

Data Extraction

Introduction

Data extraction is the process of retrieving data from data sources for further data processing or storage

Various techniques used for data extraction.

- Web scraping
- Using APIs
- Query languages
- Open source datasets.

DO's and DON'Ts of Data Extraction

- Never Extract Data From Web Without First Checking For The Legalities Related To The Type Of Data You Are Extracting.
- Extract Only The Data Needed (because the extraction process is labour intensive and costly)
- Extracted Data Should Accurately Reflect The Problem Statement And Provide Sufficient Information To Understand And Perform Analysis

Data Preparation

Descriptive Data Summarization

- Assessing the usefulness of each variable for modelling.
- Deciding the usage/ interpretation of the variables
 - Simple descriptive Statistics.
 - Distributions

Data Visualization

Visualization of data

Data visualization is the representation of data through use of common graphics, such as charts, plots, infographics, and even animations. These visual displays of information communicate complex data relationships and data-driven insights in a way that is easy to understand.

The advantages and benefits of good data visualization

- Fast decision-making.
- More people involved.
- Higher degree of involvement.
- Better understanding.
- Obtain critical data in real time.
- Quickly process large amounts of data.

Key Questions

The start of any data visualization depends on the data set and what we want to know. Broadly speaking, there are five main questions we can ask.

- How does my one thing correspond to another?
- How is this data linked to other data?
- How is that data scattered?
- What is this data composed of, and
- How does this data appear on a map?

Types of Data Visualization.

Common general types of data visualization:

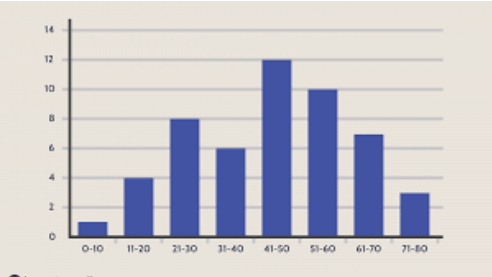
- Charts
- Tables
- Graphs
- Histograms
- Scatter plots
- Heat maps
- Pie Charts

Selecting the Right Chart

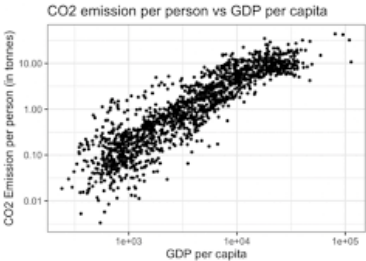
- How many variables do you want to show in a single chart? One, two, three, many?
- How many items (data points) will you display for each variable? Only a few or many?
- Will you display values over a period of time, or among items or groups?

