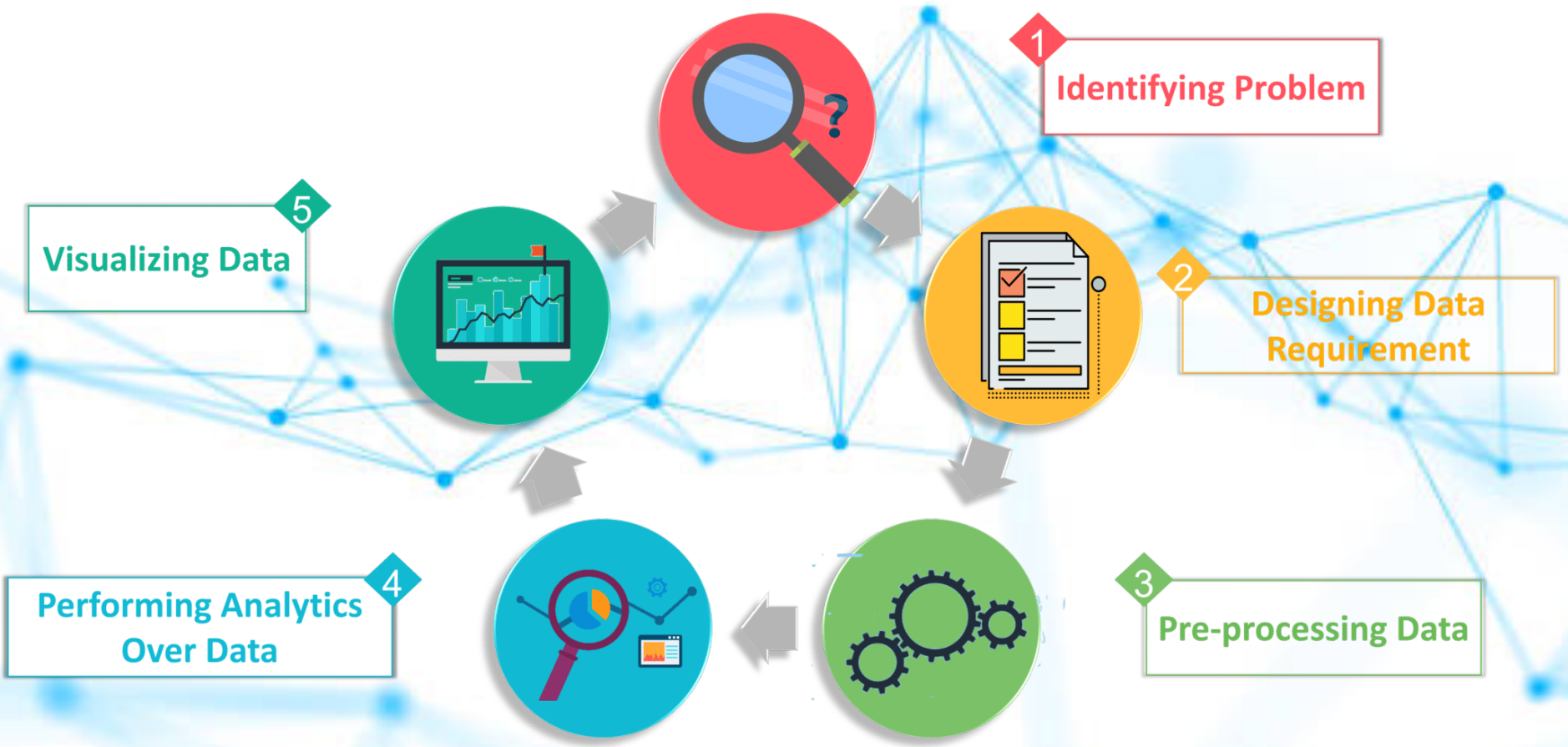
Fundamentals of Big Data Analytics

Lecture 4- Data & Exploratory Data Analysis

Dr. Iqra Safder
Assistant Professor
FAST NUCES, Lahore

Data Analytics: Process and Tasks

The Analytics Process



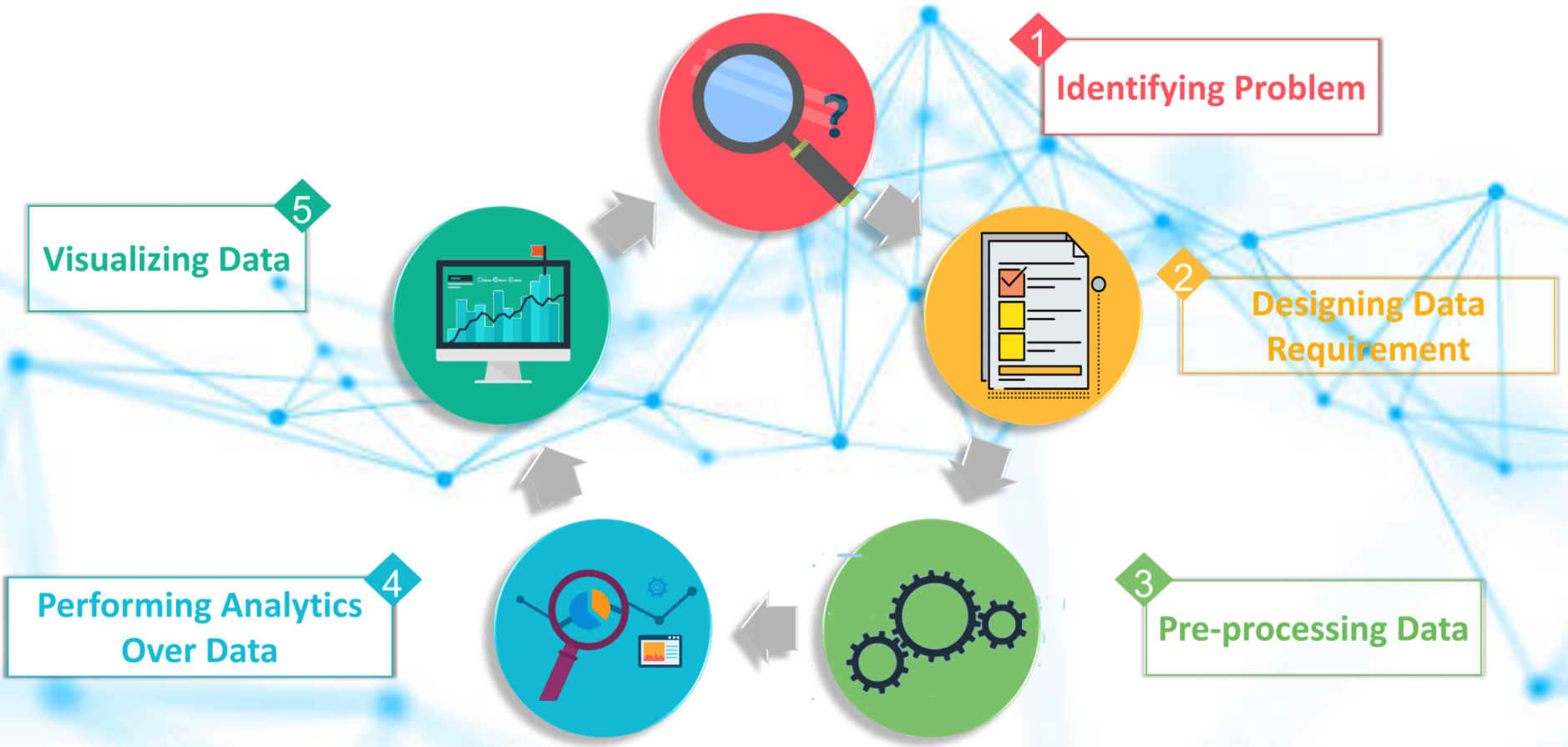
The Analytics Process

- **Business Objective**
  - Why we are seeking data analytics in the first place?
  - How can we reduce production costs without sacrificing quality?
  - What are some ways to increase sales with our current resources?
  - Do customers view our brand in a favorable way?
- **Data Collection**
  - What data is needed and available?
  - Identify sources of data and relevance of data
  - Are there enough instances, are all relevant features there?
  - Identify datasets, acquire and retrieve
  - Sources RDBMS, .txt, webservices (soup), RSS, tweets
  - Experiments, synthetic data generation, Survey

