A+ is definitely better than B grade.

These categories help us deciding which encoding strategy can be applied to which type of data. Data encoding for Qualitative data is important because machine learning models can't handle these values directly and needed to be converted to numerical types as the models are mathematical in nature.
For nominal data type where there is no comparison among the categories, one-hot encoding can be applied which is similar to binary coding considering there are in less number and for the ordinal data type, label encoding can be applied which is a form of integer encoding.

- **Binary:** These are data elements that can only take on two values. Examples include gender, employment status.

- **Quantitative Data:** This data type tries to quantify things and it does by considering numerical values that make it countable in nature. The price of a smartphone, discount offered, number of ratings on a product, the frequency of processor of a smartphone, or ram of that particular phone, all these things fall under the category of Quantitative data types.

  The key thing is that there can be an infinite number of values a feature can take. For instance, the price of a smartphone can vary from x amount to any value and it can be further broken down based on fractional values. The two subcategories which describe them clearly are:
  - **Discrete**: The numerical values which fall under are integers or whole numbers are placed under this category. The number of speakers in the phone, cameras, cores in the processor, the number of sims supported all these are some of the examples of the discrete data type.
  - **Continuous:** The fractional numbers are considered as continuous values. These can take the form of the operating frequency of the processors, the android version of the phone, wifi frequency, temperature of the cores, and so on.

- Appropriately distinguishing between these different data elements is of key importance to start the analysis when importing the data into an analytics tool. For example, if marital status were to be incorrectly specified as a continuous data element, then the software would calculate its mean, standard deviation, and so on, which is obviously meaningless.

Analytics –Data Exploration and Exploratory Statistical Analysis
http://uc-r.github.io/gda

- Data visualization is a critical tool in the data analysis process.
- Visual data exploration is a very important part of getting to know your data in an "informal" way.
- It allows us to get some initial insights into the data, which can then be usefully adopted throughout the modelling.
- visual data exploration is about investigating the characteristics of your data set.
- Different plots/graphs can be useful here. A first popular example is **pie charts**. A pie chart represents a variable's distribution as a pie, whereby each section represents the portion of the total percent taken by each value of the variable.
- The following Figure represents a pie chart for a housing variable for which one's status can be own, rent, or for free (e.g., live with parents). By doing a separate pie chart analysis for the goods and bads, respectively, one can see that more goods own their residential property than bads, which can be a very useful starting insight.

