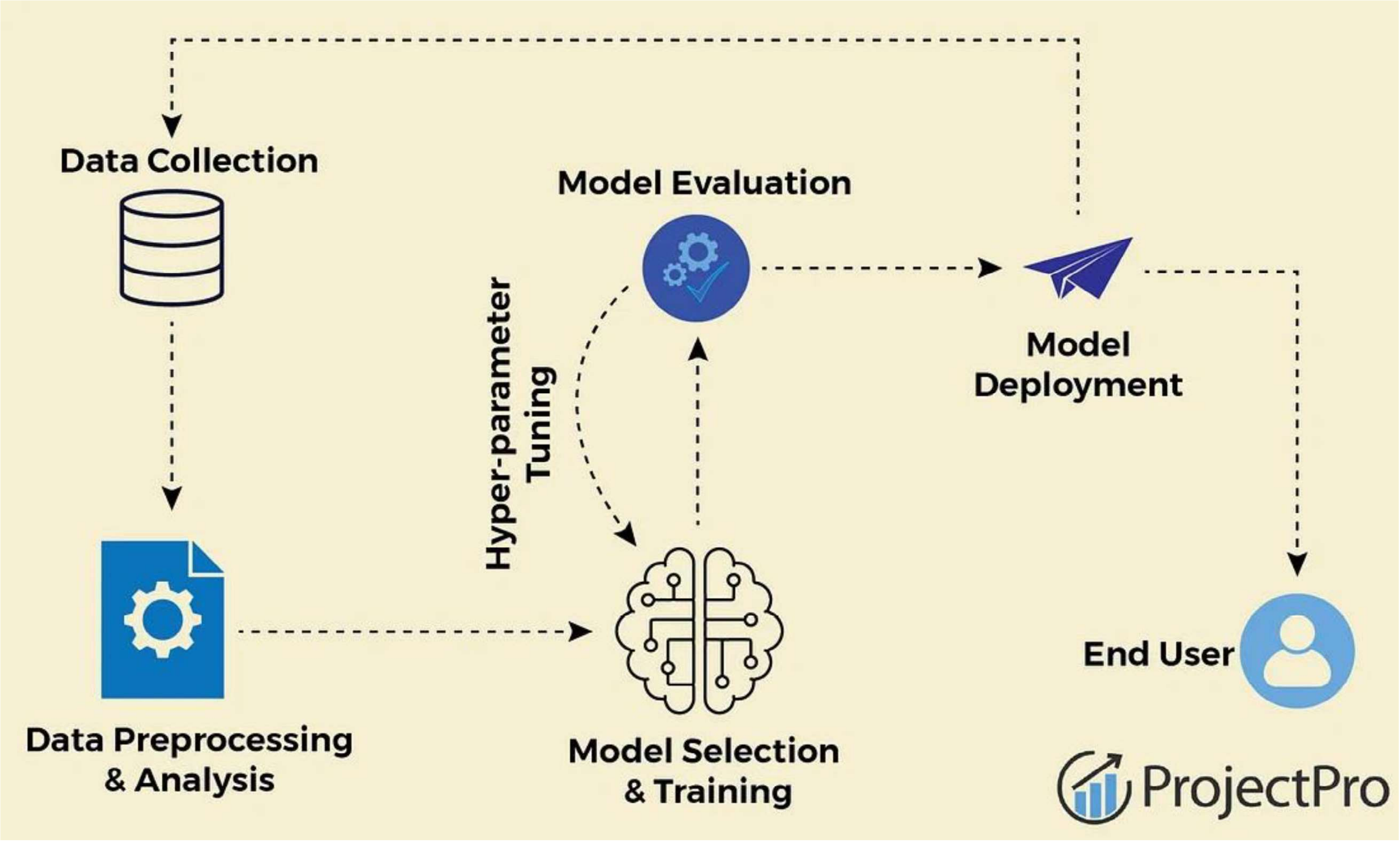
Outline

- Data – Getting the data,
- preparing the data
- Selecting and Training a Model
- Fine tuning a Model:
  - Grid Search
  - Randomized Search
- visualizing the data
- Main Challenges:
  - Data Inadequacy
  - Non-representativeness
  - Irrelevant features
  - Overfitting the Model
  - Underfitting the Model

Select and Train a Model

- A model represents the working of systems in real life.
- It uses mathematical formulas and calculations to predict what is likely to happen based on data recorded about what actually did happen in the past.
- A machine learning model is the output of the training process (Algorithm) and is defined as the mathematical representation of the real-world process.
- Machine Learning models can be understood as a program that has been trained to find patterns within new data and make predictions.
- Ex: Fraud Detection Model, Stock Market Prediction Model
- Algorithm - Machine learning algorithms are procedures that are implemented in code and are run on data. Ex: Regression, Classification, Clustering etc.

Machine Learning Cycle



https://www.projectpro.io/article/machine-learning-model-deployment/872

Hyperparameter tuning

Hyperparameter

- Parameter
  - A Machine Learning model is defined as a mathematical model with several ***parameters*** that need to be learned from the data. By training a model with existing data, we can fit the model parameters.
- Hyperparameters
  - However, there is another kind of parameter, known as Hyperparameters, that cannot be directly learned from the regular training process. They are usually fixed before the actual training process begins. These parameters express essential properties of the model, such as its complexity or how fast it should learn.

