ASSIGNMENT-1

① Write a brief note on principal component Analysis.

Principle Component Analysis:

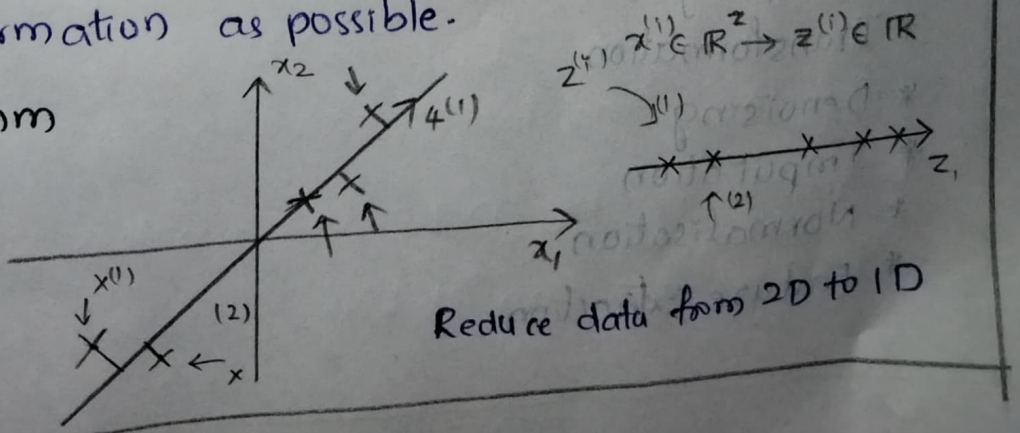
* Principle Component Analysis or PCA is a dimension-ality reduction method that is often used to reduce the dimensionality of large data sets by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

* Reducing the number of variables of a data set naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity.

* Because similar smaller data sets are easier to explore and visualise and make analysing data much easier and faster for machine learning algorithms without extraneous variables to process.

* So to sum up, the idea of PCA is simple- reduce the number of variables of a data set, while preserving as much information as possible.

PCA Algorithm



For 2D data draw a line along z axis and draw/project points onto it and Now plot them on a straight line using distance among each projected point on z axis which reduces 2D into 1D.

Similarly for 3D reducing into 2D we have to consider a plane and project points onto it and draw it separately which makes it 2D.

---

② Explain all the data pre-processing steps in detail.

Data Preprocessing:

* Data pre-processing a component of data preparation describes any type of processing performed on raw data to prepare it for another data preprocessing procedure
* It is important for the data mining process.
* Data preprocessing techniques have been adapted for training machine learning models and AI models and for running inferences against them.

Steps in Pre-processing:

* Sampling
* Transformation
* Denoising
* Imputation
* Normalisation
* Feature Extraction.

Sampling:

* A method which selects a representative subset from a large population of data.
* This sample of data should be representing whole population with small data.

Transformation:

* In this step the raw data is manipulated to get a single input.

Denoising: which removes noise from data.

Imputation: which synthesizes statistically relevant data for missing values.

Normalisation: which organises data for more effecient access and

Feature extraction: which pulls out a relevant feature subset that is significant in particular context.

* In few domains like medical imaging, data preprocessing is not recomended as it may lose some important information while pre-processing.
* If there exist any sort of missing data, or outliers which causes error in computation can be avoided by following necessary pre-processing steps depending on the visualisation techniques.

→ Missing values
→ Discard bad record
→ Assign a sentinel value

→ Assign the average value

→ Assign value based on Nearest Neighbour

→ Compute a substitute value.

3) Explain the importance of data visualisation in detail

→ Data Visualisation is the representation of data through the use of common graphics, such as charts, plots, infographics and even animations.

→ These visual display of information communicate complex data relationships and data-driven insights in a way that is easy to understand

→ Data Visualization is essential to assist business in quickly identifying data trends, which would otherwise be a hassle. The pictorial representation of data sets allows analysts to visualise concepts and new patterns.

→ A dashboard, graph, infographics, map, chart, video, slide etc all these mediums can be used for visualizing and understanding data.

→ Visualising the data enable decision-makers to interrelate the data to find better insights and reap the importance of data visualisation, which are,

* Analyzing the Data in a better way.
* Faster Decision Making.

