why not change the world?®

Introduction to Analytic Methods, Types of Data Mining for Analytics & Introduction to Group 2

Ahmed Eleish

Data Analytics ITWS-4600/ITWS-6600/MATP-4450/CSCI-4960 Group 1, Module 4, September 16th, 2025

Tetherless World Constellation
Rensselaer Polytechnic Institute

Contents

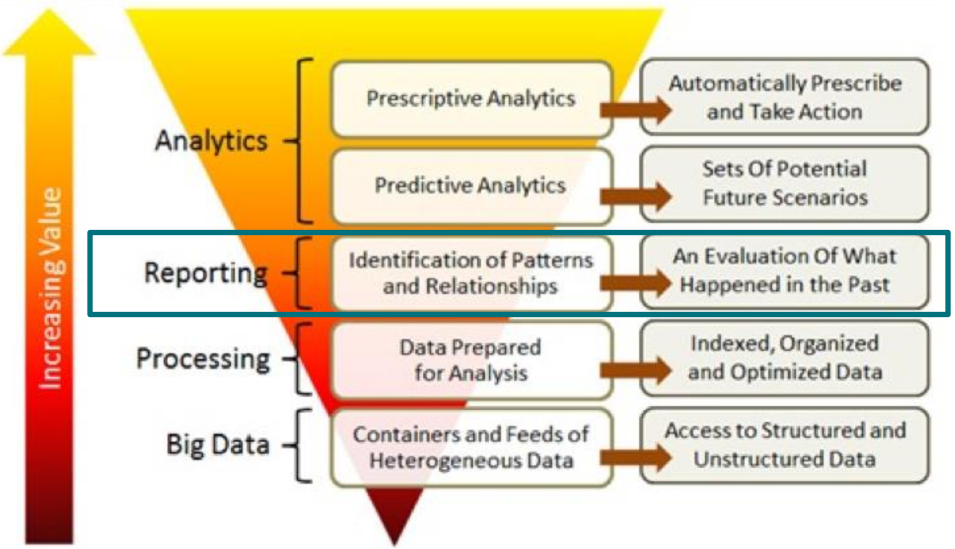
- Reminder: preliminary/exploratory data analysis, models

- Patterns/ Relations via “Data mining”

- Interpreting results

- Saving the models

- Proceeding with applying the models



Preliminary Data Analysis

- Relates to the sample v. population

- Also called **Exploratory Data Analysis**

– "EDA is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those we believe will be there" (John Tukey)

- Distribution analysis and comparison, visual 'analysis', model testing, i.e. pretty much the things you did last lab and will do more of!

Models

- Assumptions are often used when considering models, e.g. as being representative of the *population* – since they are so often derived from a *sample* – this should be starting to make sense (a bit)

- Two key topics:

– N = *all* data points .. and the open world assumption

– Model of the thing of interest *versus* model of the data (data model; structural form)

- “All models are wrong but some are useful” (*generally attributed to the statistician ~ George Box*)

Art or science?

• The form of the model, incorporating the hypothesis determines a “form”.

• Thus, as much art as science because it depends both on your world view and what the data is (are?) telling you (or not).

• We will however, be giving the models nice mathematical properties.

Patterns and Relationships

- Stepping from elementary/distribution analysis to algorithmic-based analysis i.e. pattern detection via data mining: classification, clustering, rules; machine learning; support vector machines, non-parametric models

- Relations: associations between/among variables/datasets

- Outcome: a model and an evaluation of its fitness for purpose

Data Mining = Patterns

- Classification (Supervised Learning)

– Classifiers are created using labeled training samples – Training samples created by ground truth / experts

– Classifier later used to classify unknown samples

- Clustering (Unsupervised Learning)
  - Grouping objects into clusters so that similar objects are in the same cluster and dissimilar objects are in different clusters
  - Discover overall distribution patterns and relationships between attributes

- Association Rule Mining

– Initially developed for market basket analysis

– Goal is to discover relationships between attributes

- Other Types of Mining

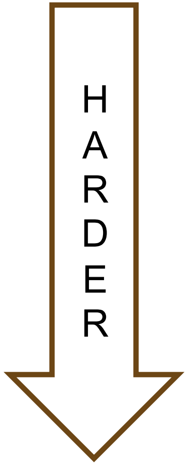
– Outlier Analysis

– Concept / Class Description

– Time Series Analysis

Models/ types

- Trade-off between Accuracy and Understandability
- Models range from “easy to understand” to incomprehensible
  - Decision trees
  - Rule induction
  - Multi-variate Regression models
  - Neural Networks
  - Deep Learning



Patterns and Relationships

- Linear and multi-variate models

- Linear Regression

- Supervised Learning (Classification)

– k-Nearest Neighbors

- Unsupervised Learning (Clustering)

– K-Means

Regression

Regression in Statistics

• Regression is a statistical process for *estimating* the relationships among continuous numerical variables.