Few Examples

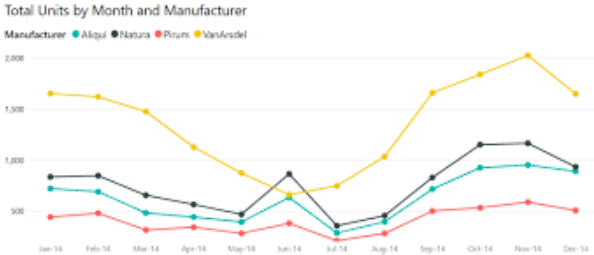
Histogram



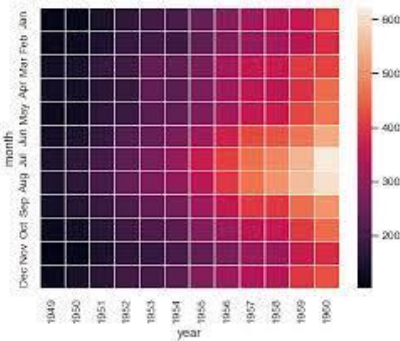
art



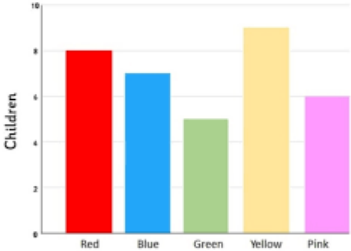
Scatter plot



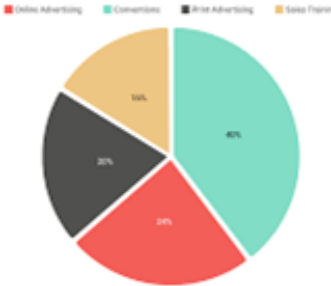
Heat map



Favourite Colour



Bar chart



Common Libraries and tools for Data Visualization

- Matplotlib - Python Library
- Seaborn - Python Library
- Dataiku - Interactive AI Tool
- Tableau - Visualization Tool
- Power BI - Visualization Tool

Descriptive Statistics






Categorical Data

- Modal Value (Mode)
 - The most frequent value
 - Multi-modal distributions have multiple values with high frequency
- Frequency Distribution of all values (Bar Chart)
 - As percentage or/and count
- Set of Unique Values

Value ▲	Proportion	%	Count
AFS Cross-Sell		8.01	6375
AFS Cross-Sell Partnership Test		7.15	5695
AFS Mortgage Customer		0.56	447
AFS Mortgage Prospecting		0.1	77
AFS Retirement		0.03	23
Auto Cross-Sell		14.22	113...
Auto Cross-Sell Prescreen		0.05	43
Auto Cross-Sell X-date		3.57	2840
Auto Cross-Sell X-date Prescreen		0.0	2
Property Cross-Sell		7.25	5772
Property Cross-Sell Prescreen		0.01	7
Prospecting		5.37	4278
Prospecting X-date		1.32	1052
Renters Cross-Sell Prescreen		0.0	3
Retention		46.25	368...
Retention Priority Customer		1.01	802
Single Service Household		0.24	189
UNDEFINED		0.01	9
Welcome		4.76	3787
Winback		0.07	59

Ordinal Data

- Minimum and Maximum Values
- Median Value
- Modal Value
- Frequency distribution
 - Bar chart
 - Histogram

Value ▲	Proportion	%	Count
1		19.17	384
2		7.64	153
3		8.89	178
4		39.44	790
5		24.86	498

- The example above shows the distribution of Age bracket
- Each value, while coded as a number is not necessarily equidistant
- Calculating a mean does not necessarily make sense

Measures of Central Tendency

- Arithmetic Mean or Average, μ
 - Sensitive to extreme values
- Weighted Average
 - When not all values are equal, each value may have an associated weight, w_i
- Trimmed Mean
 - Mean obtained by removing (giving a weight of 0) to values at the top and bottom $s\%$ (the value of s is normally small)
- Median
 - Less sensitive to skew (outliers)
 - If n is odd, the median is the middle value of the ordered set
 - If n is even then the median is the average of the two middle values
 - Approximate value may be calculated from grouped data (see next slide)
- Mode
 - The value that occurs the most frequently
 - Data may be multi-modal (have multiple values with maximum frequency)

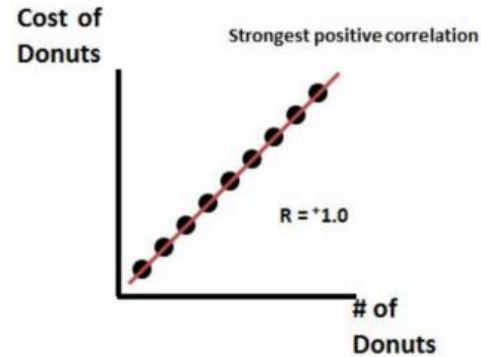
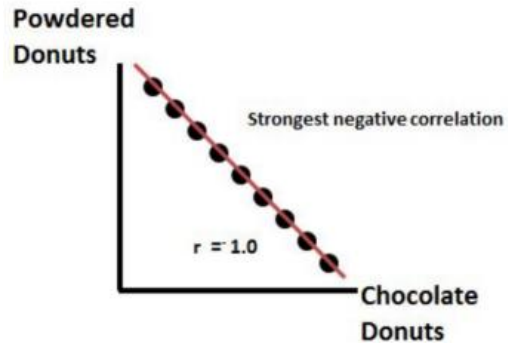
Data correlation