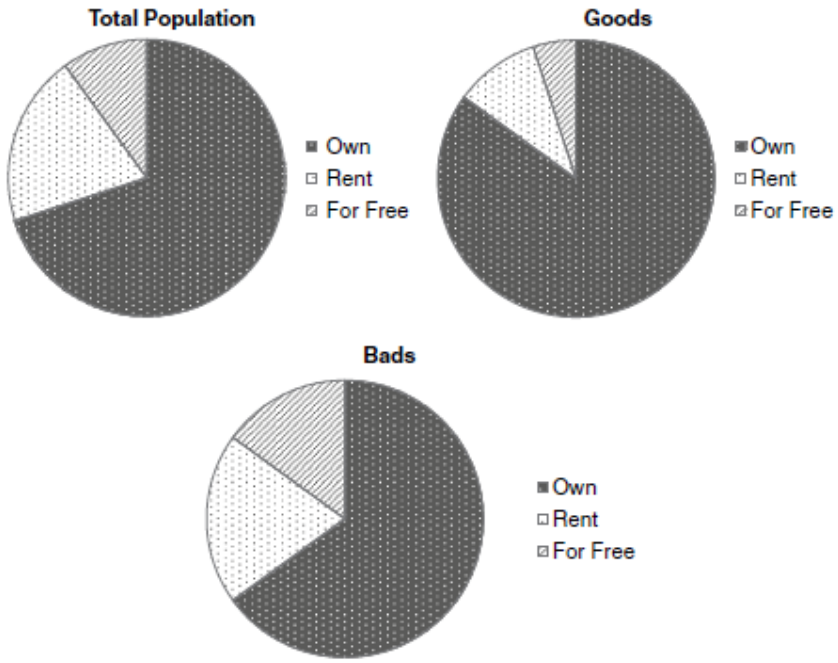
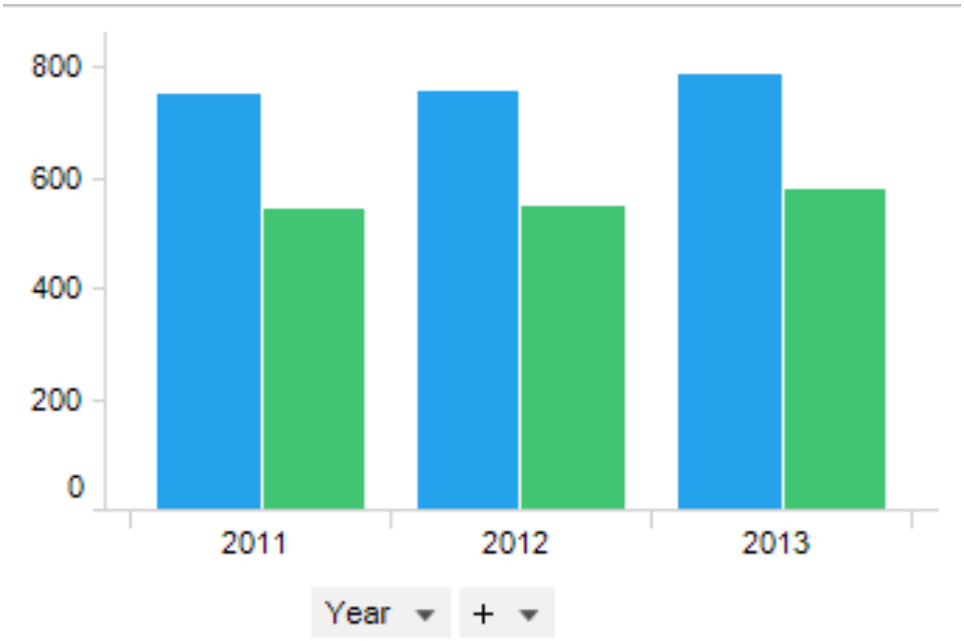
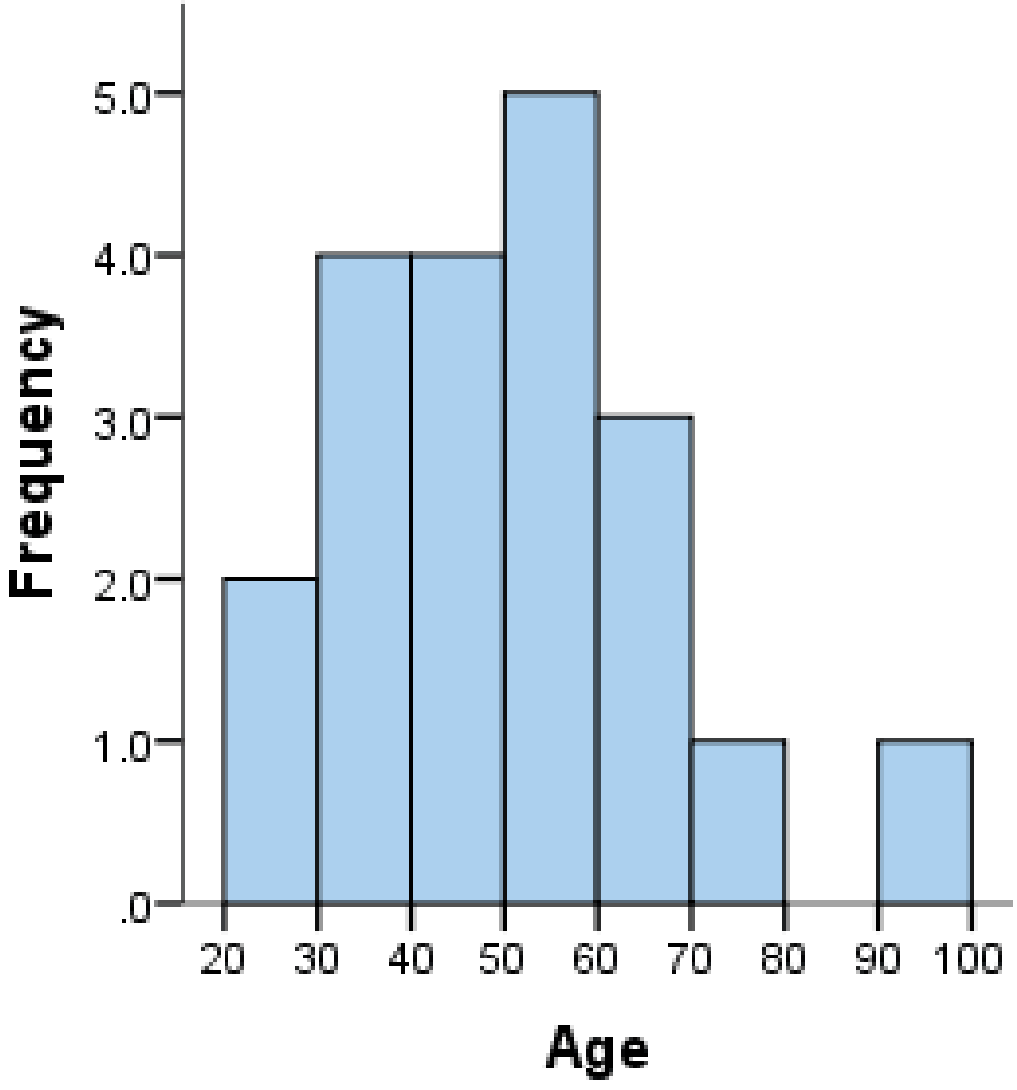