• In regression analysis the focus is on the relationship between a dependent variable and one or more independent variables.

• Independent variables are also called predictors, covariates.. these are inputs.

• Dependent variables are also called response variables, these are outputs.

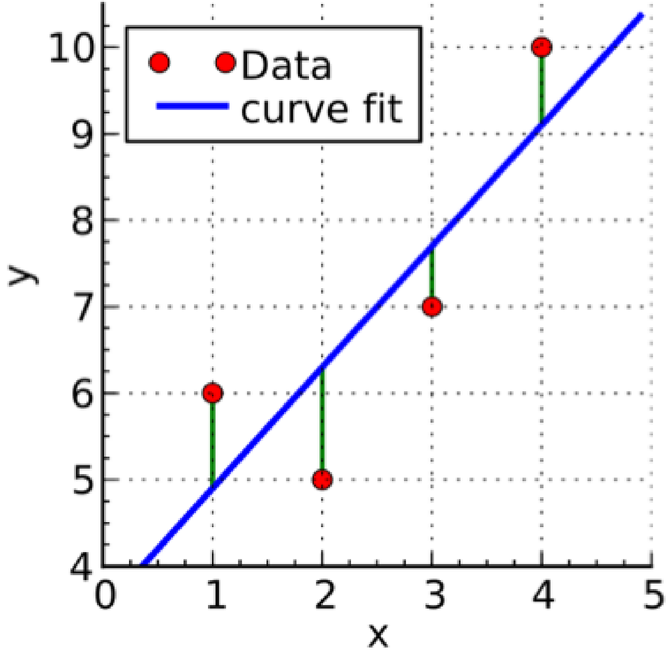
• Estimation is often done by constraining an objective function.

• Must be tested for significance, confidence.

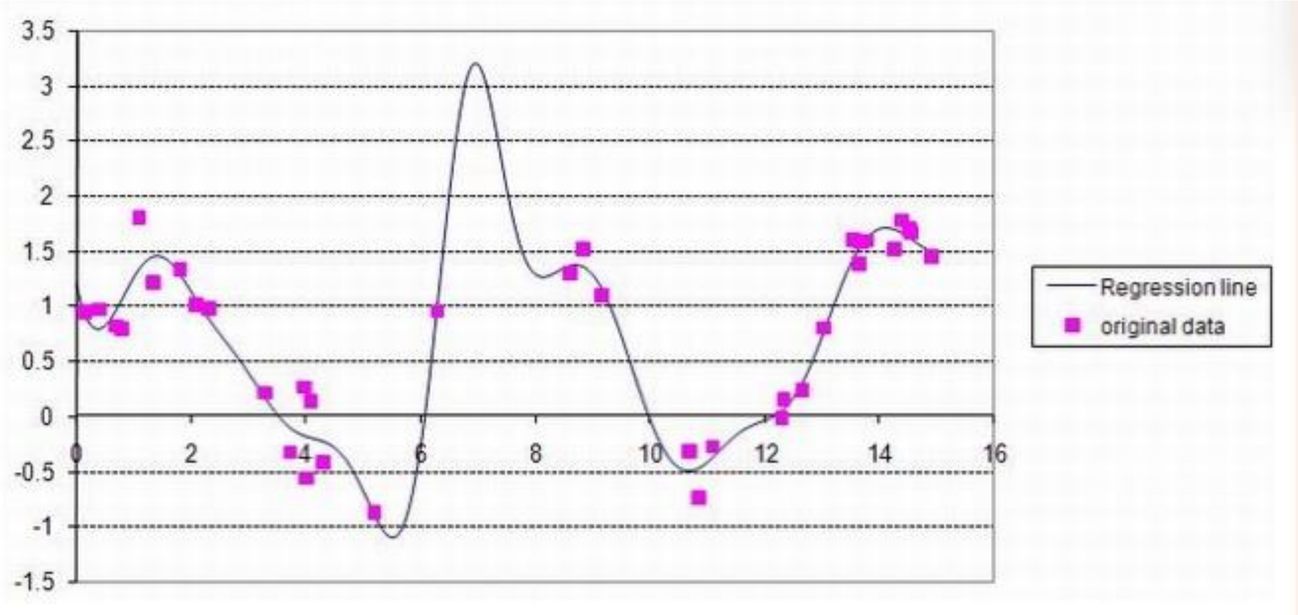
Objective function

Constraint function(s)

Regression



Regression - when it gets complex…



Supervised Learning

k-nearest neighbors (knn)

- Can be used in both regression and classification ("non-parametric")
  – Supervised learning, i.e. model is trained

- kNN is a method for classifying objects based on the closest training examples in the feature space.