Hyperparameter

- KNN
  - K neighbours
- SVM
  - Kernel, penalty
- Decision tree classifier
  - Attribute selection measure, max depth, minimum number of samples required at a leaf
- How many trees should I include in my random forest.
- Clustering
  - #clusters, distance metric, density measures
- Linear model
  - Degree of polynomial
- ANN
  - # iterations, learning rate, dropout, #layers, batch size, pooling

Hyperparameter tuning

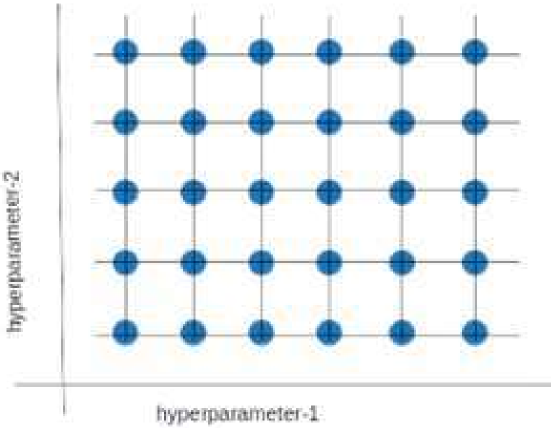
- Manual Search
- Grid Search
- Random Search
- Halving
  - Grid Search
  - Randomized Search
- Automated Hyperparameter tuning
  - Bayesian Optimization
  - Genetic Algorithms
- Artificial Neural Networks Tuning
- HyperOpt-Sklearn
- Bayes Search

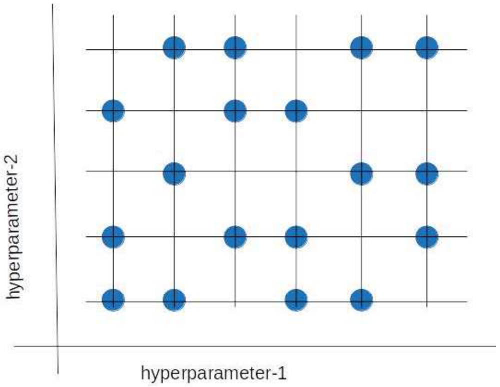
Grid Search

- Manual Search
  - While this helps to gain intuition into the decision surface and identify promising regions in the search space, this involves trial and error and is computationally expensive. It requires that researchers have domain expertise for the given data and experience working with ML/DL techniques.
- Grid Search
  - Grid search is a sort of "brute force" hyperparameter tuning method.
  - A grid of possible discrete hyperparameter values is created, and then fit the model with every possible combination.
  - The model performance for each set is recorded, and then select the combination that has produced the best performance.
  - Grid search of the hyperparameter space is a popular method that is simple to implement and parallelize and provides insight into the search space.
  - Grid search is a popular technique for hyperparameter tuning which performs an exhaustive search with every possible combination of the HPs. It becomes computationally expensive as the number of HPs increases.
  - Early stopping to the rescue
  - Can take advantage of parallelism

Random search

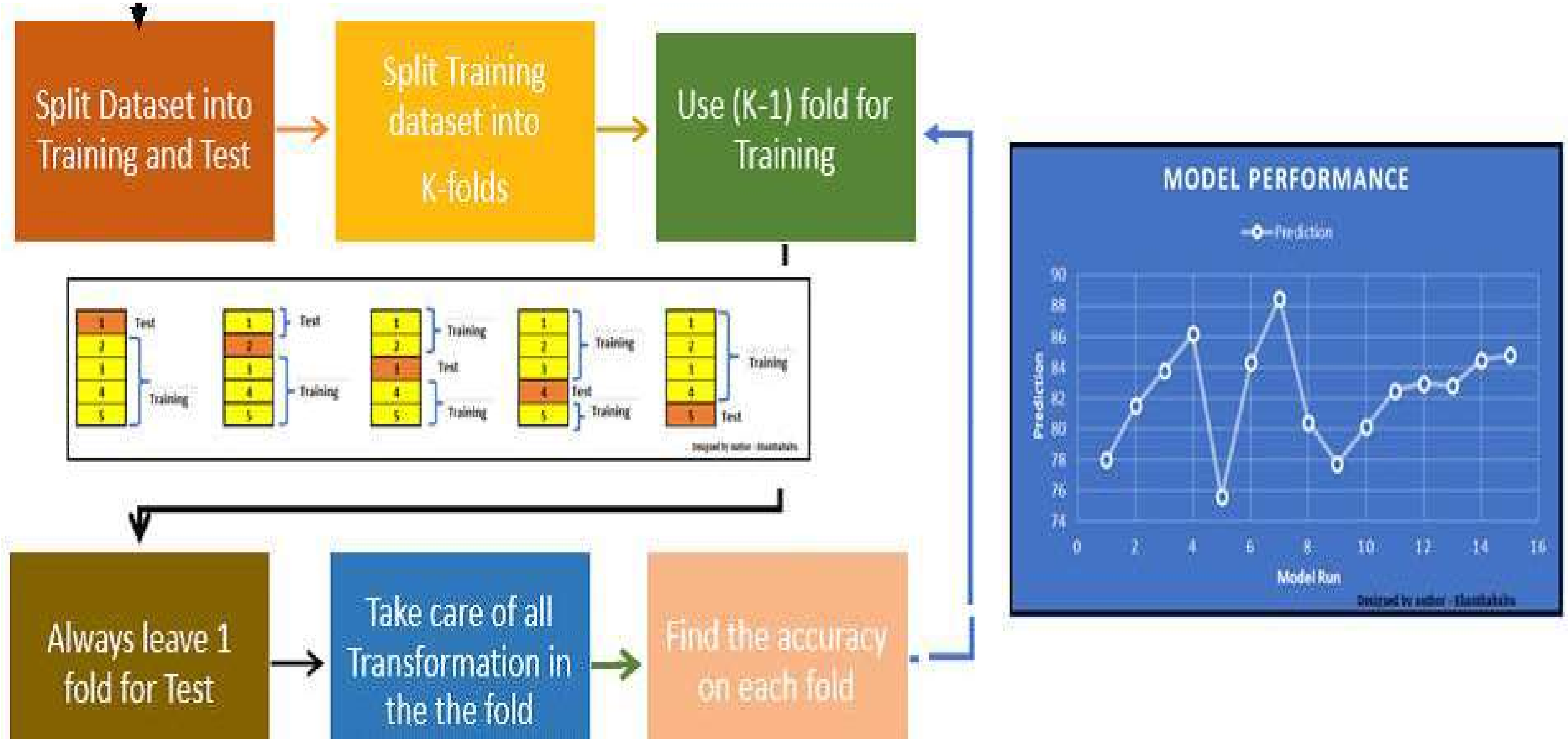
- Random search draws independent sets from the HP search space using statistical distributions of the HPs. When the number of HPs is high, the random search can effectively search a larger space compared to grid search.
- Random search tries a random combination of hyperparameters in each iteration and records the model performance. After several iterations, it returns the mix that produced the best result.