* Making sense of complicated Data

Applications:

* Most common use today is as a business intelligence (BI) reporting tool.
* Users can setup visualisation tools to generate automatic dashboards that track company performance across key performance indicators (KPIs) and visually interpret the results.
* Many business departments implement data visualisation softwares to track their own initiatives
* Being used as front ends for more sophisticated big data environments.

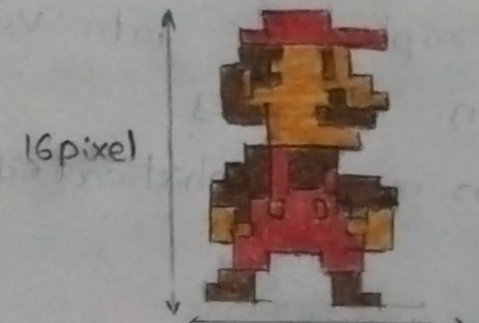
Most Widely used D.V. softwares

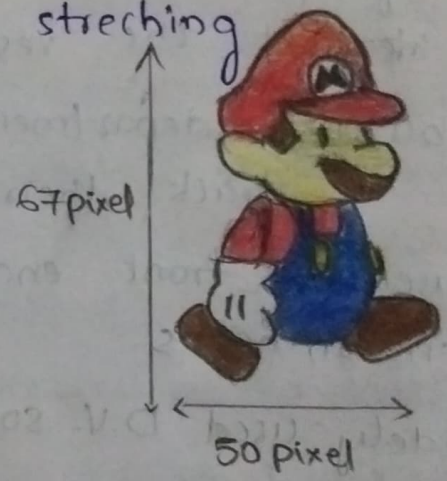
1. IBM Cognos Analytics
2. Qlik Sense and Qlik View
3. Microsoft Power B.I
4. Oracle VisualAnalyser
5. Google charts.

4) Discuss in detail about pixel normalisation, raster and vector images.

* Pixel is a computer term for 'picture element'. The ideal is that each pixel is only one color, and color is the detail in the image.

* In computer graphics a pixel, dots, or picture element is a physical pant in a picture.
* A pixel is simply the smallest addressable element of a picture represented on a screen.
* In image processing, normalisation is a process that changes the range of pixel intensity values. Normalisation is sometimes called contrast streching or histogram streching





Raster Graphics:

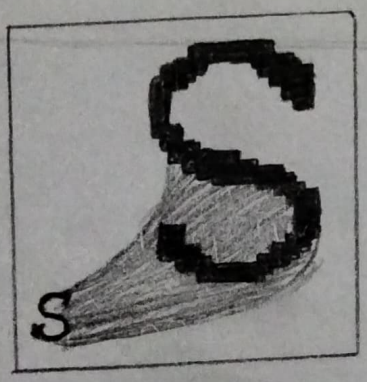
Raster images use bit maps to store information. This means a large file needs a large bitmap.

In computer graphics, objects are typically represented by sets of connected, planar polygons and the task is to create a raster (pixel level) image representing these objects, their surface properties, and their interactions with light source and other objects.

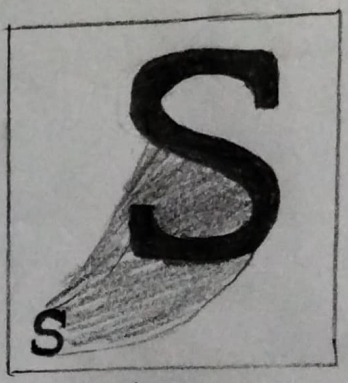
* In computer graphics and digital photography, a raster graphic represents a two-dimensional image as a rectangular matrix or grid of square pixels, viewable via a computer display, paper, or another display medium.

Vector Graphics:

Vector graphics, as a form of computer graphics, is the set of mechanisms for creating visual images directly from geometric shapes defined on a Cartesian plane, such as points, lines, curves and polygons.



Raster
.jpeg .gif .png



Vector
.svg

Reasons for Transformation

* Compressing the contents for transmission. A vertex and edge list is almost always more compact than a raster image
* Comparing the contents of two or more images. It is generally easier and more reliable to compare higher-level features of images, rather than their pixels.

* Transforming the data. Affine transformations such as rotation and scaling are easier to apply to vector representations than to raster.

* Segmenting the data. Isolating regions by drawing boundaries around them is an effective method for interactive exploration and model building.