- **An object is classified by a majority vote of its neighbors. k is always a positive integer**. The neighbors are taken from a set of objects for which the correct classification is known.

- It is usual to use the Euclidean distance, though other distance measures such as the Manhattan distance could in principle be used instead.

Algorithm

- The algorithm on how to compute the k-nearest neighbors is as follows:

– Determine parameter *k* = number of nearest neighbors, beforehand. This value **is up to you**.

– Calculate the distance between the query-instance and all the training samples. You can use **any distance** algorithm.

– Sort the distances for all the training samples and determine the nearest neighbors based on the *k* shortest distances.

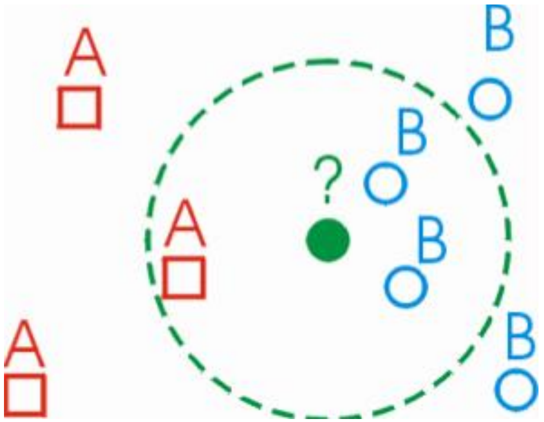
– Since this is supervised learning, get the classes for the *k* nearest neighbors from the training set.

– Use the majority of nearest neighbors as the prediction value.

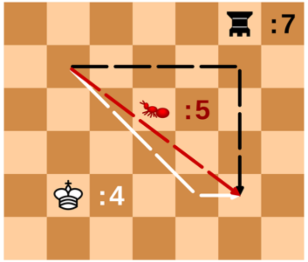
Choice of k?

- Don’t you hate it when the instructions read: the choice of ‘k’ is all up to you ??

- Loop over different k, evaluate results...



Distance metrics

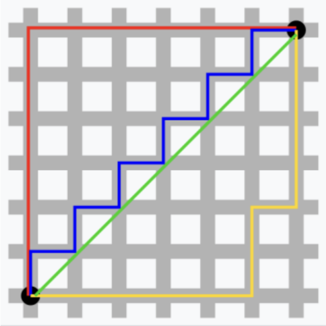


- **Euclidean** distance is the most common use of distance. Euclidean distance, or simply 'distance', examines the root of the sum of square differences between the coordinates of a pair of objects. This is most generally known as the Pythagorean theorem.

Image credit: Cmglee
license: CC BY-SA 4.0 – no changes

- The **taxicab** metric is also known as **rectilinear** distance, L1 distance or L1 norm, city block distance, **Manhattan** distance, or Manhattan length, with the corresponding variations in the name of the geometry. It represents the distance between points in a city road grid. It examines the absolute differences between the coordinates of a pair of objects.

https://en.wikipedia.org/wiki/Taxicab_geometry



More generally

- **Chebyshev** distance is also called the Maximum value distance, defined on a vector space where the distance between two vectors is the greatest of their differences along any coordinate dimension. In other words, it examines the absolute magnitude of the differences between the coordinates of a pair of objects.

- The general metric for distance is the **Minkowski** distance. When *p* is equal to 1, it becomes the city block distance, and when *p* is equal to 2, it becomes the Euclidean distance. The special case is when *p* is equal to infinity (taking a limit), where it is considered as the Chebyshev distance.

https://en.wikipedia.org/wiki/Chebyshev_distance

$$D(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}$$

What does “Near” mean??

- More on this in the next topic but ...
  - DISTANCE – and what does that mean
  - RANGE – acceptable, expected?
  - SHAPE – i.e. the form

Training and Testing

- We are going to do much more on this going forward...

- Supervision means not all the data are used to train because you want to test on the untrained set (before you predict for new values)

– What is the ‘sampling’ strategy for training?
– Is the sample representative of the population?
– What is the optimum split of the dataset for training/testing?