Cross-Validation

- K-fold cross-validation
- Leave one our cross-validation


Split Dataset into Training and Test
Split Training dataset into K-folds
Use (K-1) fold for Training
1 2 3 4 5
Test
Training
Always leave 1 fold for Test
Take care of all Transformation in the the fold
Find the accuracy on each fold
MODEL PERFORMANCE
Prediction
Prediction
90
88
86
84
82
80
78
76
74
0
2
4
6
8
10
12
14
16
Model Run
Designed by author

Data visualization

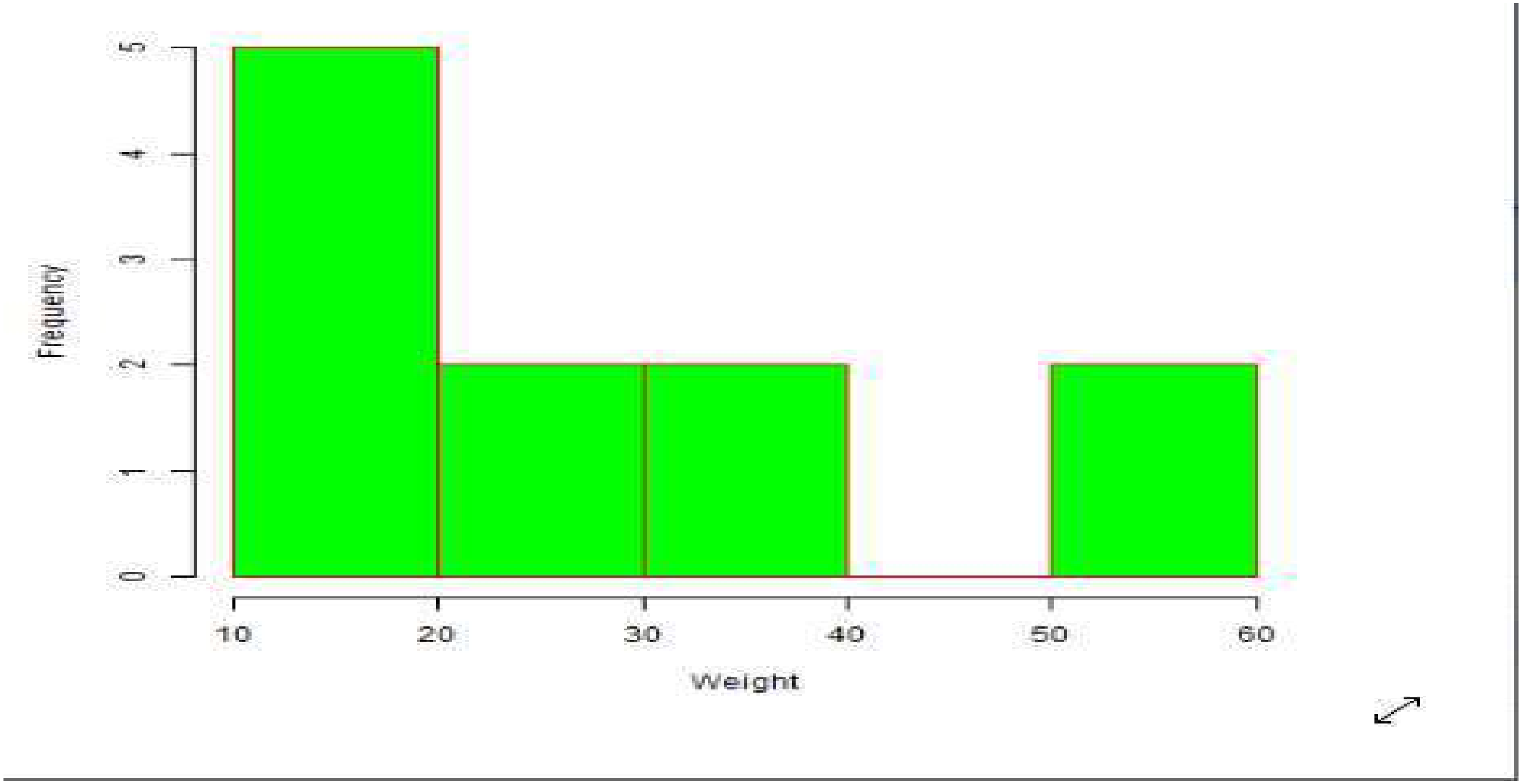
- For gaining insight about data
- Easily obtain hidden patterns in data
- Fast and productive way to convey the message

