Knowledge Graphs, Data Analysis I: Concepts, Statistics, Visualizations, Regression

Ahmed Eleish

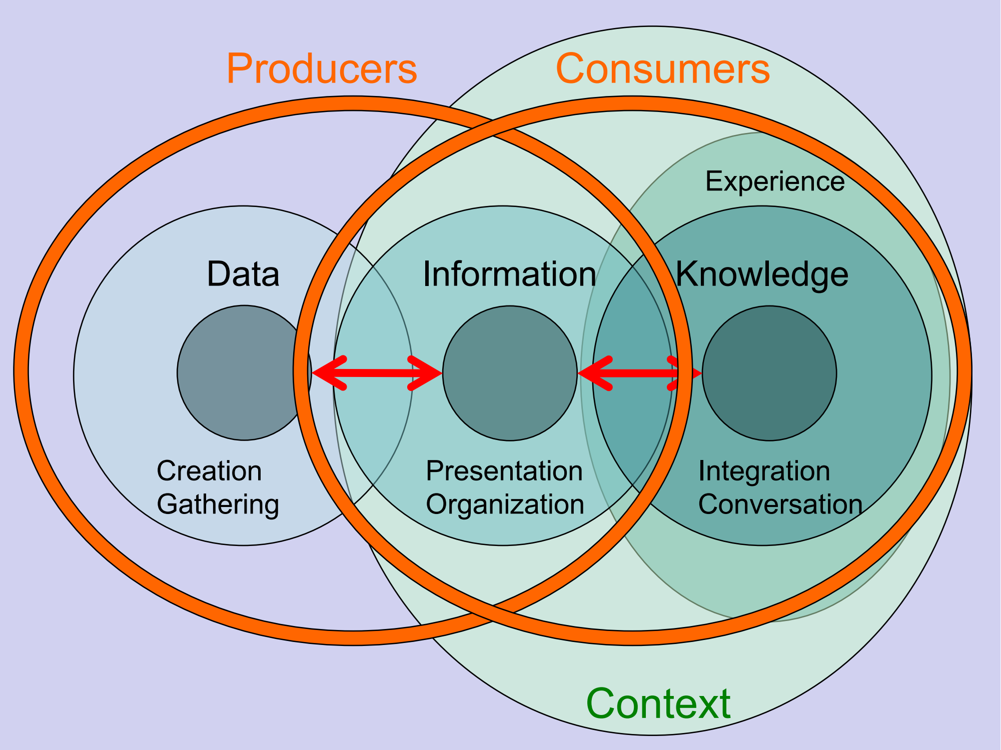
Data Science – ITWS/CSCI/ERTH 4350/6350 Module 4, September 24, 2025

Tetherless World Constellation
Rensselaer Polytechnic Institute

Contents

- Reading review
- Module 2 & 3 review
- Knowledge Graphs
- Data Analysis Concepts, Exploratory Analysis
- Visualizations, Distributions, Statistics, Regression
- Assignment 2

Data-Information-Knowledge Ecosystem



Modes of collecting data, information

- Observation
- Measurement
- Generation

- Driven by
  - Questions
  - Research idea
  - Exploration

Data Management

- Creation of logical collections
- Physical data handling
- Interoperability support
- Security support
- Data ownership
- Metadata collection, management and access
- Persistence
- Knowledge and information discovery
- Data distribution and publication

Data Management

- Creation of logical collections

e.g. a catalogue mapping data objects to files/database views with standardized naming conventions.

- Physical data handling

e.g. storing the data in relational database tables with foreign key constraints linking the tables, backups monthly placed in a “backup” directory.

- Interoperability support

e.g. backups exported as CSVs, metadata follow naming convention from a controlled vocabulary.

Data Management

- Security support

e.g. data stored on a network shared storage device with access and permissions managed through Microsoft Active Directory.

- Data ownership

e.g. data products are assigned to specific team member for quality control and metadata injection.

- Metadata collection, management and access

e.g. metadata associated with lab experiments are recorded in plain text file using the LINCS Phase II Extended Metadata Standards.

Data Management

- Persistence

e.g. Data package will be submitted to RPI Institute Archives and Special Collections for archival, archival package must conform to the OAIS standard.

- Knowledge and information discovery

e.g. An interactive visualization allows researchers to examine a portion of the data before downloading the entire dataset OR statistics computed from the dataset are published along with the data files.