Figure: Pie Charts for Exploratory Data Analysis

- **Bar charts** represent the frequency of each of the values (either absolute or relative) as bars.



- Other handy visual tools are **histograms and scatter plots**. A histogram provides an easy way to visualize the central tendency and to determine the variability or spread of the data. It also allows you to contrast the observed data with standard known distributions (e.g., normal distribution). Scatter plots allow you to visualize one variable against another to see whether there are any correlation patterns in the data.



- Also, Online analytical processing (OLAP) based multidimensional data analysis can be usefully adopted to explore patterns in the data.
- A next step after visual analysis could be inspecting some basic statistical measurements, such as averages, standard deviations, minimum, maximum, percentiles, and confidence intervals. One could calculate these measures separately for each of the target classes to see whether there are any interesting patterns present.

Analytics – Missing Values

- Missing values can occur because of various reasons.
- The information can be non-applicable. For example, when modelling time of churn, this information is only available for the churners and not for the non-churners because it is not applicable there.
- The information can also be undisclosed. For example, a customer decided not to disclose his or her income because of privacy. Missing data can also originate because of an error during merging (e.g., typos in name or ID).
- Some analytical techniques (e.g., decision trees) can directly deal with missing values.
- Other techniques need some additional pre-processing.
- The following are the most popular schemes to deal with missing values:
  - **Replace (impute).** This implies replacing the missing value with a known value (e.g., consider the example in Table below). One could impute the missing credit bureau scores with the average or median of the known values. For marital status, the mode can then be used. One could also apply regression based imputation whereby a regression model is estimated to model a target variable (e.g., credit bureau score) based on the other information available (e.g., age, income). The latter is more sophisticated, although the added value from an empirical viewpoint (e.g., in terms of model performance) is questionable.
  - **Delete.** This is the most straightforward option and consists of deleting observations or variables with lots of missing values. This, of course, assumes that information is missing at random and has no meaningful interpretation and/or relationship to the target.
  - **Keep.** Missing values can be meaningful (e.g., a customer did not disclose his or her income because he or she is currently unemployed). Obviously, this is clearly related to the target (e.g., good/bad risk or churn) and needs to be considered as a separate category.

Table 2.1 Dealing with Missing Values

ID	Age	Income	Marital Status	Credit Bureau Score	Class
1	34	1,800	?	620	Churner
2	28	1,200	Single	?	Nonchurner
3	22	1,000	Single	?	Nonchurner
4	60	2,200	Widowed	700	Churner
5	58	2,000	Married	?	Nonchurner
6	44	?	?	?	Nonchurner
7	22	1,200	Single	?	Nonchurner
8	26	1,500	Married	350	Nonchurner
9	34	?	Single	?	Churner
10	50	2,100	Divorced	?	Nonchurner

- As a practical way of working, one can first start with statistical testing whether missing information is related to the target variable (using, for example, a chi-squared test, discussed later). If yes, then we can adopt the keep strategy and make a special category for it. If not, one can, depending on the number of observations available, decide to either delete or impute.