Standard Graphics

- Histogram
- Scatterplots
- Heat map
- Piecharts
- Barplots

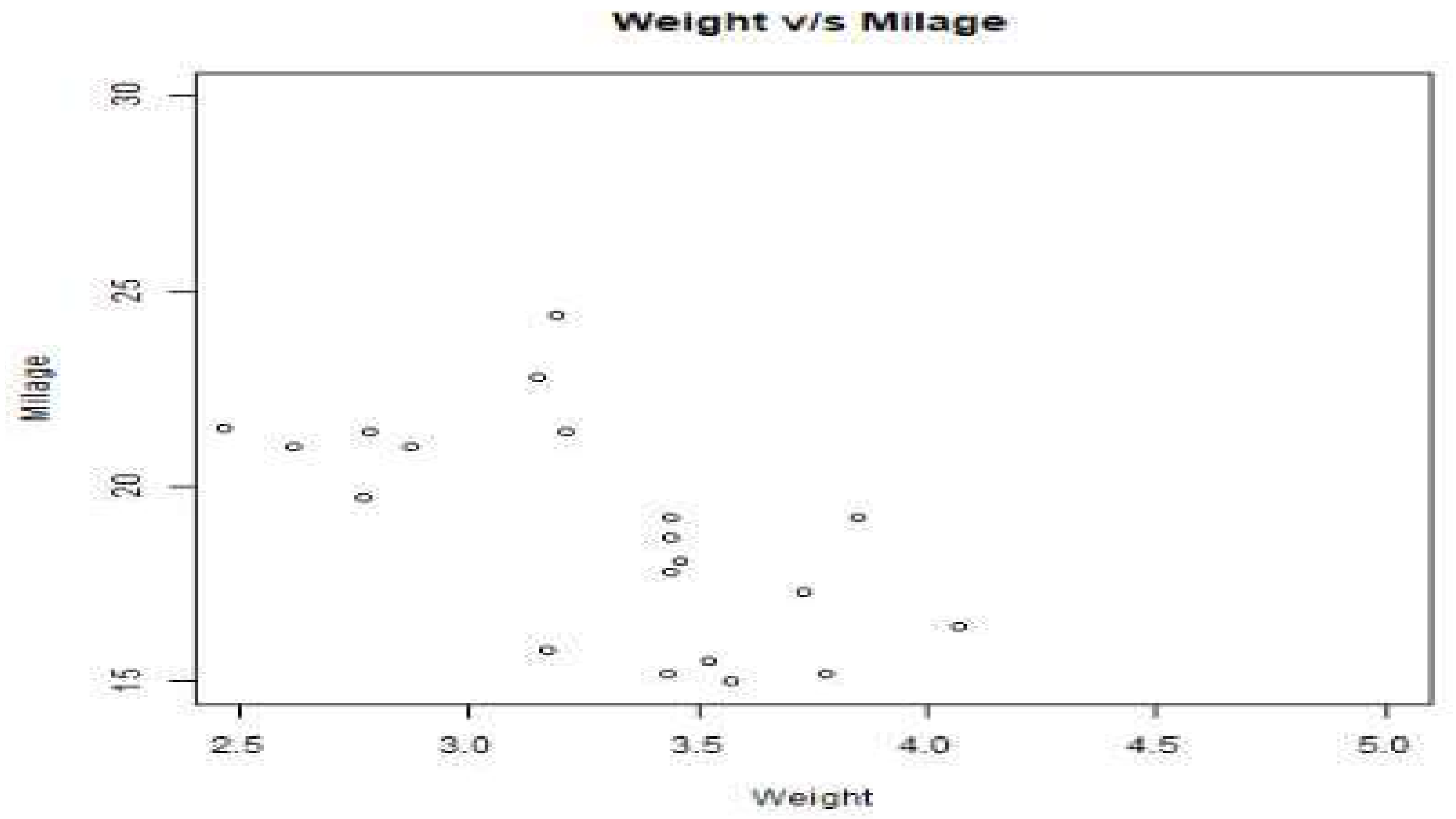
R Histogram

- analyze the distribution of data
- Type of bar chart - shows the frequency of the number of values which are compared with a set of values ranges



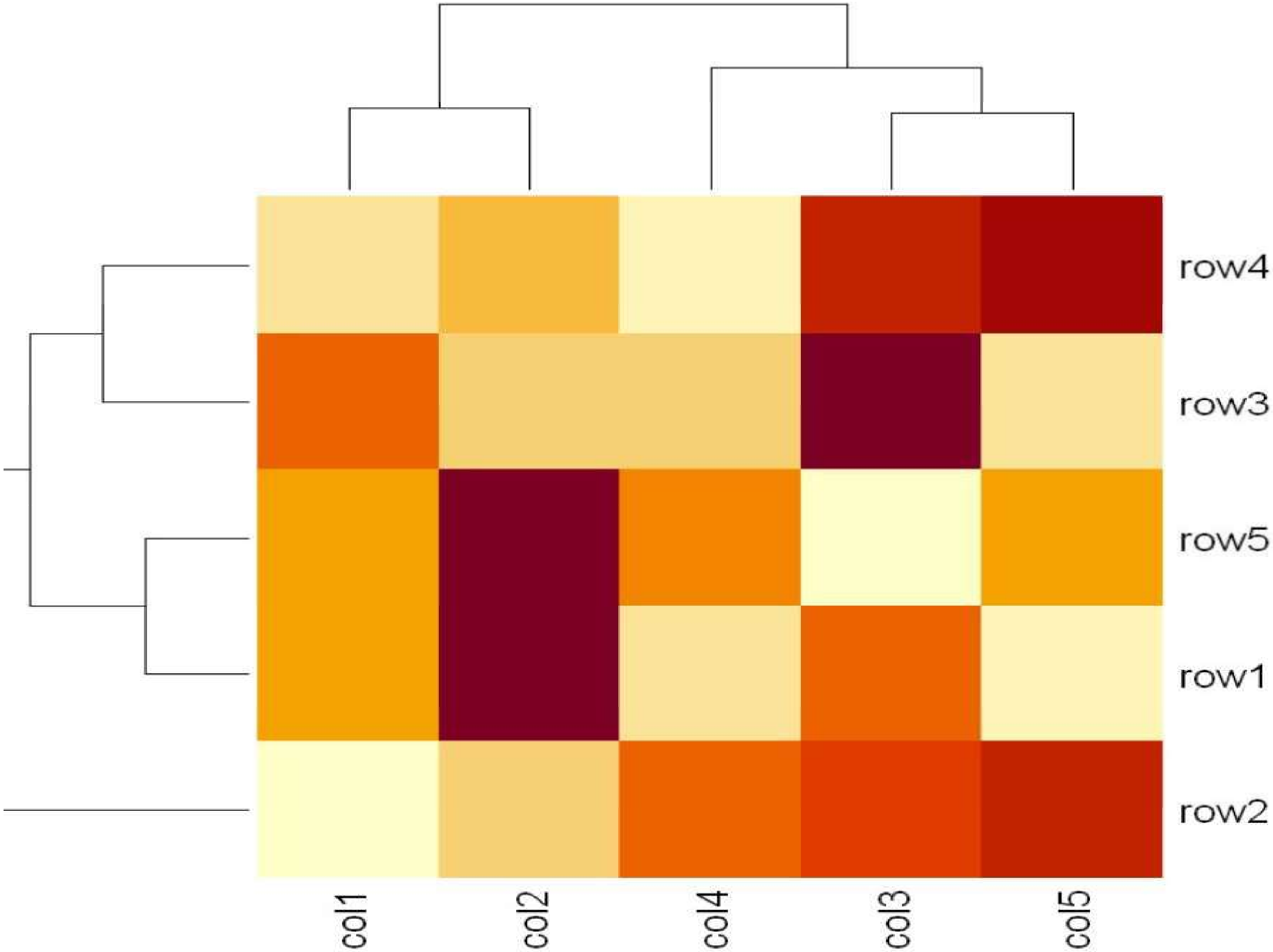
R Scatterplots

- visualize the relationship between two continuous variables.
- A comparison between variables is required when we need to define how much one variable is affected by another variable
- The data is represented as a collection of points