Summing up ‘knn’

- **Advantages**

– Robust to noisy training data (especially if we use inverse square of weighted distance as the “distance”)
– Effective if the training data is large

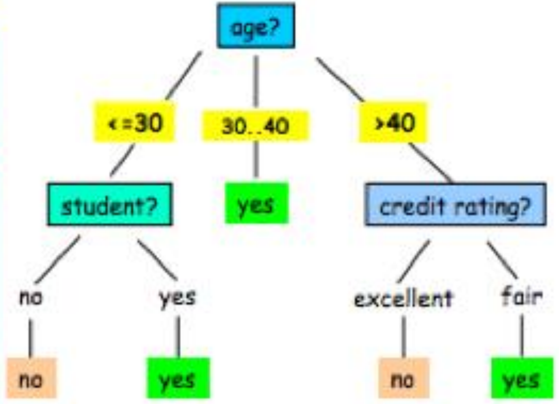
- **Disadvantages**

– Need to determine value of parameter k (number of nearest neighbors)
– Distance based learning is not clear on which type of distance to use and which attribute to use to produce the best results. Shall we use all attributes or certain attributes only?
– Computation cost is quite high because we need to compute distance of each query instance to all training samples.

Decision tree classifier

Classification by Decision Tree Induction

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no



buys_computer ?

More on this later in Group 2 ...

Unsupervised Learning

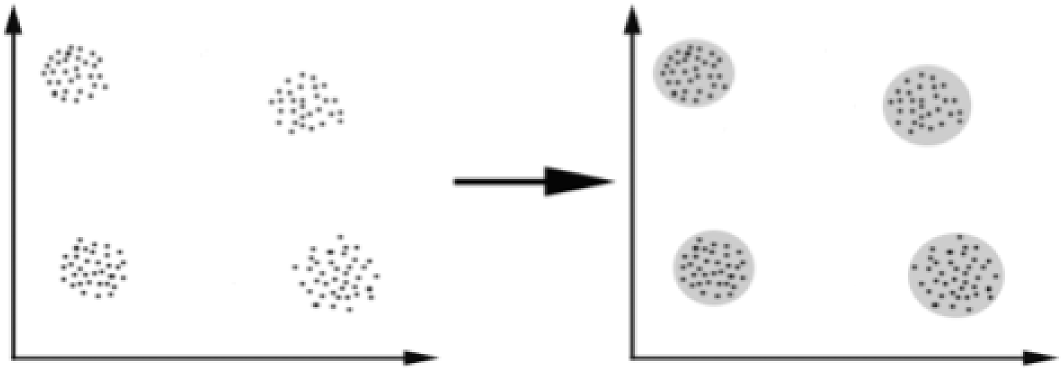
K-means

- Unsupervised Learning - Clustering, i.e. no class labels known beforehand

- Types:
  - Hierarchical: Successively determine new clusters from previously determined clusters (parent/child clusters).
  - Partitional: Establish all clusters at once, at the same level.

Similarity/Distance Measure

- Clustering is about finding “similarity”.
- To find how similar two objects are, one needs a “distance” measure.
- Similar objects (same cluster) should be close to one another (short distance).



Distance Measure

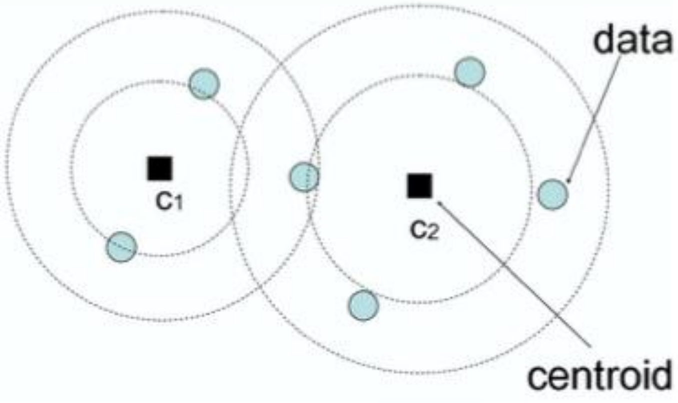
- Many ways to define distance measure.

- Some elements may be close according to one distance measure and further away according to another.