The Analytics Process

- Data Preparation
  - Make the data ready for analytics
  - Exploratory Data Analysis Describe, Summarize, Visualize
  - Pre-process: Improve data quality, clean data, transformation, standardization, normalization
- Data Analysis
  - Apply analytical techniques
  - Supervised and unsupervised learning, Graph analytics
- Report and Deployment
  - Communicate results and findings, and apply conclusions to gain benefit

The Analytics Process



Data Analytics Tasks and Methods

Data Analytics is the process

- to discover patterns in data to
- find relationships in data
- to (automatically) extract knowledge from data
- to summarize data in ways that are understandable and useful

Discovering knowledge form data often requires learning

Data Analytics Tasks and Methods

Descriptive Analytics

- Uncover patterns, correlations, trends & trajectories describing data, Explanatory in nature
- Require post-processing to validate and explain the results
- Clustering/grouping the data or Detecting outliers (anomalies) in data

Predictive Analytics

- Predict value of an attribute based on values of other attributes
- Predicted attribute: Target/dependent/response variable
- Attributes used to predict: Predictor/explanatory/independent variables
- Classification: nominal target attribute (class labels)
- Regression: numeric target attribute

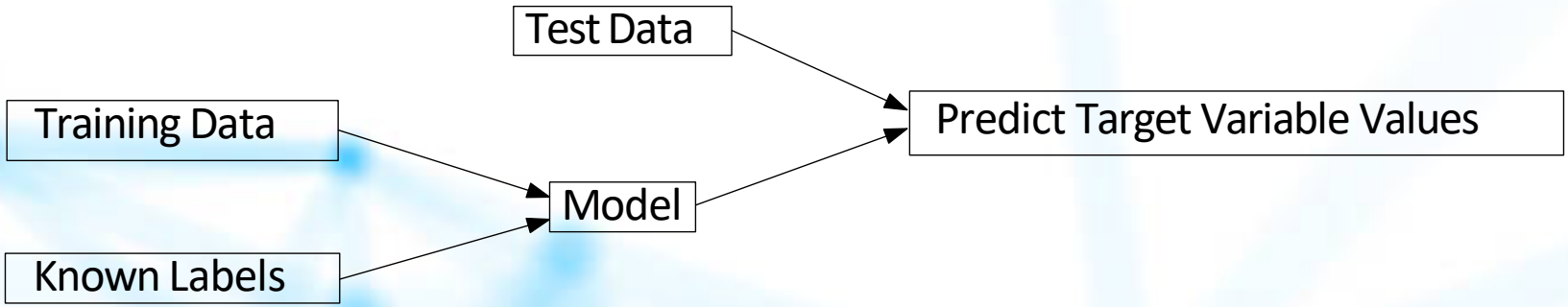
Data Analytics Taks

- Clustering:Partition data into meaningful groups
- Outlier Detection:Detect points that are unusual (unlike others)
- Classification:Assign (predefined) class labels to each object
- Regression:Find a function that models (continuous) target variable
- Association Analysis:Find patterns in data that describe relationships
- Recommendation:Predict an unknown rating based on known ratings
- Community Detection:Find (overlapping) communities of nodes in networks
- Centrality and Important nodes:Find important (or evaluate importance of) nodes in networks

Machine Learning for Data Analytics

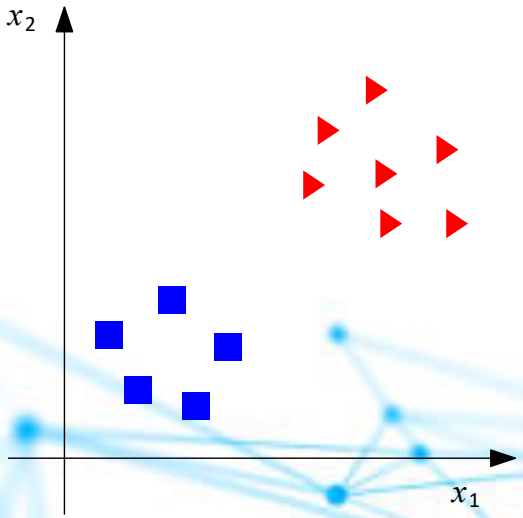
Supervised Learning

- For some data items the correct results (values of the target variable) are given(ground truth)
- We want to learn a model that generalizes i.e. the model is able to perform accurately on new/unseen/unlabeled data items
- Classification, where the target is a categorical attribute
- Regression, where the target is a continuous attribute