Is a way to understand the relationship between multiple variables and attributes in your dataset. Using Correlation, you can get some insights such as:

- One or multiple attributes depend on another attribute or a cause for another attribute.
- One or multiple attributes are associated with other attributes.
- Usually ranges from -1 to 1
 - 0 implies no relationship
 - -1 implies perfect negative linear relationship
 - 1 implies perfect linear relationship

Types of correlation

- Positive Correlation
- Negative Correlation
- No Correlation



Covariance

- Measures the tendency for two attributes to co-vary
 - Variance of an attribute is defined as the average squared deviation from the mean
 - Covariance of two attributes X and Y is defined as

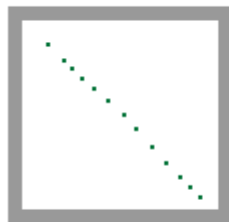
$$c(X, Y) = \sum_{i=1}^n \frac{(x_i - \bar{x}) \times (y_i - \bar{y})}{n - 1}$$

- Properties of $c(X, X)$
 - If attributes X and Y tend to increase together, then $c(X, Y) > 0$
 - If attributes X tends to decrease when attribute Y increases, then $c(X, Y) < 0$
 - If attributes X and Y are independent then $c(X, Y) = 0$
- Notice that
 - $c(X, X) = \text{Variance of X}$

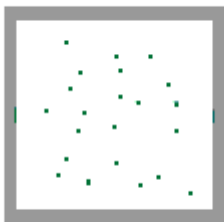
Covariance

- Positive Covariance
- Negative Covariance
- No Covariance

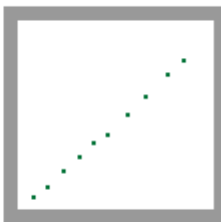
COVARIANCE



Large Negative
Covariance



Near Zero
Covariance



Large Positive
Covariance

Feature Selection

Feature selection is the process of isolating the most consistent, non-redundant, and relevant features to use in model construction. Methodically reducing the size of datasets is important as the size and variety of datasets continue to grow. The main goal of feature selection is to improve the performance of a predictive model and reduce the computational cost of modeling.

Standardisation

Standardization is scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

Here's the formula for standardization:

- Where μ is the mean of the feature values.
- And σ is the standard deviation of the feature values

$$z_x = \frac{x - \mu}{\sigma}$$

Data Transformation

TRANSFORMING DATA TYPES

- Some modelling techniques can only handle categorical attributes
 - Example, some decision tree implementations
- Others can only handle numeric attributes
 - Example, Neural Networks
- Yet other can handle only data types of one type at a time (without introducing a bias), all categorical or all numeric
 - Example, Distance based Techniques such as Clustering and Nearest Neighbour

MAPPING CATEGORICAL VALUES TO NUMERIC SCALE

- Some modelling techniques cannot handle categorical attributes e.g. neural networks
- Prior to discovery, categorical attributes must be mapped onto a numeric scale
 - Note that there must be a rationale for any mapping
- Domain based Mapping
 - Zip Code can be mapped to latitude and longitude
 - Colour can be mapped to RGB values

1-HOT ENCODING

- For each category value, create a new binary attribute that takes the value 1 if the categorical attribute has that value set, else it takes the value 0

Fruit	Categorical value of fruit	Price
apple	1	5
mango	2	10
apple	1	15
orange	3	20