Analytics – Outlier Detection and Treatment

- One of the most important steps as part of data pre-processing is detecting and treating the outliers as they can negatively affect the statistical analysis.
- An Outlier is an observation in a given dataset that lies far from the rest of the observations.
- In other words, outliers are extreme observations that are very dissimilar to the rest of the population.
- An outlier may occur due to the variability in the data, or due to experimental error/human error. They may indicate an experimental error or heavy skewness in the data.
- In general, two types of outliers can be considered:
  Outliers are generally classified into two types: **Univariate** and **Multivariate**.

  Univariate Outliers – A univariate outlier is a case with an extreme value that falls outside the expected population values for a **single variable. Example Marks scored by students.**

  Multivariate Outliers – A multivariate outlier is a combination of unusual scores on at least two variables. **Example income and age**

  Both types of outliers can influence the outcome of statistical analyses.

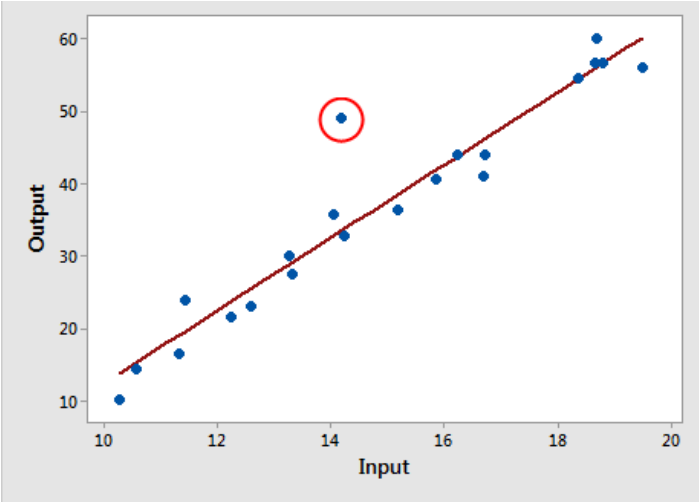


Figure: Univariate Outliers

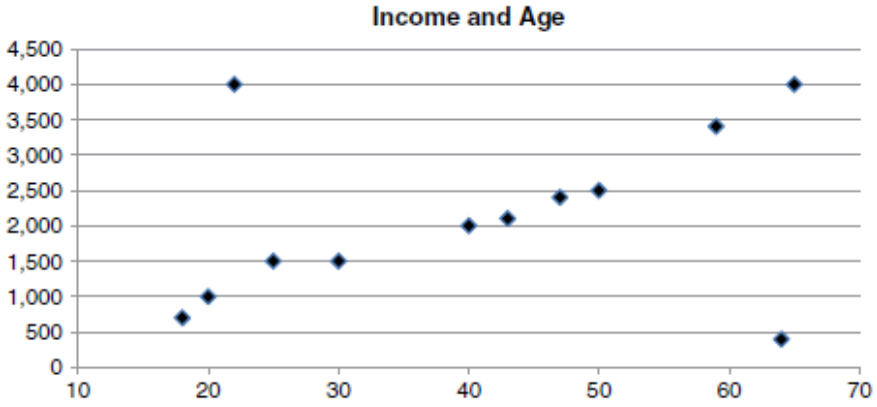


Figure: Multivariate Outliers

- Two important steps in dealing with outliers are detection and treatment. A first obvious check for outliers is to calculate the minimum and maximum values for each of the data elements.

Detecting Outliers:

- If our dataset is small, we can detect the outlier by just looking at the dataset. But what if we have a huge dataset, how do we identify the outliers then? We need to use visualization and mathematical techniques.
- Univariate outliers can be detected using Histograms, Boxplots, Z-Scores while multivariate outliers can be detected by fitting regression lines and inspecting the observations with large errors (using, for example, a residual plot).
- Below are the techniques of detecting univariate outliers
    1. **Histograms**
    2. **Boxplots**
    3. **Z-score**

- A histogram is a graphical representation of data points organized into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted visual by taking many data points and grouping them into logical ranges or bins. The figure shown below presents an example of a distribution for age whereby the circled areas clearly represent outliers.