- Select a good distance measure is an important step in clustering.

k-Means Clustering

- Separate the objects (datapoints) into k clusters.
- Cluster center (centroid) = the mean (average) of all the data points in the cluster.
- Assigns each data point to the cluster whose centroid is nearest (using distance function).



k-Means Algorithm

1. Place k points into the space of the objects being clustered. They represent the initial group centroids.

2. Assign each object to the group that has the closest centroid.

3. Recalculate the positions of the k centroids.

4. Repeat Steps 2 & 3 until the group centroids no longer move.

k-Means Algorithm Example

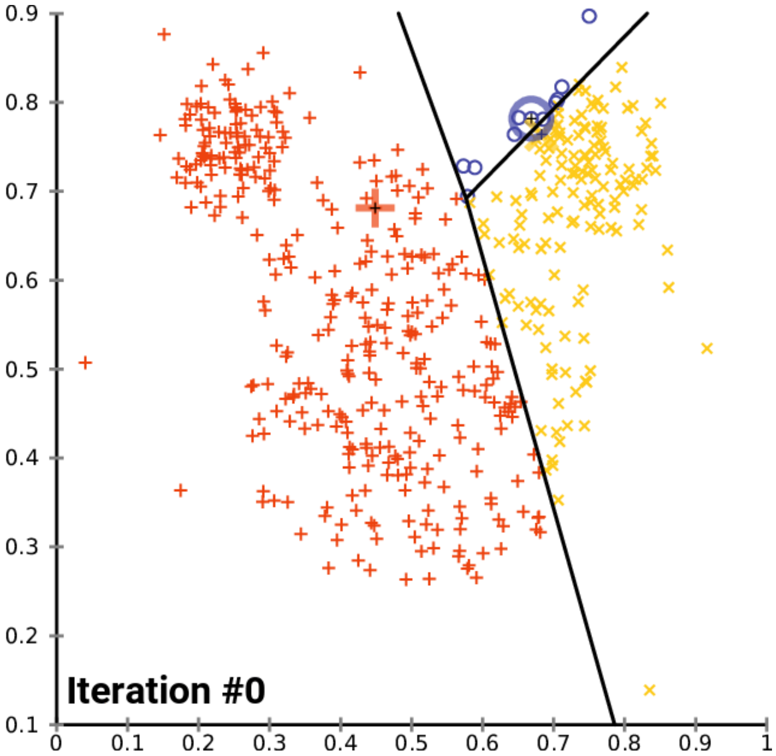
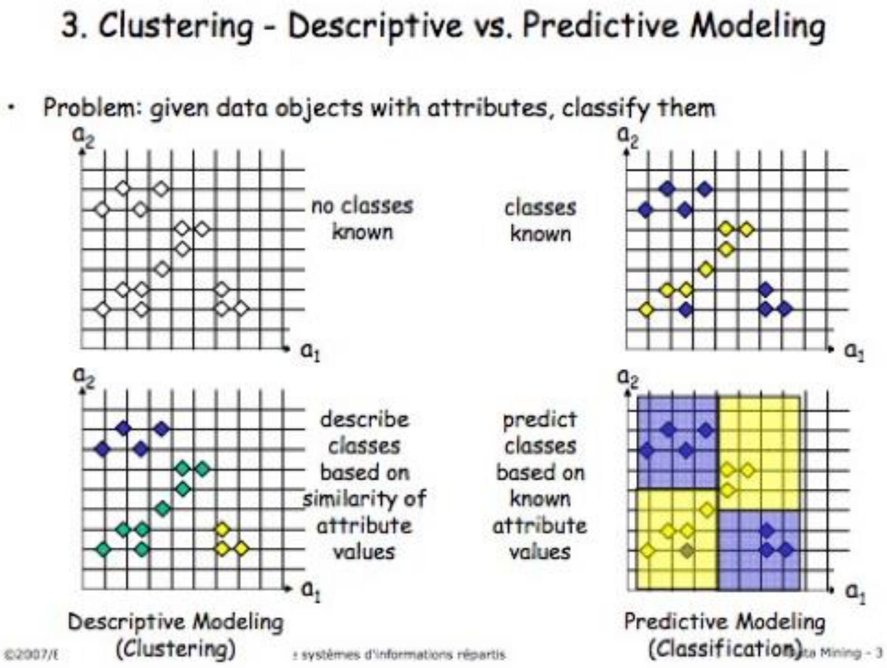