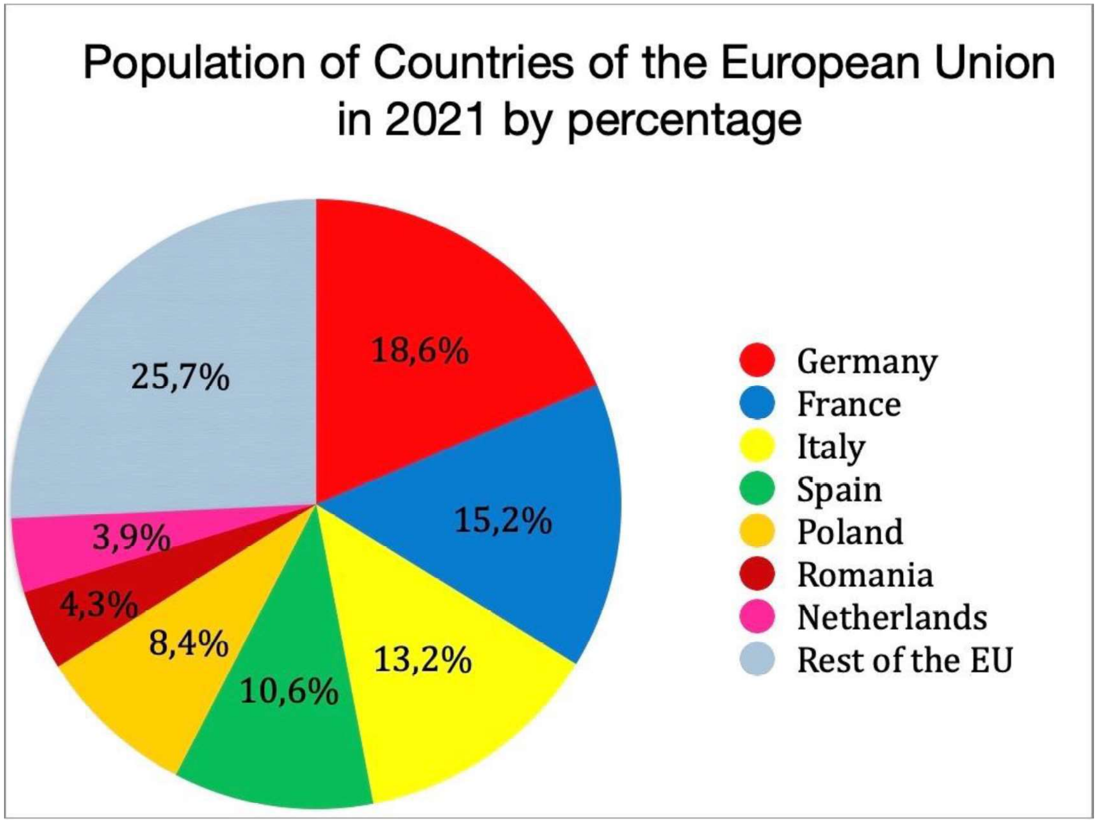
Heat Map

- Use it for the intensity of colours
- Display a relationship between two or three or many variables in a two-dimensional image
- Explore two dimensions of the axis and the third dimension by an intensity of colour


row4
row3
row5
row1
row2
col1
col2
col4
col3
col5

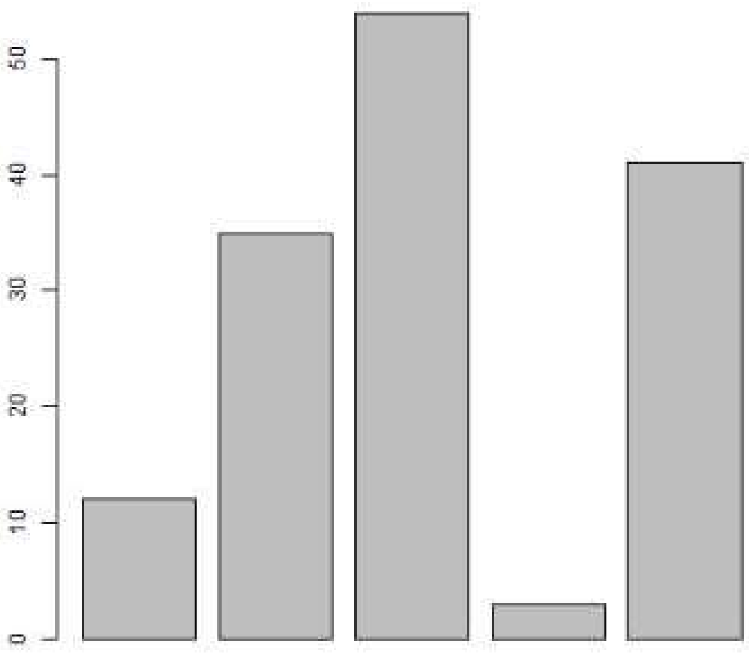
Pie Charts

- A pie-chart is a representation of values in the form of slices of a circle with different colors
- Slices are labeled with a description, and the numbers corresponding to each slice are also shown in the chart