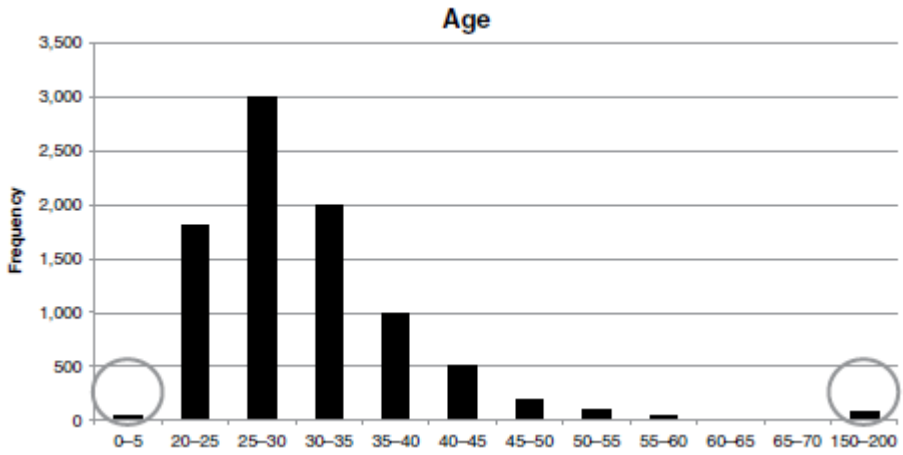
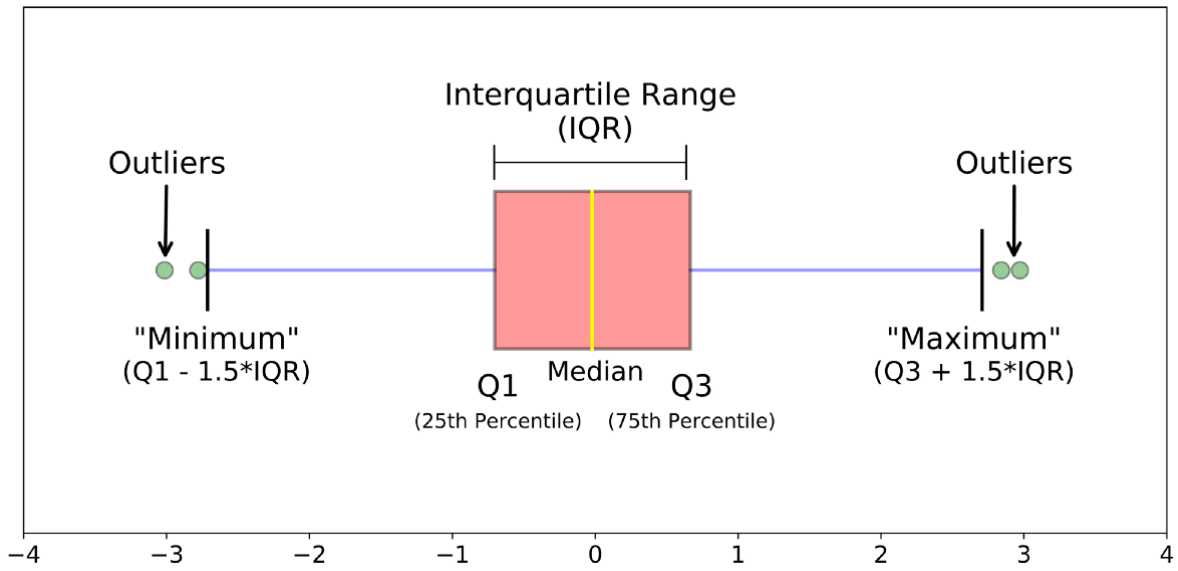