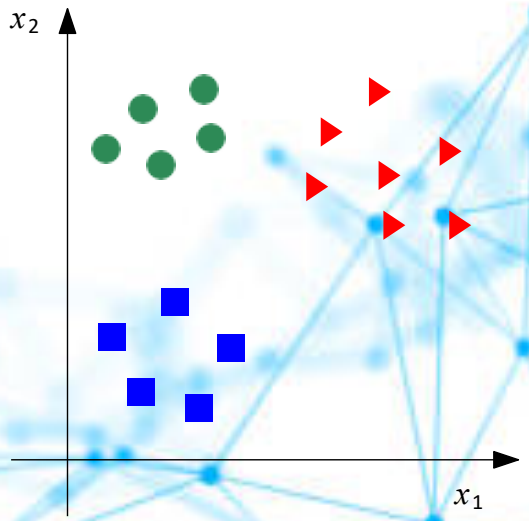


Machine Learning for Data Analytics

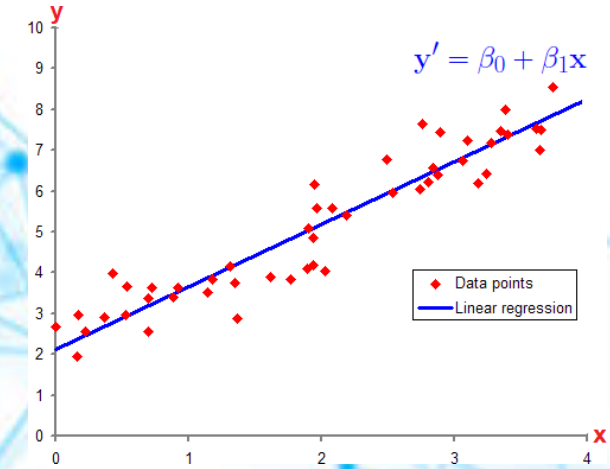
Binary Classification



Multi-Class Classification



Regression



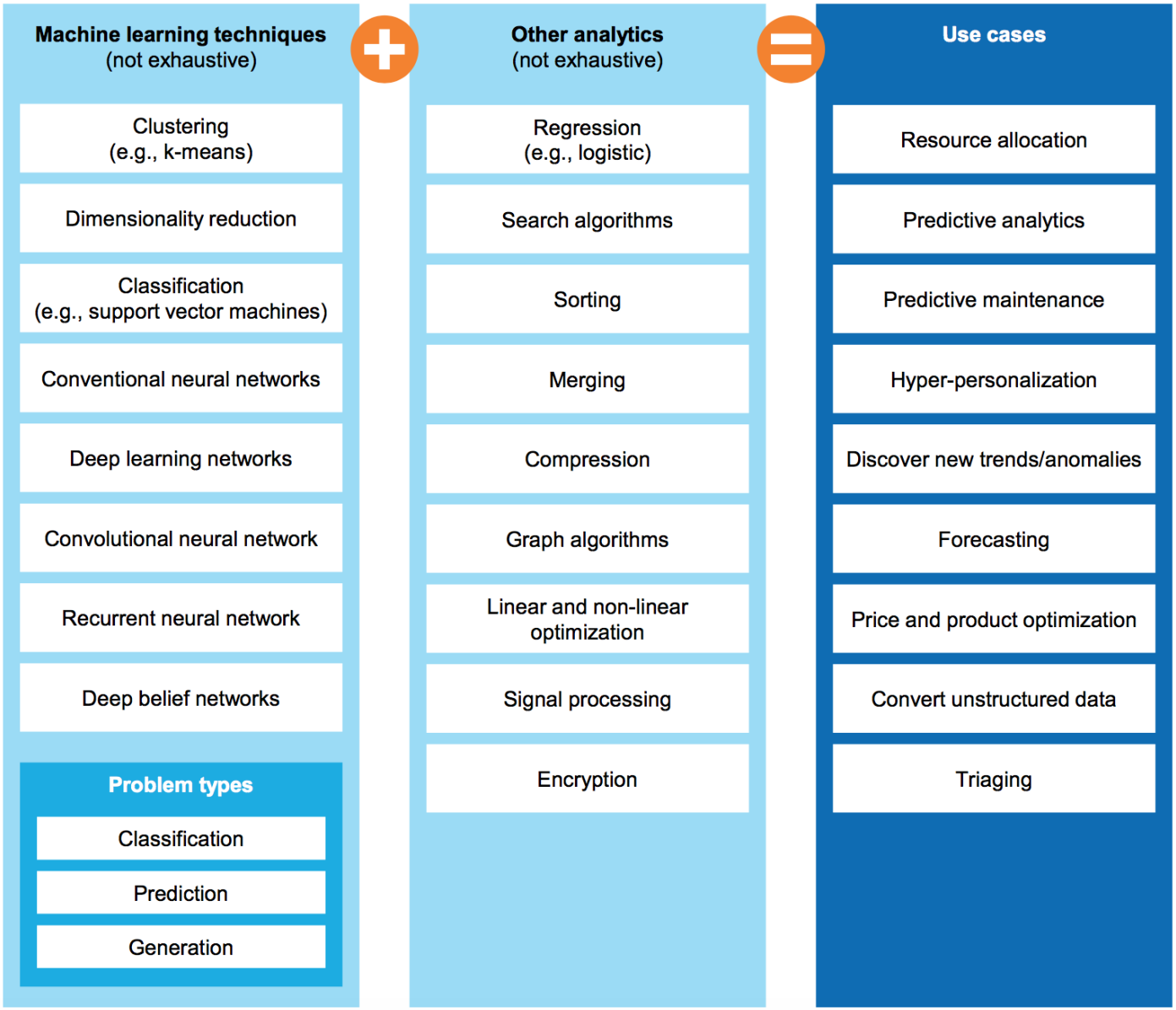
Machine Learning for Data Analytics

Unsupervised Learning

- No correct output is provided
- Learning and analytics is done using statistical properties of data
- Clustering Outlier detection
- Modeling the density Data
- Dimensionality Reduction

Data Analytics Tasks and Methods

Machine learning can be combined with other types of analytics to solve a large swath of business problems



SOURCE: McKinsey Global Institute analysis

Graphical EDA

Diagrammatic Representations of Data

- Easy to understand: Numbers do not tell all the story. Diagrammatic representation of data makes it easier to understand
- Simplified Presentation: Large volumes of complex data can be represented in a simplified and intelligible diagram
- Reveals hidden facts: Diagrams help in bringing out the facts and relationships between data not noticeable in raw/tabular form
- Easy to compare: Diagrams make it easier to compare data

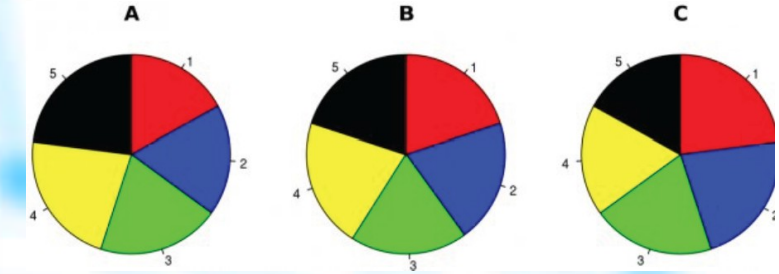
Types of Diagrams

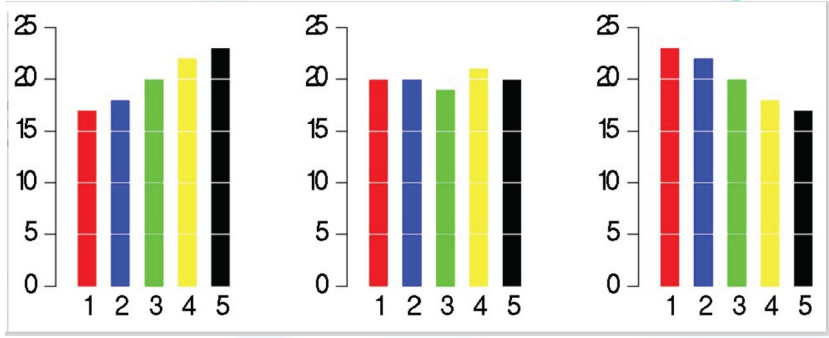
We will briefly discuss and use the following types of diagrams

- Bar Charts
- Histogram
- Box Plot
- Scatter Plot
- Heat map
- Line Graph
- Parallel Axis Plot
- Word-Cloud

Bar charts

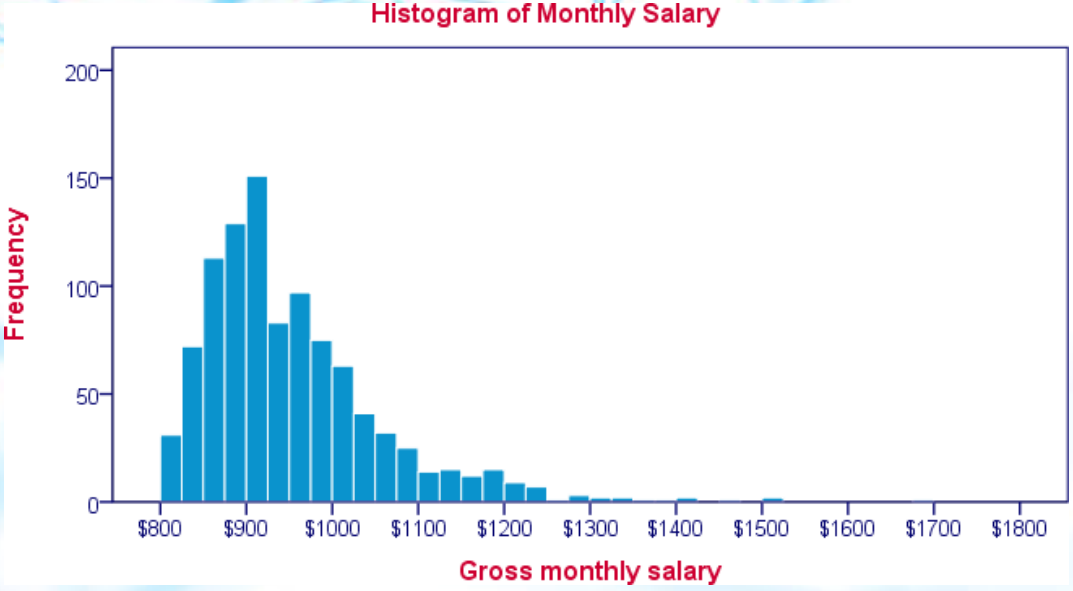
- Generally used for a nominal and ordinal variables
- Different bars (usually colored/shaded differently) for distinct values (levels, categories, symbols) of the variable.
- Height of bar represent frequencies of each symbol (value).
- Can reveal variables that have no or limited information e.g. constants
- Note that we can use pie charts for the same purpose too
- Humans perceive difference in lengths better than in angles





Histograms

- Represent distribution of data in a numeric/continuous variable (estimates probability distribution of a numeric variable)
- Group values by a series of intervals (bins -usually consecutive non-overlapping subintervals covering range of data)
- Plot the number of values falling in each bin (represented by the height of the bar)
- Normalized histogram shows proportion of values in each bin



Histograms

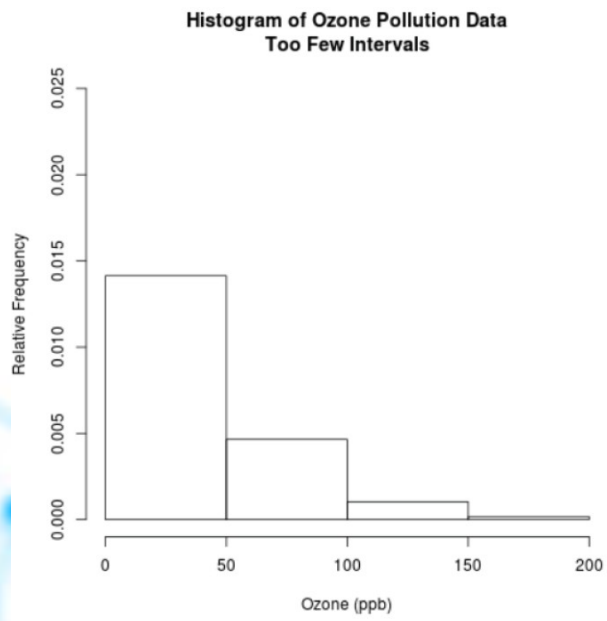
A histogram with appropriate number/length of bins reveals