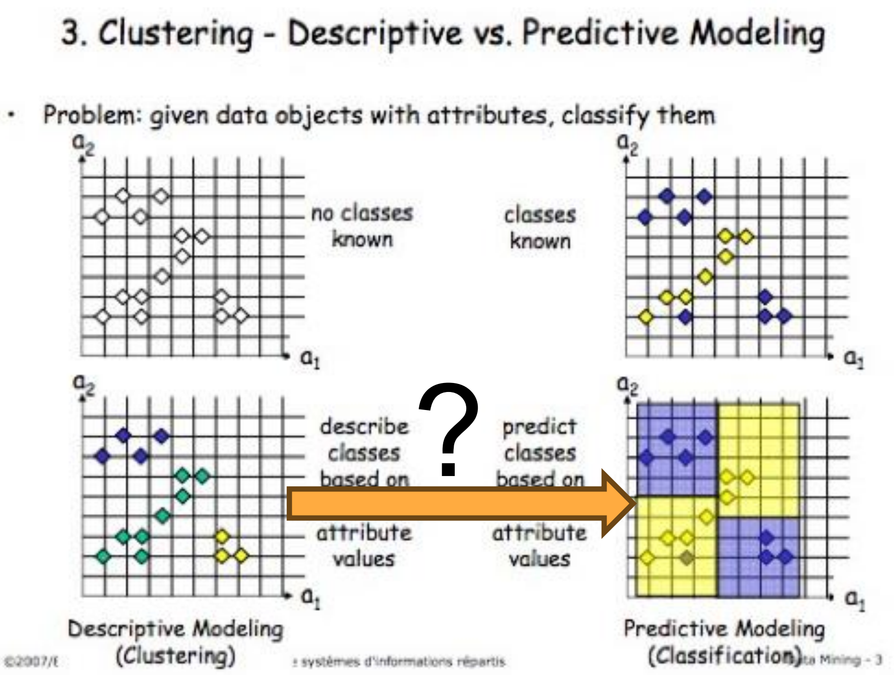
Image credit: Chire license: CC BY-SA 4.0 – no changes

https://en.wikipedia.org/wiki/K-means_clustering

Describe v. Predict



Describe v. Predict



More on this later in Group 2 ...

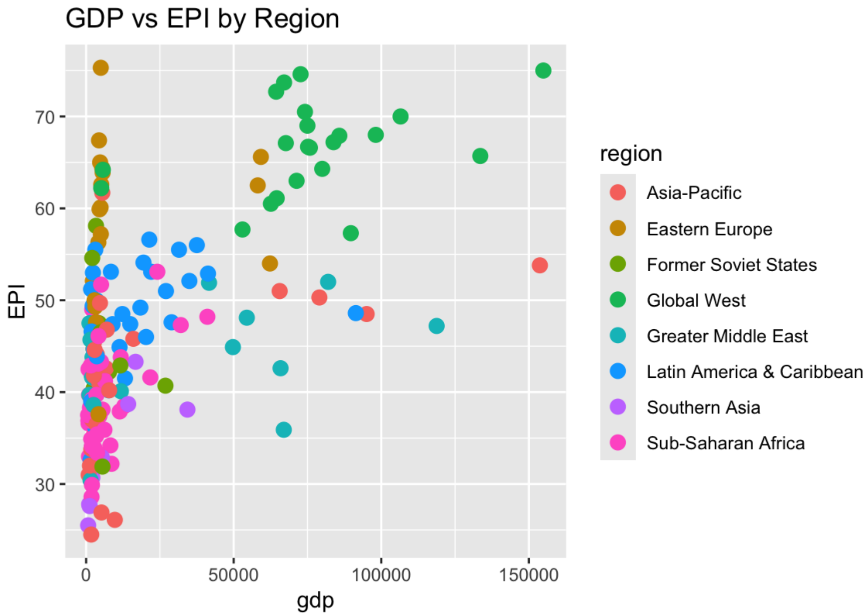
Visualizations

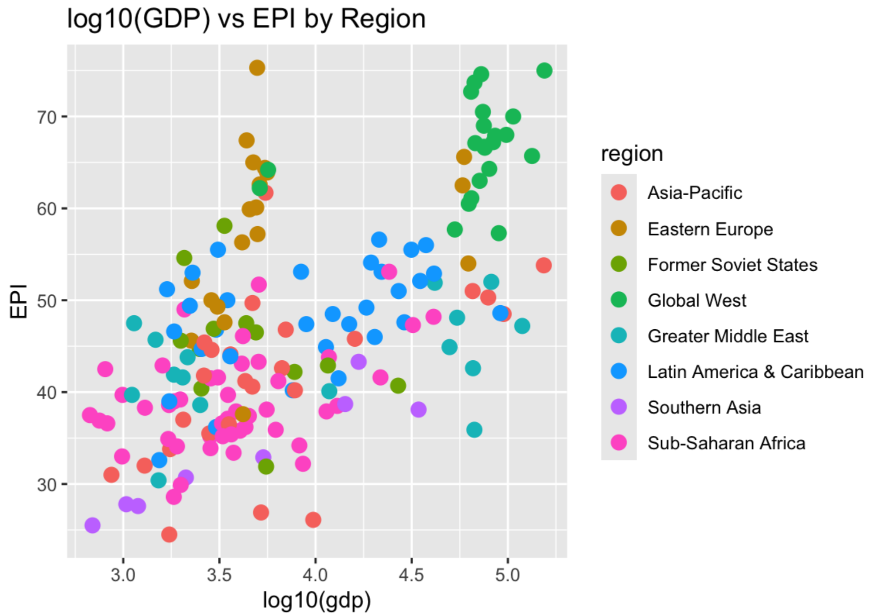
- Scatter Plot – Paired data (x,y)