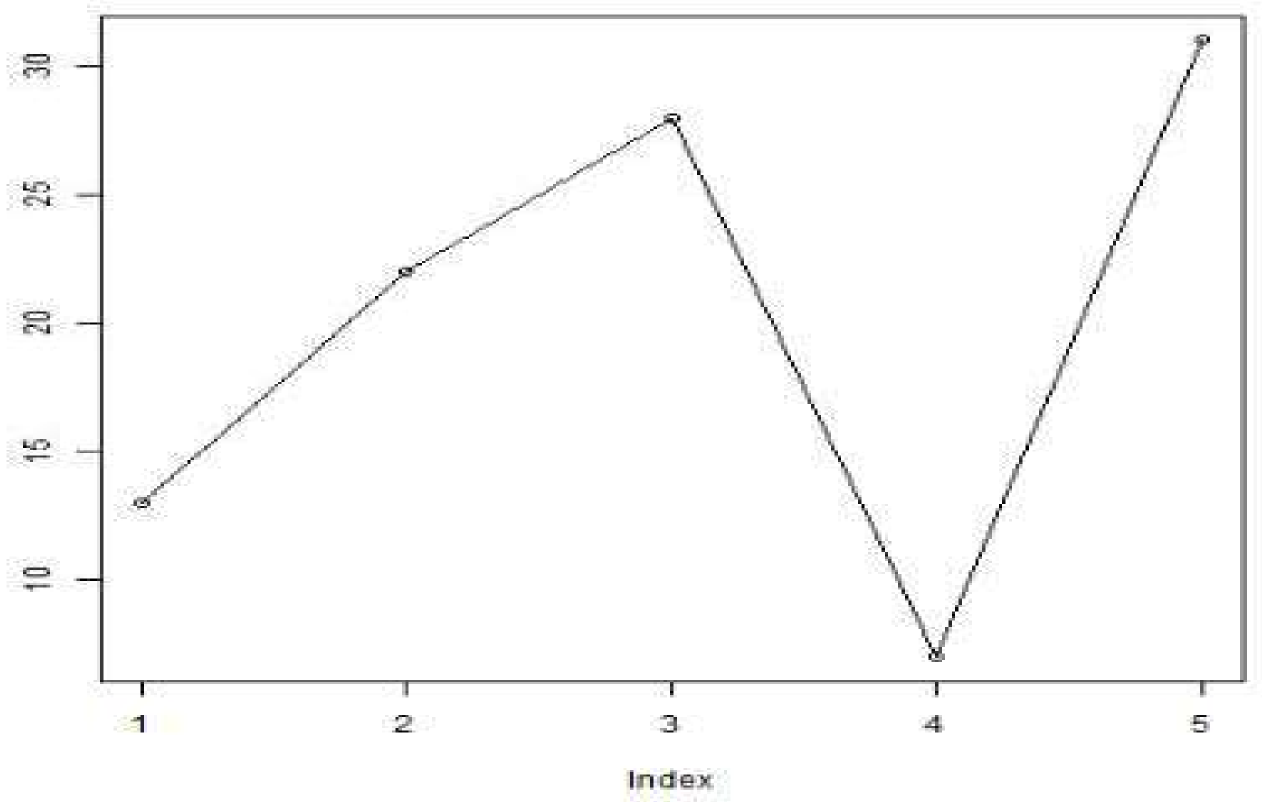
Bar Charts

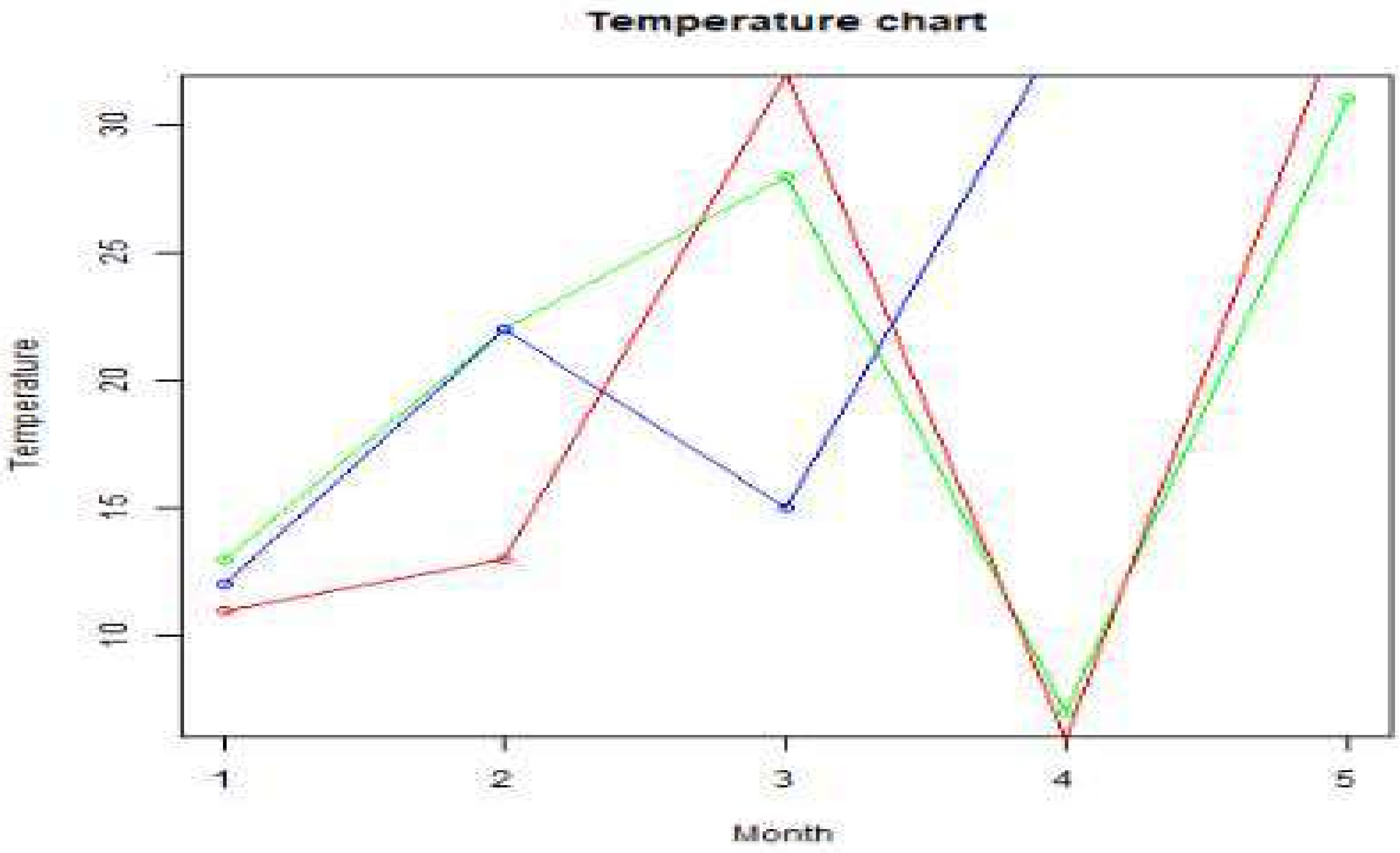
- A bar chart is a pictorial representation in which numerical values of variables are represented by length or height of lines or rectangles of equal width