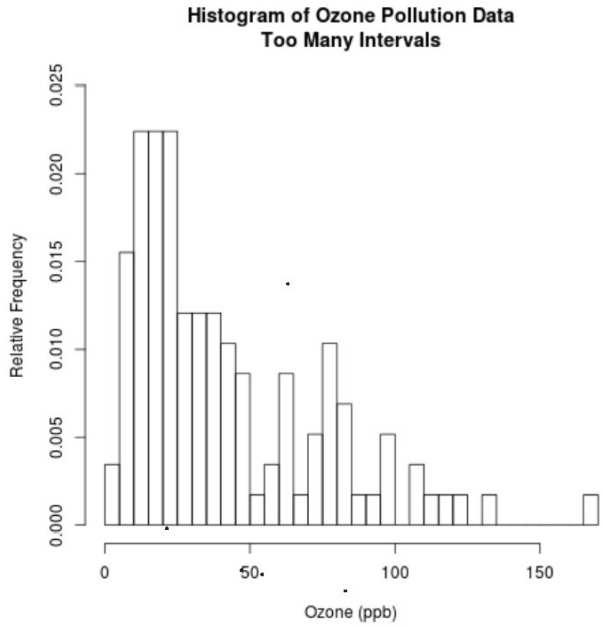
- Where is the data located
- Where/what are the extremes
- What is the distribution of the data
- How the data is spread out
- If the distribution is symmetric or have skew (left or right)
- Can also detect outliers in the data if any

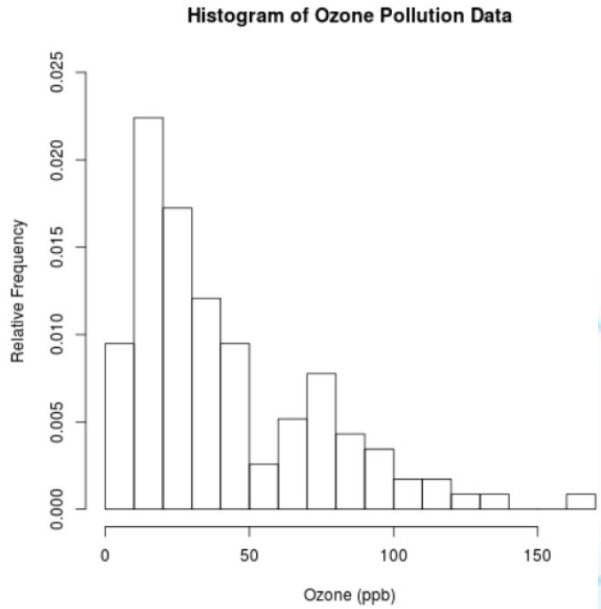
Histograms

- Number and sizes of bins are important considerations
- Bins do not have to be of equal sizes
- For unequal bin sizes height of the bar is not the frequency of values in the bin, it is the frequency density
  - Area of the bar is proportional to the frequency
  - Number of items per unit of the variable of *x*-axis
- Too many bins in histogram gives too much unnecessary details (shows too much noise)
- Too few bins give almost nothing, obscure the underlying patterns

Histograms

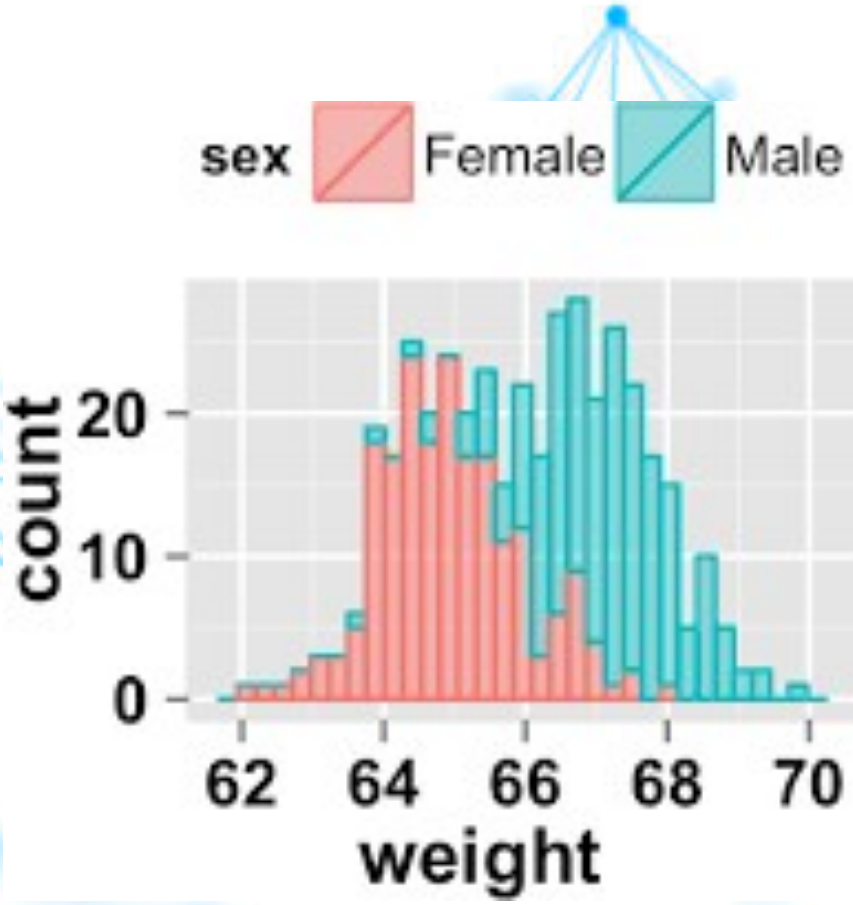

Histogram of Ozone Pollution Data
Too Few Intervals
Relative Frequency
0.000
0.005
0.010
0.015
0.020
0.025
0
50
100
150
200
Ozone (ppb)


Histogram of Ozone Pollution Data
Too Many Intervals
Relative Frequency
0.000
0.005
0.010
0.015
0.020
0.025
0
50
100
150
Ozone (ppb)


Histogram of Ozone Pollution Data
Relative Frequency
0.000
0.005
0.010
0.015
0.020
0.025
0
50
100
150
Ozone (ppb)

Overlapping Histograms

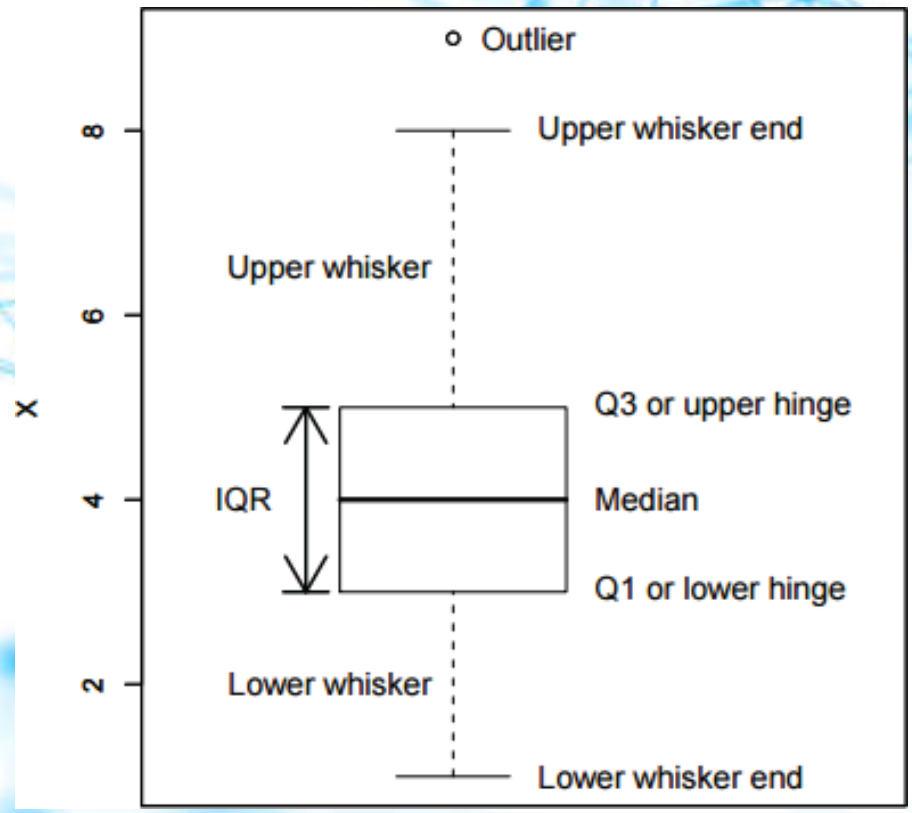
Useful in observing distribution of values with respect to a nominal variable