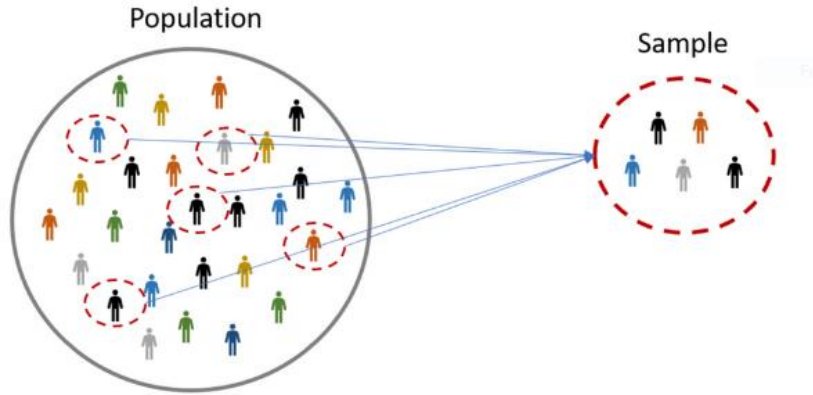


apple	mango	orange	price
1	0	0	5
0	1	0	10
1	0	0	15
0	0	1	20

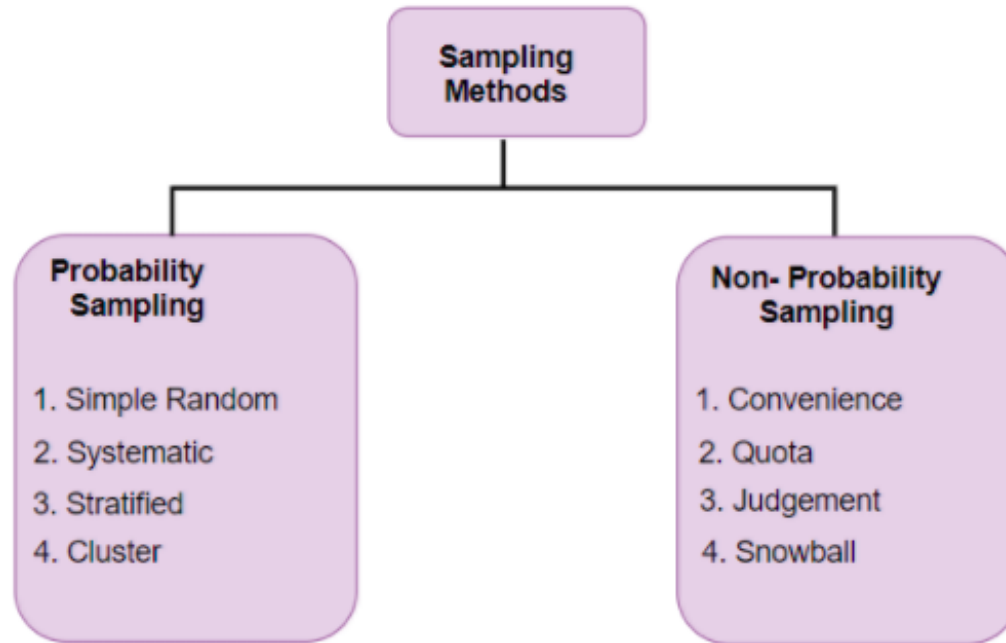
- Dimensionality can increase dramatically
- Density of some values can be very low
 - modelling technique may not be able to use an attribute effectively with such a skewed distribution

Sampling

“Sampling is a **method that allows us to get information about the population based on the statistics from a subset of the population** (sample), without having to investigate every individual”



Sampling methods



MIN-MAX NORMALIZATION

- Linear transformation

$$f(x) = \frac{x - \min}{\max - \min}$$

- $f(x)$ is the new normalised value, belonging to the range [0,1]
 - min and max are the minimum and maximum values for the attribute being normalised
- More generally if the new range for the attribute is [nmin, nmax]

$$f(x) = \frac{x - \min}{\max - \min} (n\max - n\min) + n\min$$

- Range Normalisation as it does not affect the distribution of the values within the range, i.e. the shape of the distribution remains unaltered

DEALING WITH OUT-OF-RANGE VALUES

- Leave as is
 - Implies the state space is not a unit state space
 - A number of mining techniques can handle non unit state spaces
- Ignore the whole record
- Clip the values
 - If the value is larger than the maximum, set it equal to the maximum
 - If the value is smaller than the minimum, set it equal to the minimum
- If the out-of-range values are indeed valid, these techniques will result in the loss of information and introduce a bias in the data
- What would be a more robust way of dealing with possible out-of-range values?
 - What additional information is required to do this?