- Describe the relationship between numerical variables.
- Make a note on the direction of the data points
  - Positive direction
  - Negative Direction
- Check for unusual observations
- See the relationship - Linear or Non-linear

Visualizations

- Scatter Plot – Paired data (x,y)





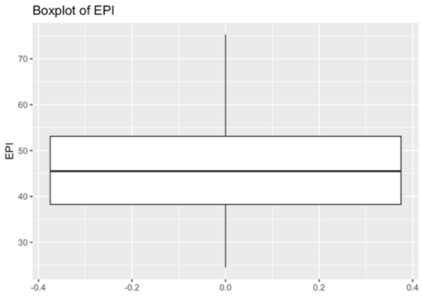
Visualizations

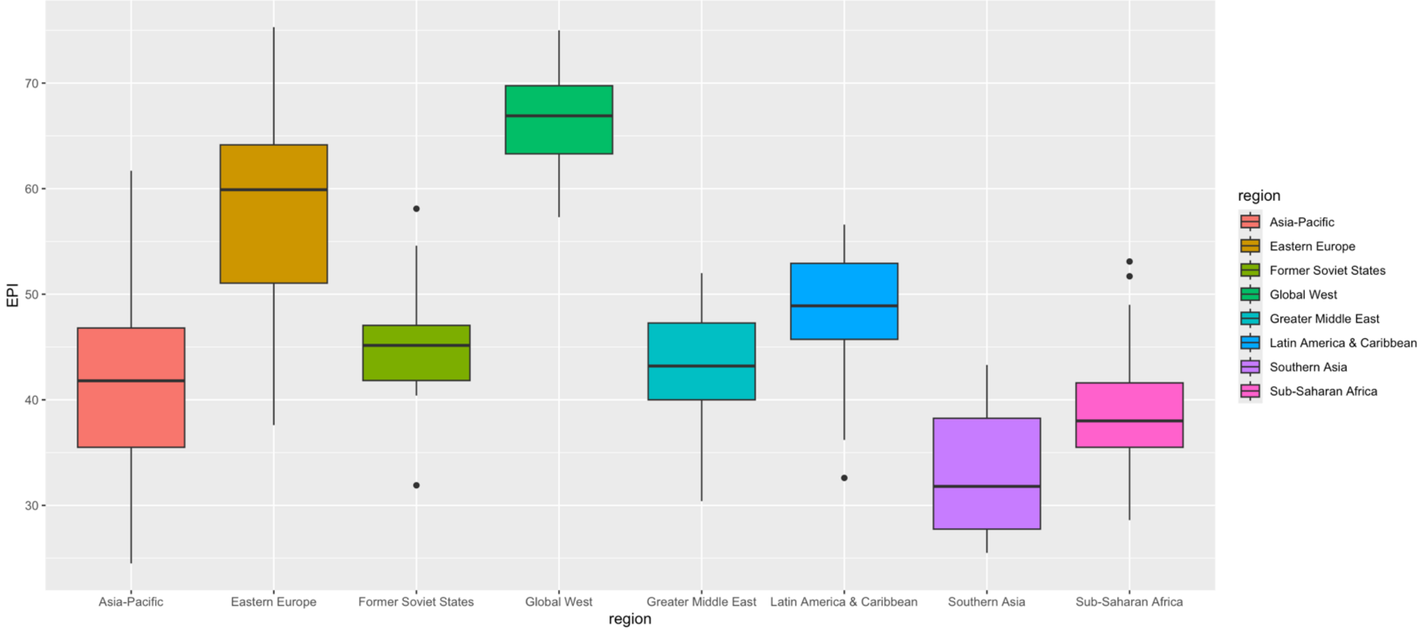
- Boxplot - box-and-whisker of a variable x

- Gives an overview of the range of the variable and where most of the observations are exist along that range
- Make a note of the minimum, maximum, quartiles
- Check for unusual observations (outliers)