Box Plots

Another way of displaying the distribution of data (somewhat)

BoxPlots or Box and Whisker diagrams



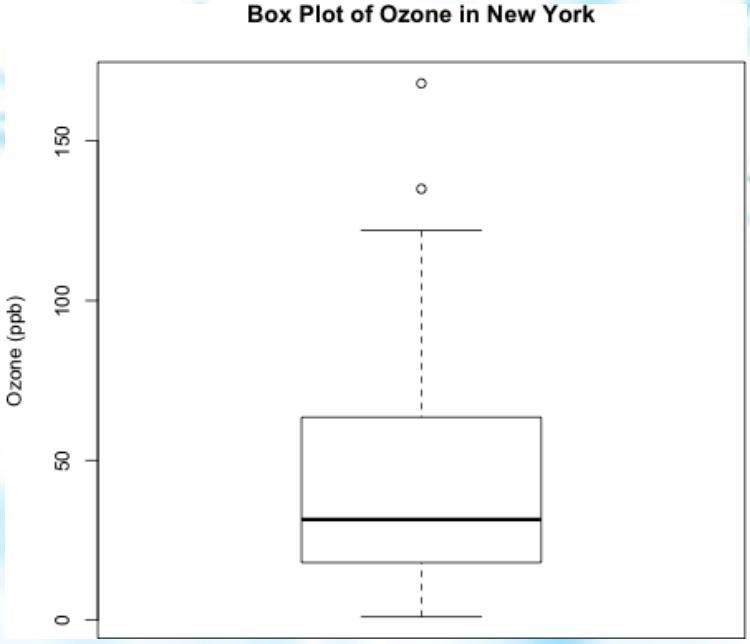
Box Plots

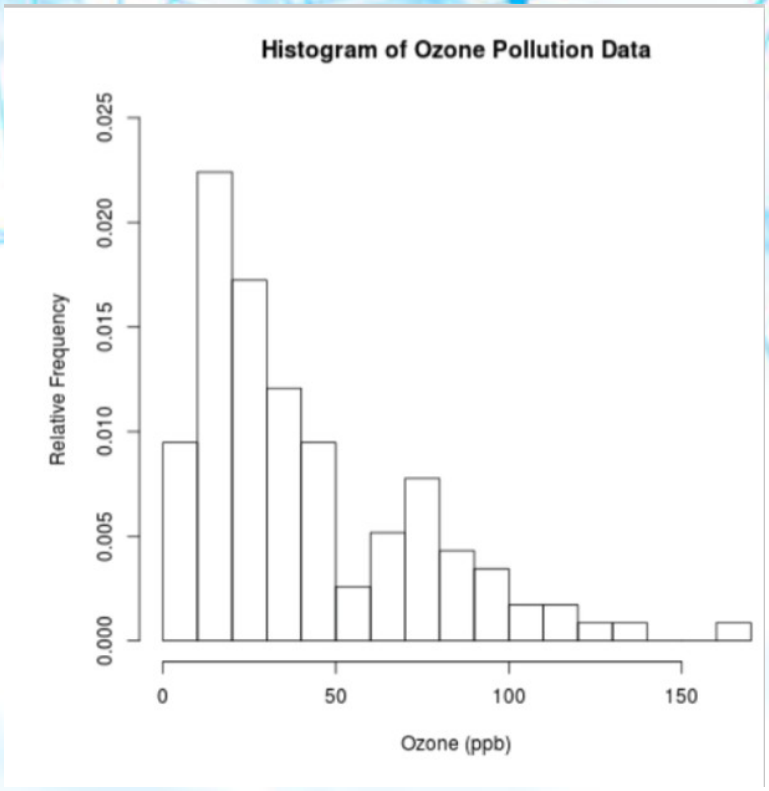
BoxPlots or Box and Whisker diagrams

- Top and bottom lines of the box are 3rd and 1st quartiles of data
- Length of the box is the inter-quartile range (midspread)
- The line in the middle of the box is median of data
- The top whisker denotes the largest value in the data that is within 1.5 times midspread ($Q3 \rightarrow 1.5 \cdot IQR$)
- Similarly the bottom whisker
- Anything above and below the whiskers are considered outliers
- Relative location of median within the box tells us about data distribution
- We find out at what end are the outliers if any

Box Plots

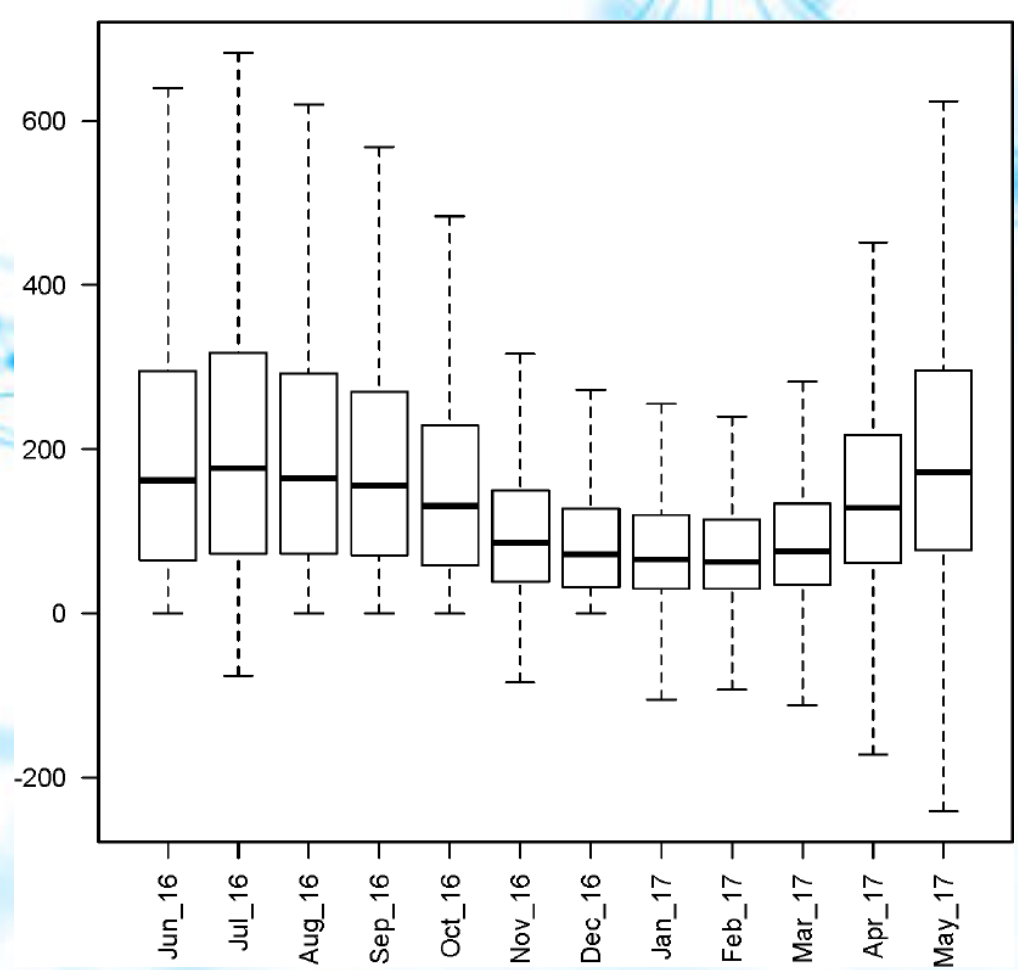
- Can get some idea of skew by observing the shorter whisker