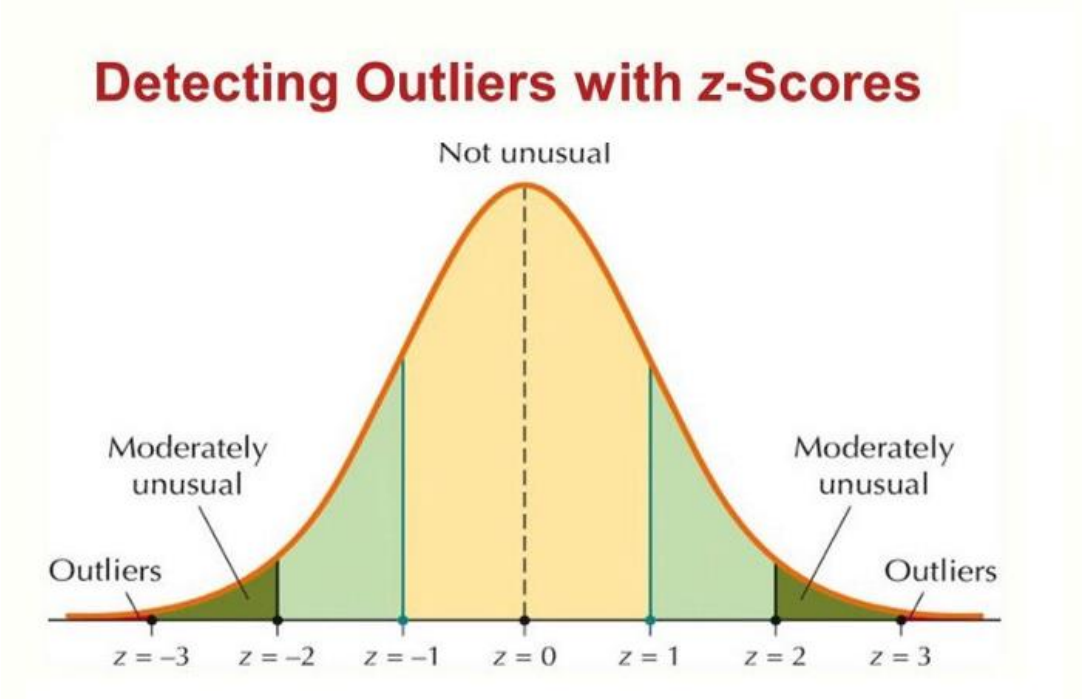
Figure: Histograms for Outlier Detection

- Another useful visual mechanism are ***Box plots***. Box plot is a data visualization plotting function. It shows the min, max, median, first quartile and third quartile. First quartile (25 percent of the observations have a lower value), the median (50 percent of the observations have a lower value), and the third quartile (75 percent of the observations have a lower value). All three quartiles are represented as a box. The minimum and maximum values are then also added unless they are too far away from the edges of the box.
- Outliers in Box Plots are then quantified as more than (1.5 * IQR) where Inter quartile Range IQR = Q3 – Q1.



- Another way is to calculate ***Z-scores***, measuring how many standard deviations an observation lies away from the mean, as follows:

$$z_i=(x_i-\mu)/\sigma$$



where **μ** represents the Mean of the variable and σ its standard deviation.

- An example is given in the below Table. Note that by definition, the *z* scores will have 0 mean and unit standard deviation. A practical rule of thumb then defines outliers when the absolute value of the *z*-score |z| is bigger than 3. Note that the *z* score relies on the normal distribution.

Table 2.2 Z-Scores for Outlier Detection

ID	Age	Z-Score
1	30	(30 – 40)/10 = –1
2	50	(50 – 40)/10 = +1
3	10	(10 – 40)/10 = –3
4	40	(40 – 40)/10 = 0
5	60	(60 – 40)/10 = +2
6	80	(80 – 40)/10 = +4
…	…	…
	$\mu = 40$ $\sigma = 10$	$\mu = 0$ $\sigma = 1$

- Some analytical techniques (e.g., decision trees, neural networks, Support Vector Machines (SVMs)) are fairly robust with respect to outliers.
  Others (e.g., linear/logistic regression) are more sensitive to outliers.

Treating Outliers:

- Various schemes exist to deal with outliers.
- It highly depends on whether the outlier represents a valid or invalid observation.
- For invalid observations (e.g., age is 300 years), one could treat the outlier as a missing value using any of the schemes discussed.
- For valid observations (e.g., income is $1 million), popular schemes like trimming the outlier or truncation/capping are used.

Analytics – Standardizing Data

- Data standardization is the process of making sure that our data set can be compared with other data sets.
- Standardizing data is a data pre-processing activity targeted at scaling variables to a similar range.
- Data standardization is about making sure that data is internally consistent; that is, each data type has the same content and format.
- **For example, suppose X_1 and X_2 went to different universities. One day, X_1 and X_2 got their final grades in Big Data Course. Assume, Professor of X_1 follows normal grading scale while Professor X_2 uses his own grading scale.**
  X_1 got a grade of 84 out of 100; where the test has a mean of 77 and a standard deviation of 6.
  X_2 got a grade of 452; where the test has a scale of 750, mean of 400, and standard deviation of 100.
  Both X_1 and X_2 scored above average, but who did better? While the main data points might not be immediately comparable, there is a way to standardize and compare the data points. Converting them to percentages shows that you came out ahead, with an 84% compared to your friend's 60%.
- Hence, it could make sense to bring them back to a similar scale.
- No matter how you standardize your data, standardization gives both a data point and full data set greater meaning.
- The following standardization procedures could be adopted:
  - Min/Max standardization

Min/max standardization

$$X_{new} = \frac{X_{old} - \min(X_{old})}{\max(X_{old}) - \min(X_{old})}(newmax - newmin) + newmin,$$

whereby newmax and newmin are the newly imposed maximum and minimum (e.g., 1 and 0).