R Line Graphs

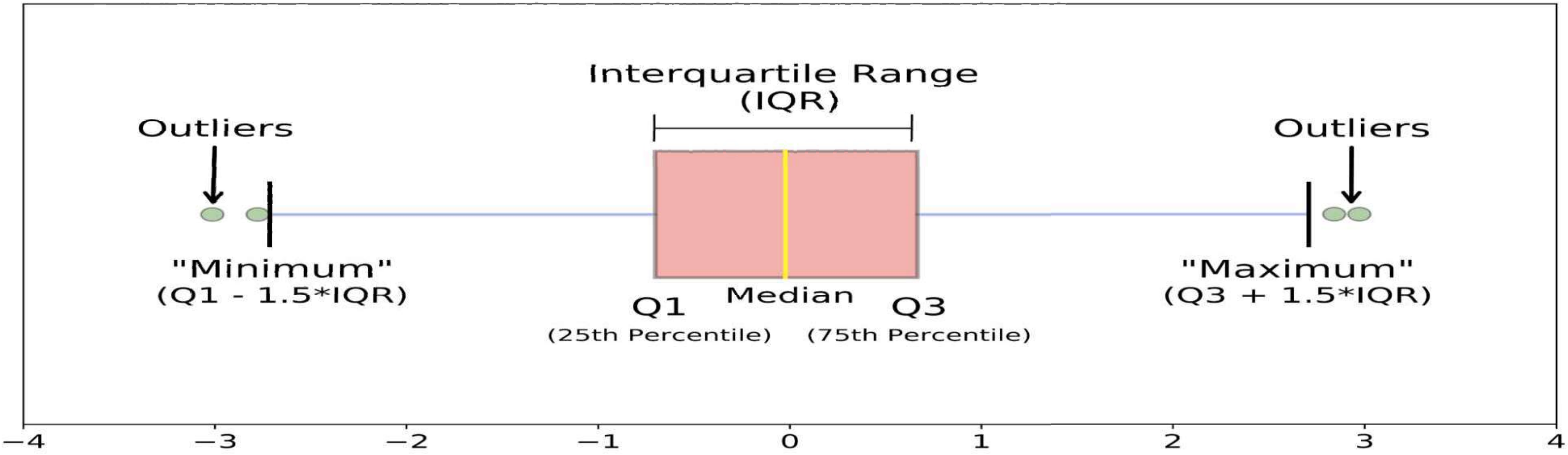
- indicate trends and evaluate how the data has changed over time.