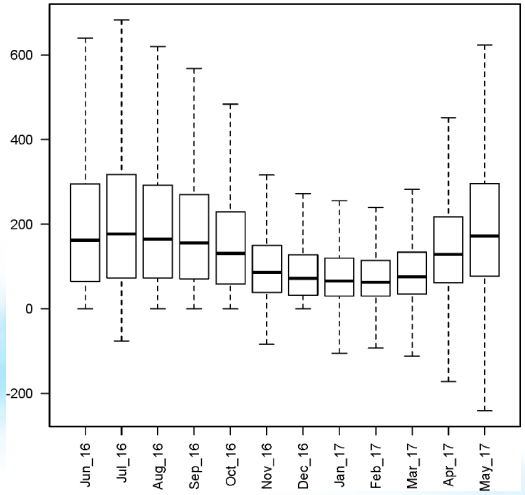
Side-by-side Box Plots

- Extremely useful for comparisons of two or more variables.
- To compare numeric variables, we draw their box-plots in parallel



Side-by-side Box Plots

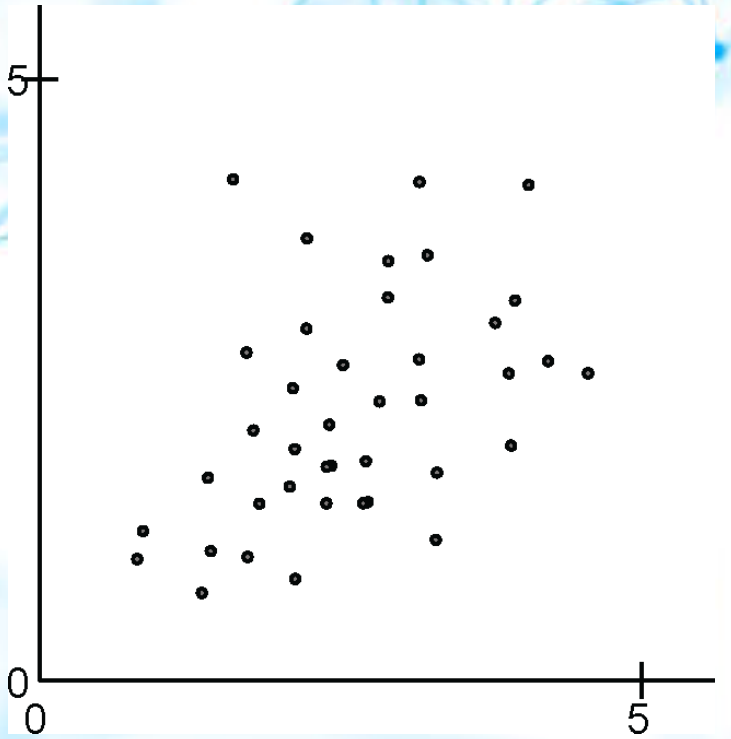
- Side by side groupwise box plots are extremely useful
- Groups are based on values of a categorical variable
- It reveals whether a factor (the categorical variable) is important
- It addresses whether the location of data differ between groups
- To some extent it also reveals whether distribution and variation differ between groups
- Overlapping histograms are more suitable for the latter question, unless there is too much overlap



Scatter Plot

Scatter Plot is the best to visualize two dimensional numeric data

This directly represent the two dimensional observations as points in R^2.
Plot one variable on *x*-axis and other on *y*-axis



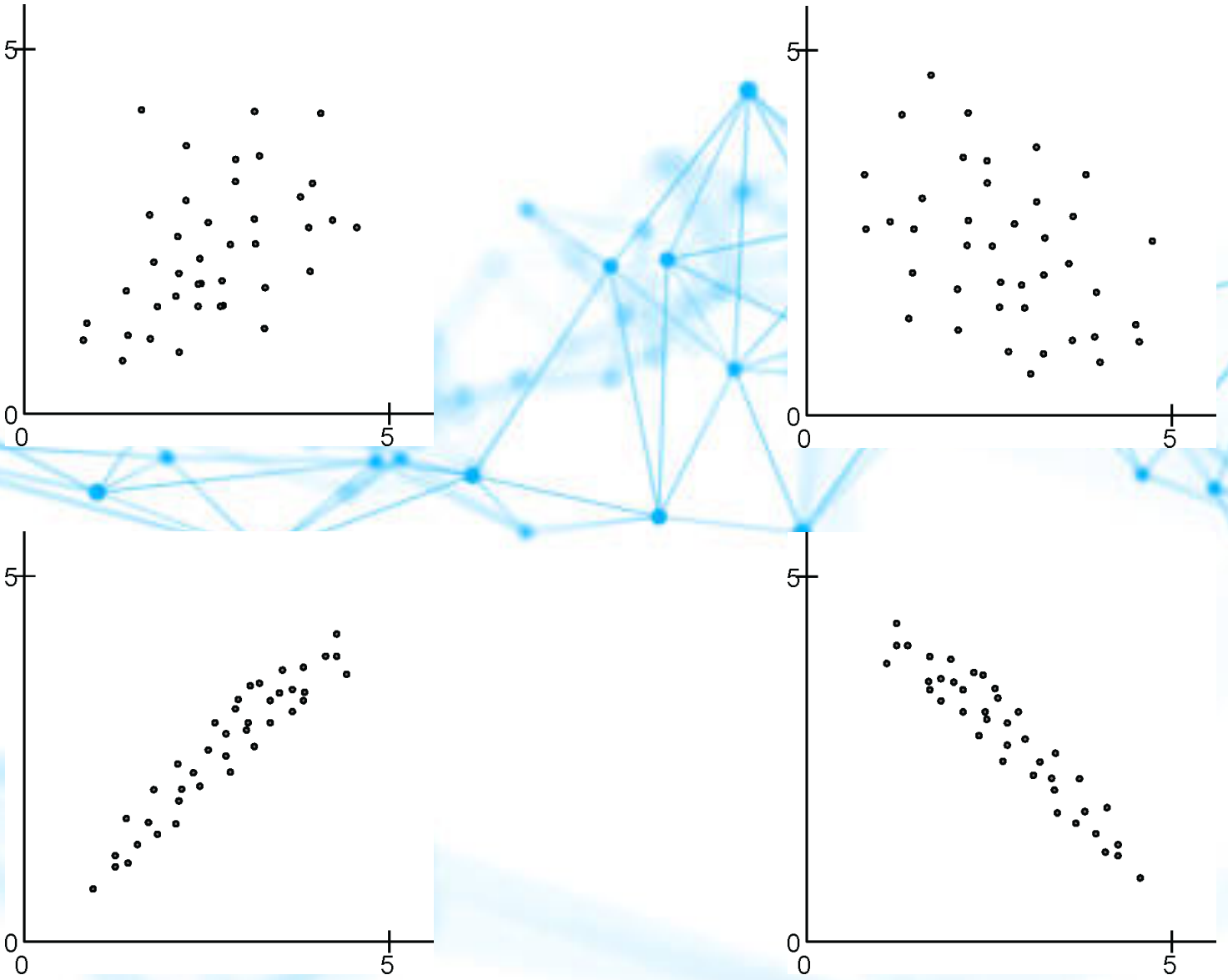
Scatter Plot

Scatter Plot is the best to visualize two dimensional numeric data

This directly represent the two dimensional observations as points in R^2.
Plot one variable on *x*-axis and other on *y*-axis