Visualizations

- Boxplot - box-and-whisker of a variable x





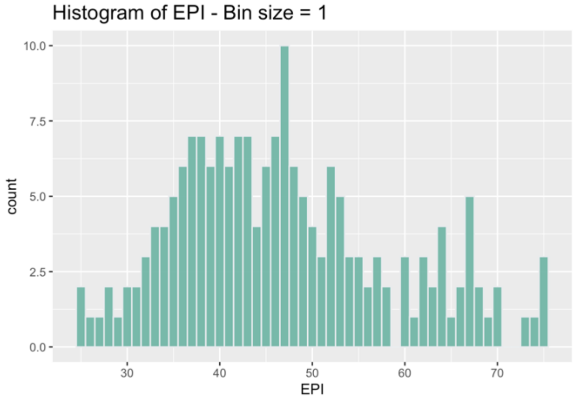
Visualizations

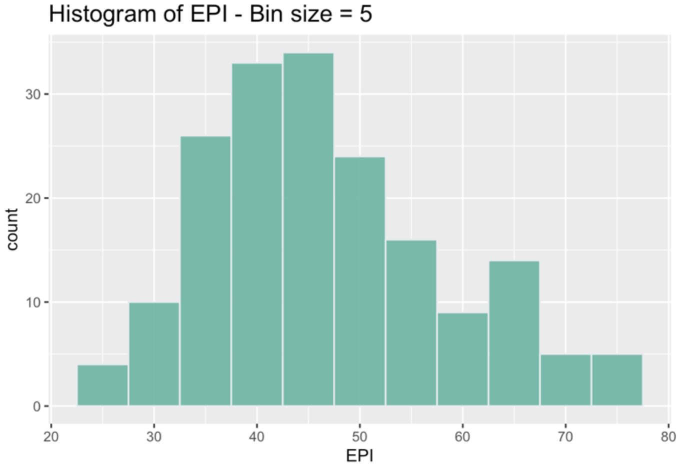
- Histogram – Distribution of variable X

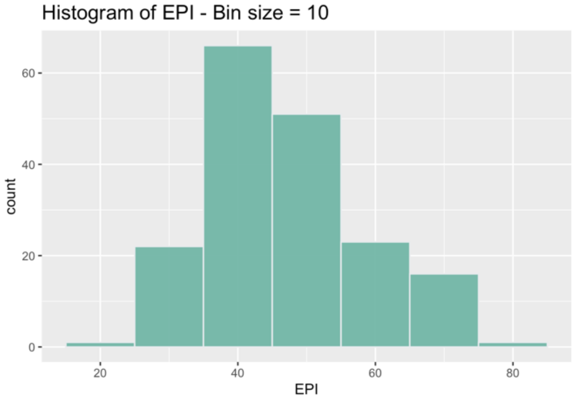
- Describe frequency of occurrence of values or ranges of values (bins) of x
- Tune bin size parameter
- Observe overall shape
- Check for mixed distributions

Visualizations

- Histogram – Distribution of variable X







We'll do more during up coming lectures/labs..

- We will move to Group 2: Patterns, relations, descriptive analytics..

Possible Project Ideas for the Data Analytics Course

- Sustainable Development Goals (SDG) using UN Data
https://sdgs.un.org/goals

- Watch:
https://www.youtube.com/watch?time_continue=4&v=0XTBYMfZyrM&feature=emb_logo&ab_channel=UnitedNations
https://data.un.org/
https://www.un.org/en/global-issues/big-data-for-sustainable-development
Watch:
https://www.youtube.com/watch?v=yobFJniliOs&feature=emb_logo&ab_channel=WesternDigitalCorporation
- Watch:
https://www.youtube.com/watch?v=v-zGHqMyd7o&feature=emb_logo&ab_channel=UNGlobalPulse

Dataset search

- If you do not have a dataset in mind for your project, please search online and select datasets using search tools such as
 https://datasetsearch.research.google.com/

- If you need help choosing a dataset, please come and talk to me during the class time or during virtual office hours, so that I can guide/help you to select datasets.

- **NOTE:6000-Level students MUST have TWO datasets (minimum two datasets) used during final project.**

More places to find data:

- US Government Data: https://www.data.gov/
- US Department of Agriculture: https://www.nass.usda.gov/Data_and_Statistics/index.php
- Center of Disease Control (CDC): https://www.cdc.gov/datastatistics/index.html
- US Financial Data: https://www.federalreserve.gov/data.htm
- European Union Open Data Portal: https://data.europa.eu/euodp/en/data/
- Nasa: https://data.nasa.gov/

Preview (optional): Chapter 3 & 5

- Chapter 3 (Linear Regression), Introduction to Statistical Learning with Applications in R, 7th Edition

- Chapter 5 (Decision Trees and Clustering), Introduction to Statistical Learning with Applications in R, 7th Edition

https://rpi.box.com/s/zr41tgyri5iybpes3233lhx0v9nvnbji

Next Class: Friday September 19th

Lab 2

Thanks!