  - *Z*-score standardization - Calculate the *z*-scores

$$z_i = \frac{x_i - \mu}{\sigma}$$

Decimal scaling

Decimal scaling

Dividing by a power of 10 as follows: $X_{new} = \frac{X_{old}}{10^n}$, with n the number of digits of the maximum absolute value.

Again note that standardization is especially useful for regression based approaches, but is not needed for decision trees, for example.

The result of standardization is that the features will be rescaled to ensure the mean and the standard deviation are 0 and 1, respectively.

Analytics – Categorization

- Categorization (also known as coarse classification, classing, grouping, binning, etc.) can be done for various reasons.
  - Used to reduce the number of categories and obtain a more robust model
  - Useful with both categorical and continuous variables
- For categorical variables, it is needed to reduce the number of categories. Consider, for example, the variable "purpose of loan" having 50 different values. When this variable would be put into a regression model, one would need 49 dummy variables (50 – 1 because of the collinearity), which would necessitate the estimation of 49 parameters for only one variable. With categorization, one would create categories of values such that fewer parameters will have to be estimated and a more robust model is obtained.

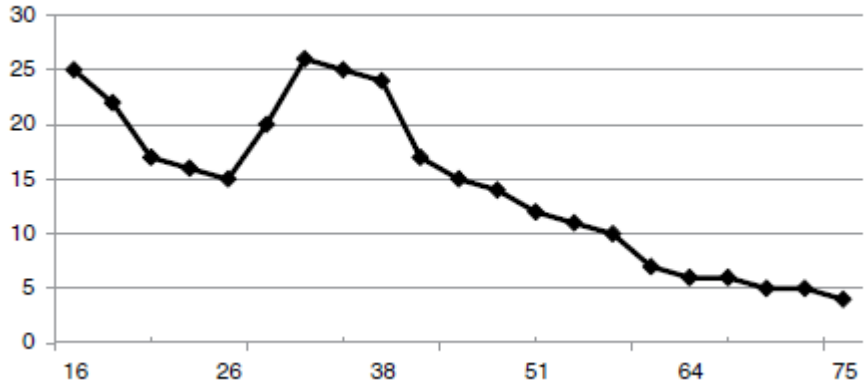


Figure: Default Risk versus Age

- For continuous variables, categorization may also be very beneficial. Consider, for example, the age variable and its risk as depicted in Figure above. Clearly, there is a non-monotonous relation between risk and age. If a nonlinear model (e.g., neural network, support vector machine) were to be used, then the nonlinearity can be perfectly modelled. However, if a regression model were to be used (which is typically more common because of its interpretability), then since it can only fit a line, it will miss out on the non-monotonicity. By categorizing the variable into ranges, part of the non-monotonicity can be taken into account in the regression. Hence, categorization of continuous variables can be useful to model nonlinear effects into linear models.

- Various methods can be used to do categorization. **Two very basic methods are equal interval binning and equal frequency binning.** Consider, for example, the income values 1,000, 1,200, 1,300, 2,000, 1,800, and 1,400. Equal interval binning would create two bins with the same range—Bin 1: 1,000, 1,500 and Bin 2: 1,500, 2,000—whereas equal frequency binning would create two bins with the same number of observations—Bin 1: 1,000, 1,200, 1,300; Bin 2: 1,400, 1,800, 2,000. However, both methods are quite basic and do not take into account a target variable (e.g., churn, fraud, credit risk).