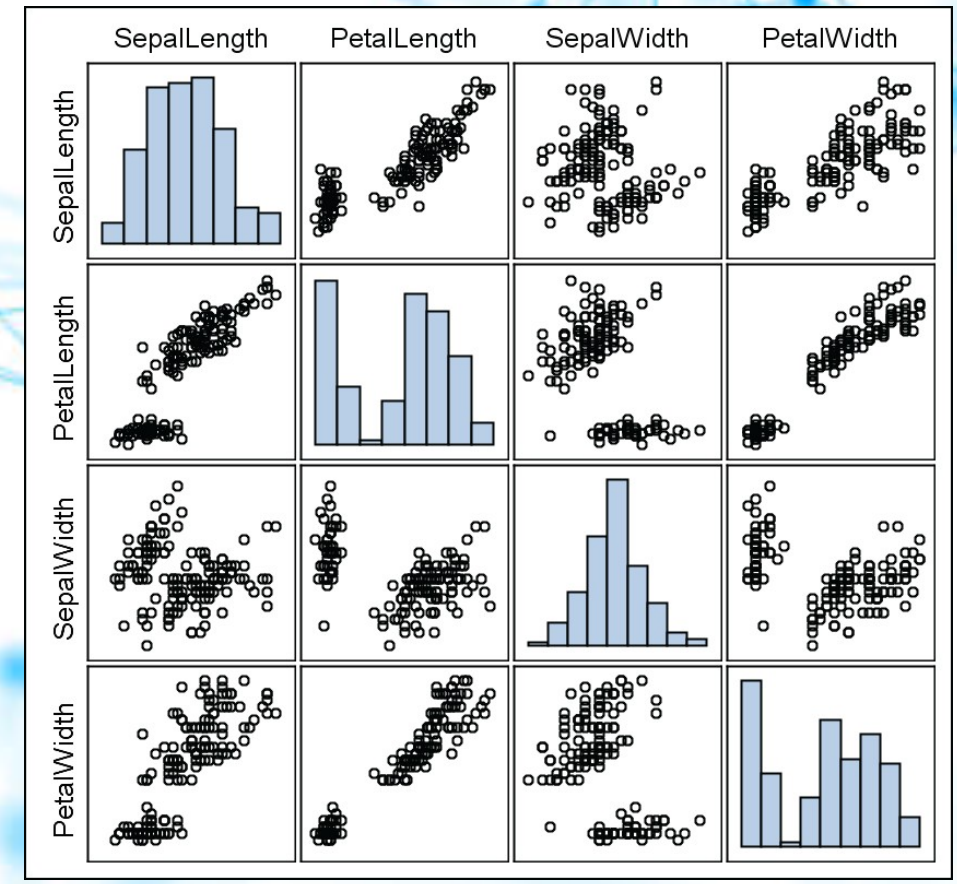
- It shows how the two variables are related to each other
  reveals correlations between the variables
- If one or both variables are highly skewed, then scatter plots are hard to examine, as bulk of the data is concentrated in a small part of plot
- For this we should use some kind of transformation, explained later on one or both the variables
- log-scaled plots can also be used in such cases

Scatter Plot



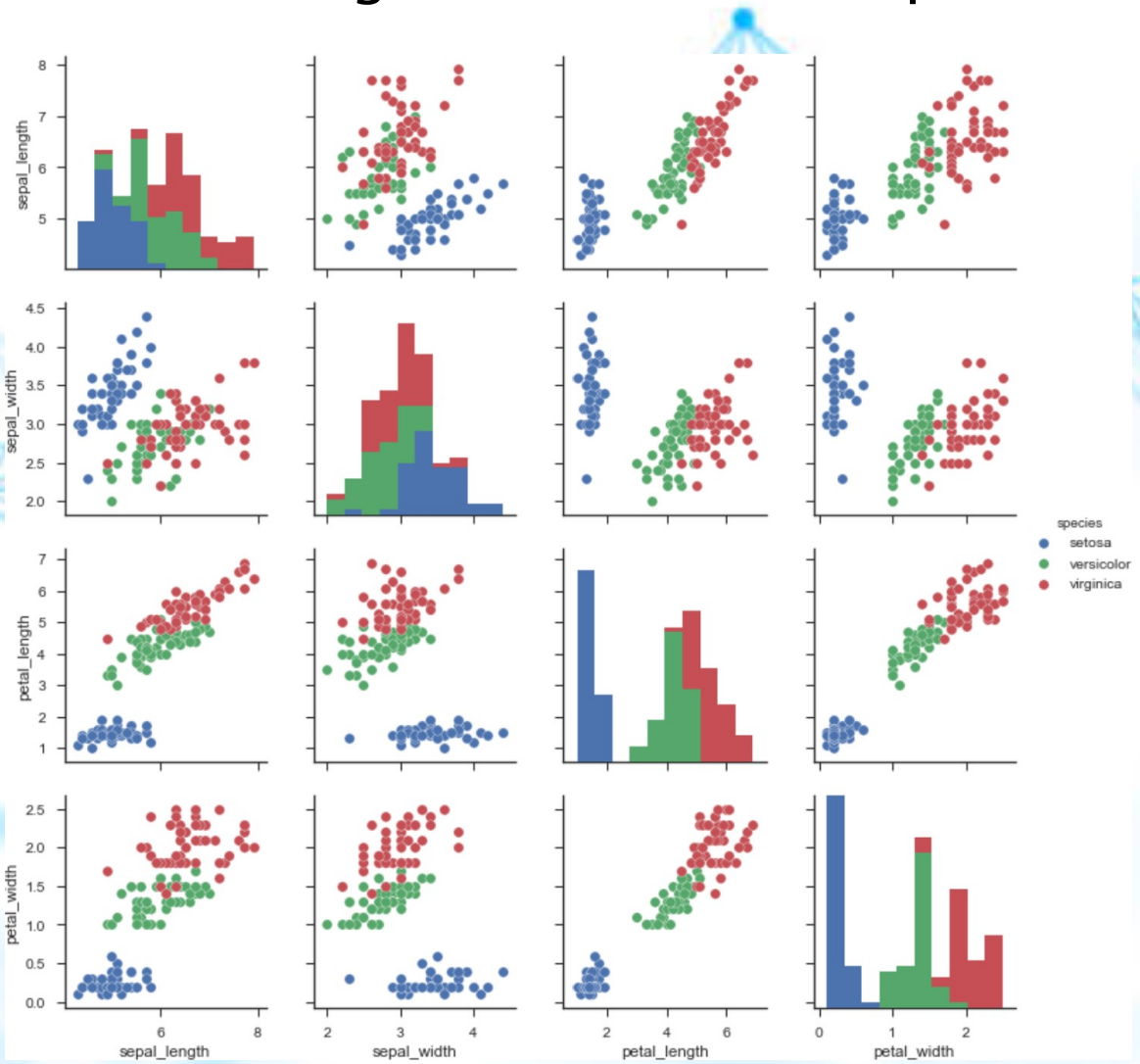
Scatter Plot Matrix

- Pairwise scatter plots, pairwise correlations and individual histograms or density plots
- Summarize the relationships of all pairs of numerical attributes