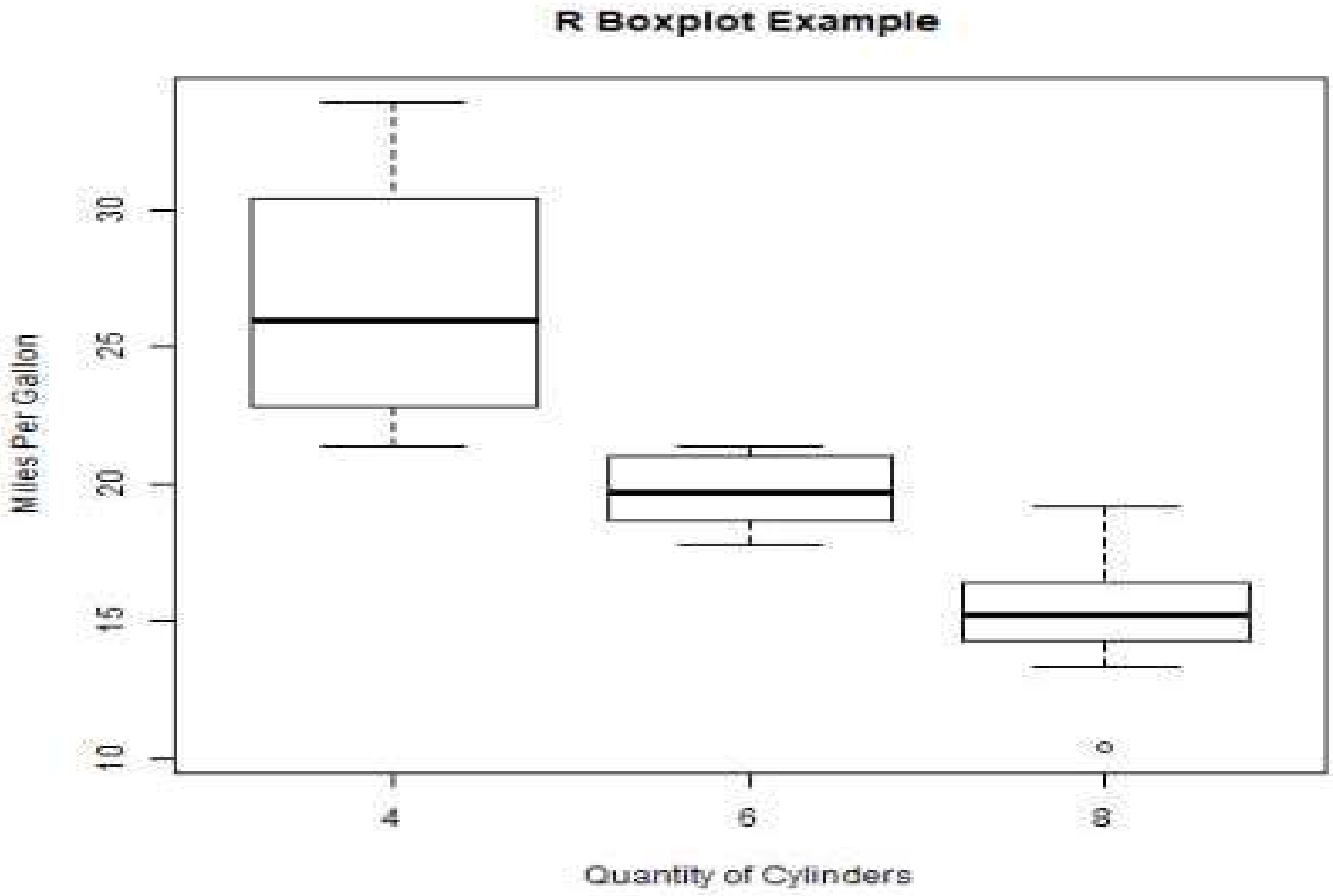



Temperature chart
Temperature
10
15
20
25
30
1
2
3
4
5
Month

Boxplots

- Measure of how well data is distributed across a data set
- divides the data set into **quartiles**
- useful in comparing the distribution of data in a data set
- used for visualizing the spread of the data and detecting outliers

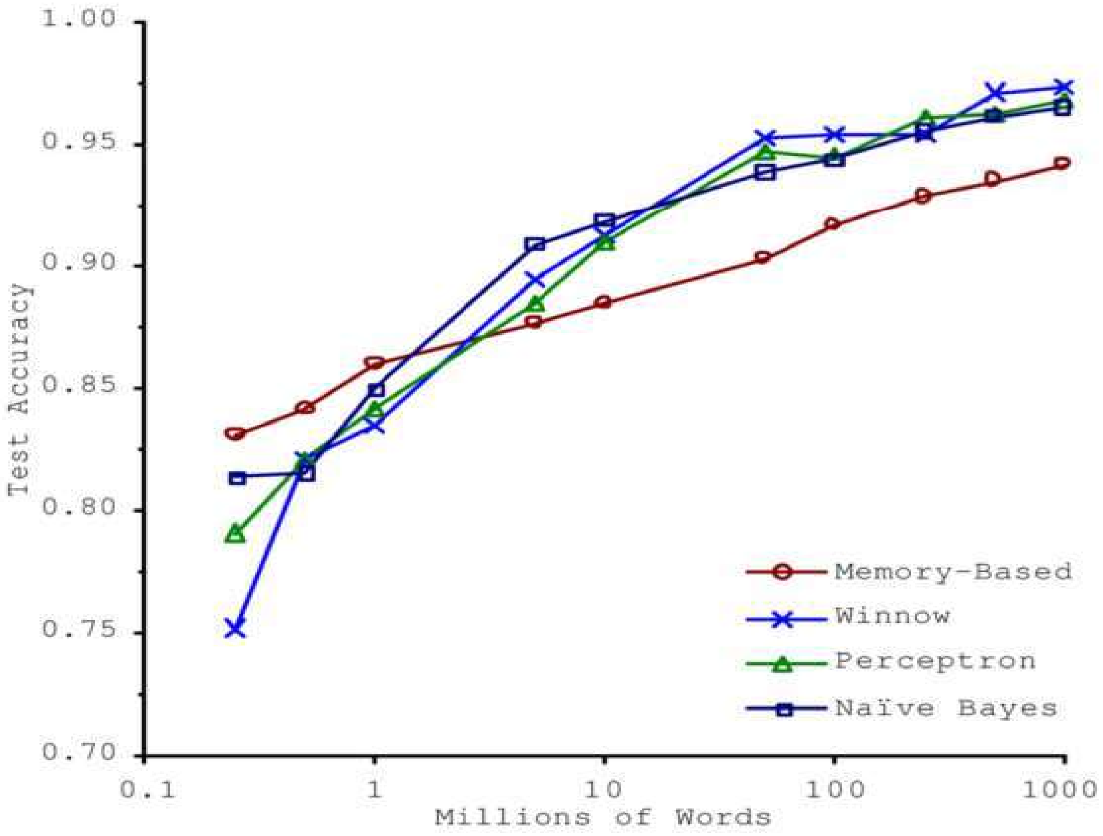



R Boxplot Example
Miles Per Gallon
10
15
20
25
30
4
6
8
Quantity of Cylinders

- **Bubble Charts-** Bubble charts are a variation of scatter charts in which the data points are replaced with bubbles. Also, an extra proportion of data is portrayed in the size of the bubbles.
- **Treemaps -** indicates hierarchical data in a nested format
- **Correlation chart** –
- **Dendrograms** show the hierarchical connection between the objects.

Main Challenges of Machine Learning

- Insufficient Quantity of Training Data



Main Challenges of Machine Learning

- Nonrepresentative Training Data
  - In order to generalize well, it is crucial that your training data be representative of the new cases you want to generalize to.
- Poor-Quality Data
  - if your training data is full of errors, outliers, and noise (e.g., due to poor quality measurements), it will make it harder for the system to detect the underlying patterns, so your system is less likely to perform well.
- Irrelevant Features
  - Your system will only be capable of learning if the training data contains enough relevant features and not too many irrelevant ones.
  - A critical part of the success of a Machine Learning project is coming up with a good set of features to train on. This process, called feature engineering

Main Challenges of Machine Learning

- Overfitting the Training Data
  - **Overfitting** means that the model performs well on the training data, but it does not generalize well.
  - Overfitting happens when the model is too complex relative to the amount and noisiness of the training data.
  - Here are possible solutions:
    - Simplify the model by selecting one with fewer parameters (e.g., a linear model rather than a high-degree polynomial model), by reducing the number of attributes in the training data, or by constraining the model.
  - Gather more training data.
  - Reduce the noise in the training data (e.g., fix data errors and remove outliers).

Main Challenges of Machine Learning

- Underfitting the Training Data
  - underfitting is the opposite of overfitting: it occurs when your model is too simple to learn the underlying structure of the data.
  - Here are the main options for fixing this problem:
  - Select a more powerful model, with more parameters.
  - Feed better features to the learning algorithm (feature engineering).
  - Reduce the constraints on the model (e.g., reduce the regularization hyperparameter).

Referance

- https://machinelearningmastery.com/hyperparameters-for-classification-machine-learning-algorithms/
- https://www.jeremyjordan.me/hyperparameter-tuning/