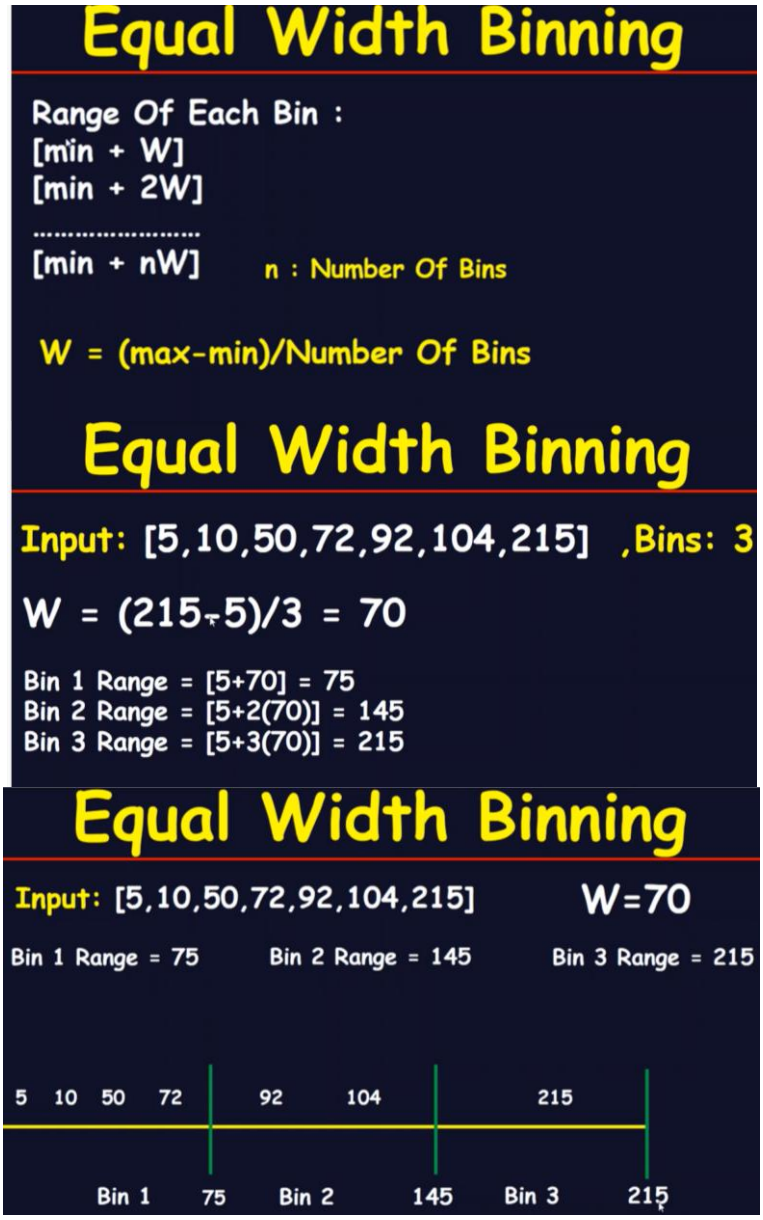
Equal frequency: Divides the data set into bins that all have the same number of samples.

- Input: [5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215]
- Output: [5, 10, 11, 13] [15, 35, 50, 55] [72, 92, 204, 215]

Equal Interval: In Equal width binning, all bins have equal width, or represent an equal range of the original variable values, no matter how many cases are in each bin.

- **Input:** `[5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215]`
- **Output:**

```
[5, 10, 11, 13, 15, 35, 50, 55, 72]
[92]
[204, 215]
```



- **Data :** 0, 4, 12, 16, 16, 18, 24, 26, 28
- **Equal width**
  - Bin 1: 0, 4 [-,10)
  - Bin 2: 12, 16, 16, 18 [10,20)
  - Bin 3: 24, 26, 28 [20,+)
- **Equal frequency**
  - Bin 1: 0, 4, 12 [-, 14)
  - Bin 2: 16, 16, 18 [14, 21)
  - Bin 3: 24, 26, 28 [21,+)

- Chi-squared analysis is another method to do categorisation.

A chi-square ($\chi 2$) statistic is a test that measures how a model compares to actual observed data.

A chi-square test is used to help determine if observed results are in line with expected results, and to rule out that observations are due to chance. A chi-square test is appropriate for this when the data being analysed is from a random sample, and when the variable in question is a categorical variable.

$\chi 2$ provides a way to test how well a sample of data matches the (known or assumed) characteristics of the larger population that the sample is intended to represent. This is known as goodness of fit.

The Formula for Chi-Square Is

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where:

$c =$ Degrees of freedom

$O =$ Observed value(s)

$E =$ Expected value(s)