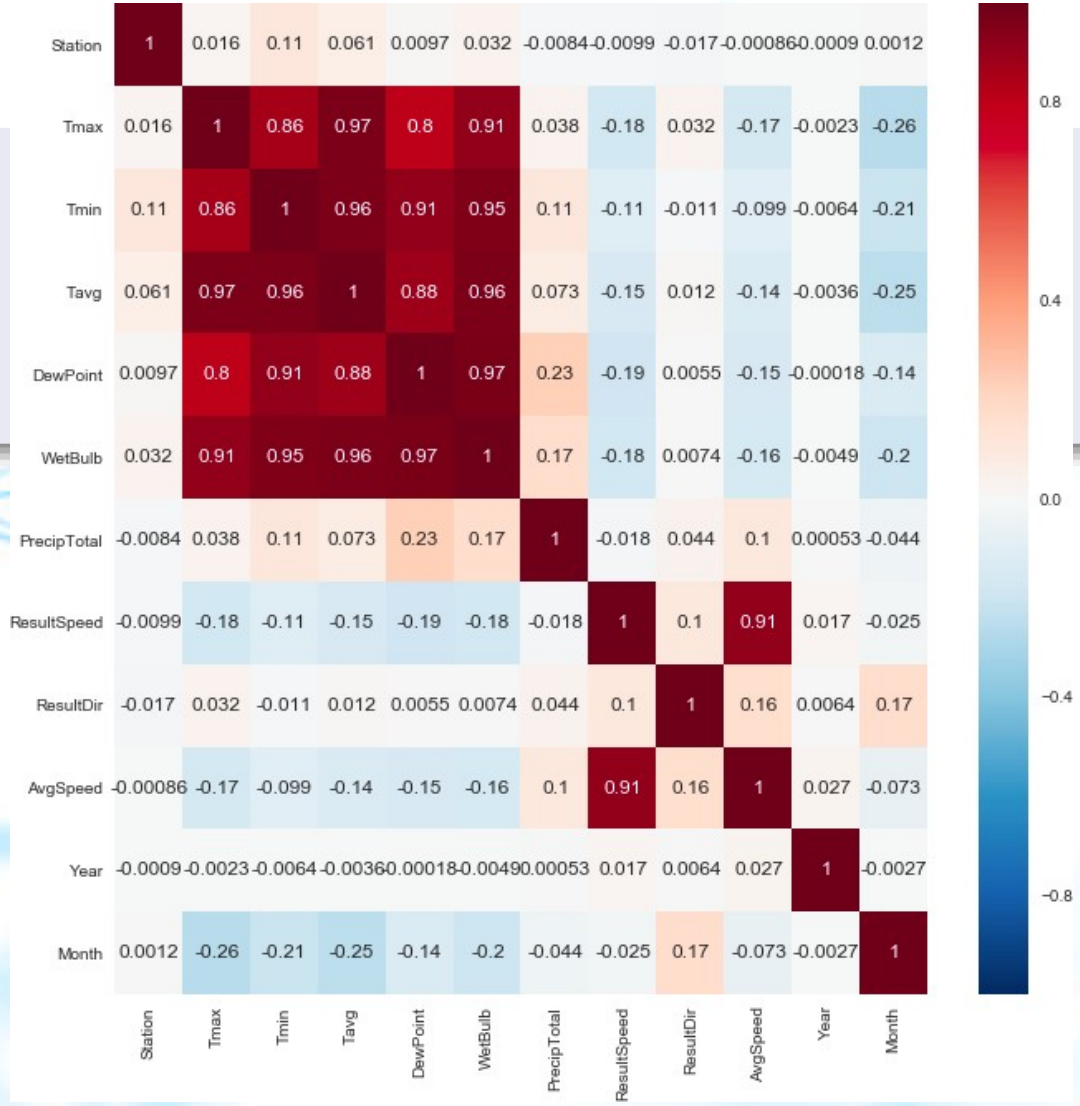
Scatter Plot Matrix

- Scatter plot (matrix) can be combined with information in a nominal attribute encoded through color or marker shape



Heat Map

Presents pairwise relationship between attributes of multivariate data

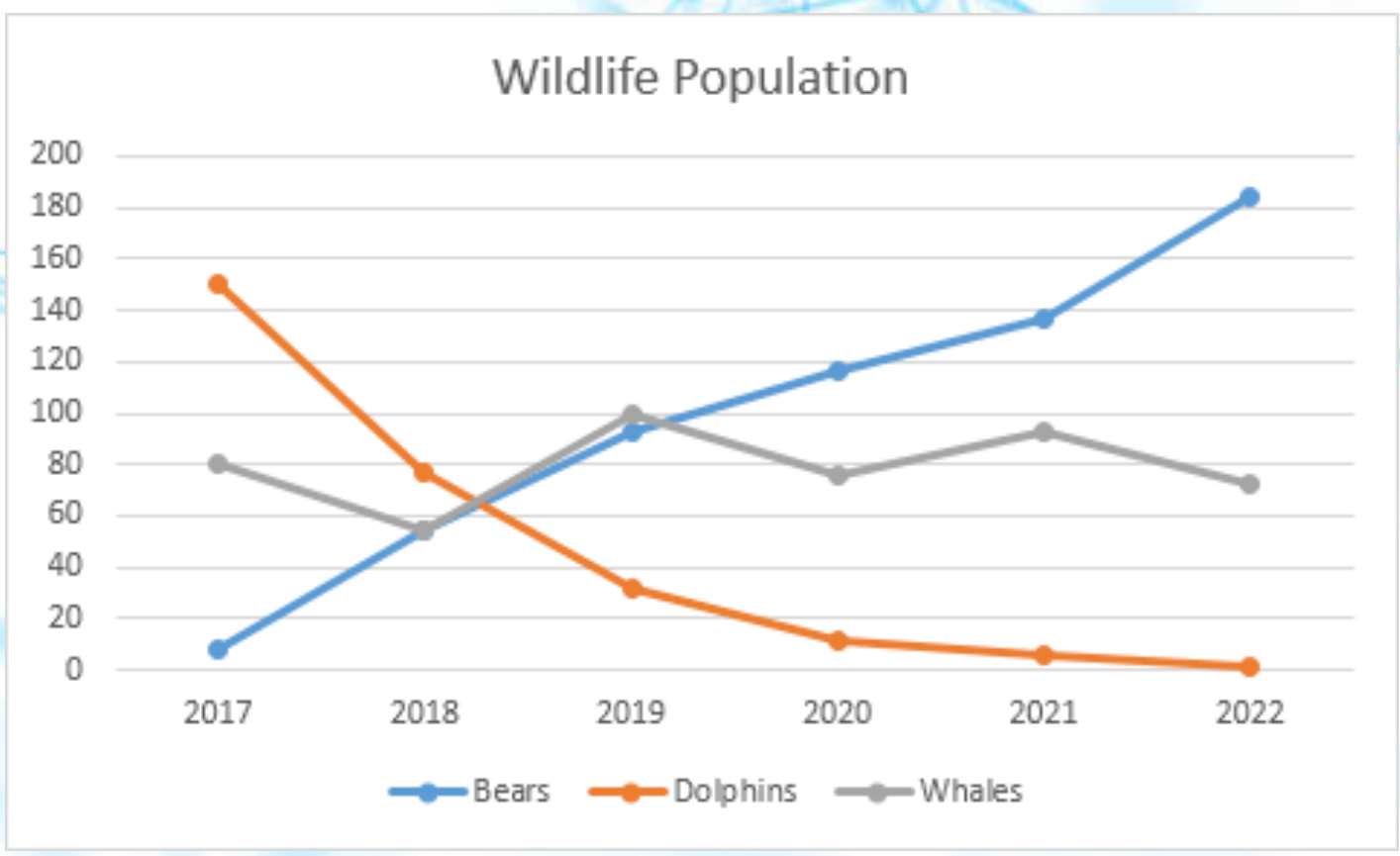


Heat Map

- Presents pairwise relationship between attributes of multivariate data
- Provides a numerical value of the correlation between each variable
- Also provides an easy to understand visual representation of those numbers (colors shades)
- Darker red showing high correlation
- Dark blue showing none or negative correlation
- Can be used to visualize any matrix

Line graphs

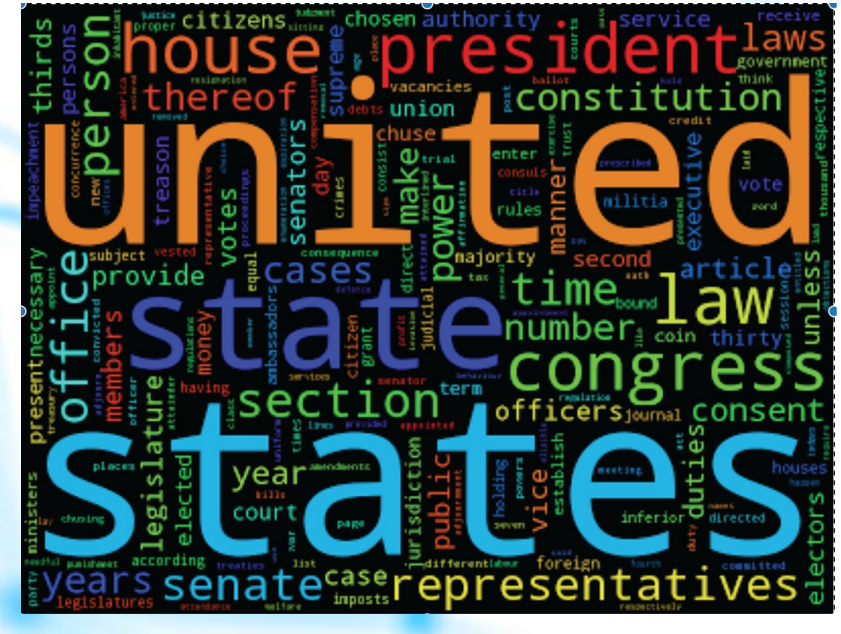
- Line graphs are used for time series e.g. player's yearly average, student's semester gpa or hourly energy consumption
- Two or more time series can be compared in different colors or markers (legend should be provided)



Word-Cloud

Very useful in text analytics

A word cloud shows words used in a text corpus (collection of documents) with size of words proportional to their importance (e.g. TF-IDF)



Quite clear that the word cloud on left is for a collection of articles about US politics, political news, while that on the right seems a corpus of astronomy/astrophysics