MISSING VALUES AND NOISE

MISSING VALUES AND NOISE

- Reasons for Missing Values
 - Data not available
 - Data Survey may show some default values input during data collection
 - Data value not appropriate
- Why do these need to be handled explicitly?
 - Some Modelling techniques cannot handle missing values
 - If they do, they may use undesirable ways to deal with them
 - Modeller had more control on how they are handled
 - There may be a domain meaning to the missing value
- Understanding Missing Value patterns similar to any classification data mining task
 - Significant patterns that allow you to predict that the attribute value is missing or not

DEALING WITH MISSING VALUES

- Unbiased Estimators
 - These are estimates for the data that do not change important data characteristics
 - What are the important characteristics?
 - Mean
 - Standard Deviation
 - Correlation with other variables
- Mean commonly used to fill numeric attributes
- Mode used to fill categorical attributes
- These generate bias within the distribution and are not necessarily valid values

USING MODELS TO FILL MISSING VALUES

- Preserving Standard Deviation is more accurate as it takes the mean and the variation around it as the basis for the prediction
- Most data mining tools can be used to predict missing values by training on instances that have no missing values
 - Less bias introduced into the data as it takes correlation among variables into account when predicting missing values
 - Can introduce bias if the missing values show a strong pattern of their own as the training data used to build the model is biased
- Use Linear Regression (or any other supervised learning algorithm) to fill missing values
 - Takes interaction between attributes into account
 - Assumes a linear relationship
 - Does fairly well unless relationship is very non-linear

DIMENSIONALITY REDUCTION

Why Dimensionality Reduction?

- Large numbers of input features can cause poor performance for machine learning algorithms.
- High Dimensional data is difficult to study and visualise.
- Data Compression: Efficient storage and retrieval.

THE CURSE OF DIMENSIONALITY

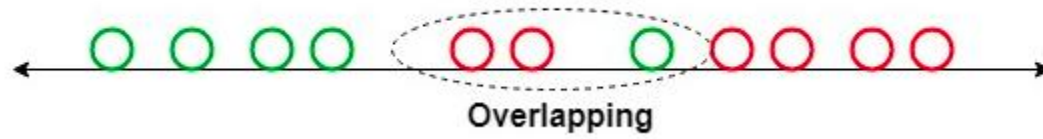
- In a two dimensional unit state space, the maximum distance between any two points, using Pythagoras theorem, is (the number of dimensions)
- As we increase the dimensionality of the state space, unless the points have the same value along the additional dimensions, the distance between them increases
- Hence adding dimensions increases the sparsity of the state space as the points in the space are pushed further apart
- Learning from sparse data is less effective, so adding dimensions makes the learning less robust – a fact referred to as the curse of dimensionality

Methods of Dimensionality Reduction

The various methods used for dimensionality reduction include:

- Principal Component Analysis (PCA)
- Linear Discriminant Analysis (LDA)
- Generalized Discriminant Analysis (GDA)

LDA



Advantages of Dimensionality Reduction

- It helps in data compression, and hence reduced storage space.
- It reduces computation time.
- It also helps remove redundant features, if any.

Disadvantages of Dimensionality Reduction

- It may lead to some amount of data loss.
- PCA tends to find linear correlations between variables, which is sometimes undesirable.
- PCA fails in cases where mean and covariance are not enough to define datasets.

Thank You