- Data distribution and publication

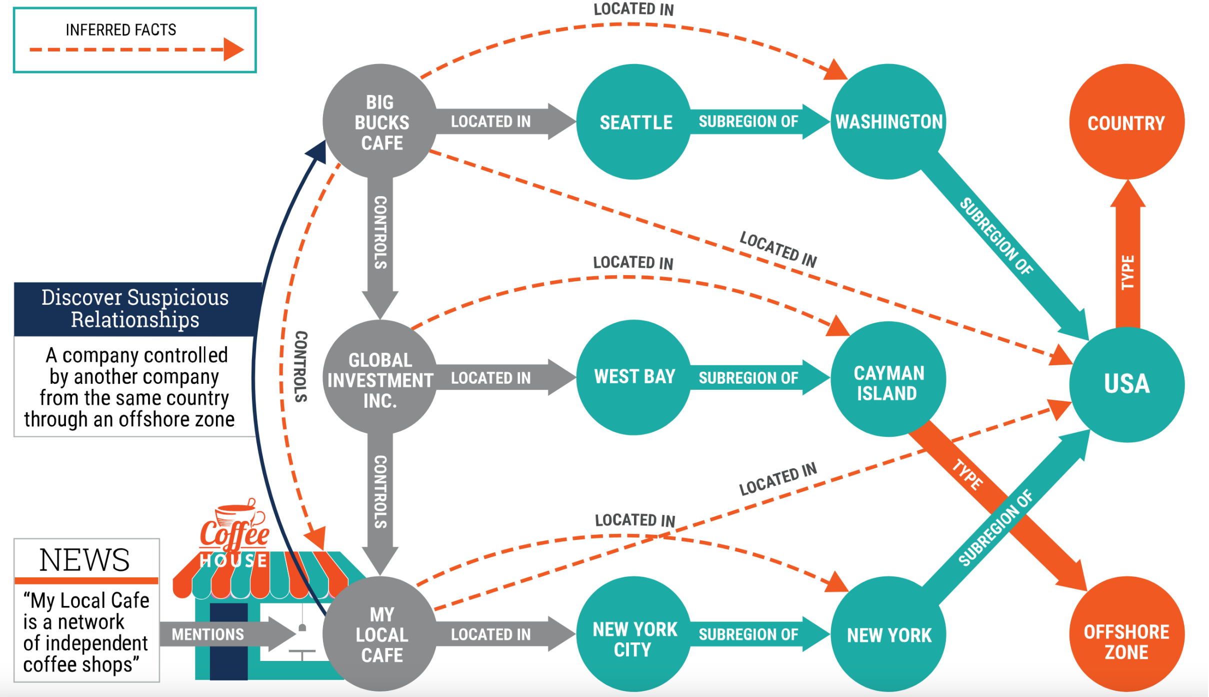
e.g. Dataset is stored on a content management system with version control that sends notifications to users when data changes.

Provenance

- Who?
- What?
- Where?
- Why?
- When?
- How?

Knowledge Graphs

What is a Knowledge Graph?



https://www.ontotext.com/knowledgehub/fundamentals/what-is-a-knowledge-graph/

What is a Knowledge Graph?

The knowledge graph (KG) represents a collection of interlinked descriptions of entities – real-world objects, events, situations or abstract concepts – where:

- Descriptions have a formal structure that allows both people and computers to process them in an efficient and unambiguous manner;

- Entity descriptions contribute to one another, forming a network, where each entity represents part of the description of the entities, related to it.

Where are Knowledge Graphs used?

- Web Search (e.g. Google)
- Data Integrations
- Answering Questions ~ before LLMs :'(

Google Knowledge Graph

- Introducing Knowledge graphs:
- Watch: https://www.youtube.com/watch?v=mmQl6VGvX-c&t=17s&ab_channel=Google
- Read: https://blog.google/products/search/introducing-knowledge-graph-things-not/

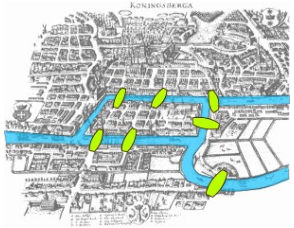
Knowledge Graphs (KGs)

Knowledge Graphs are a type of **graph**.

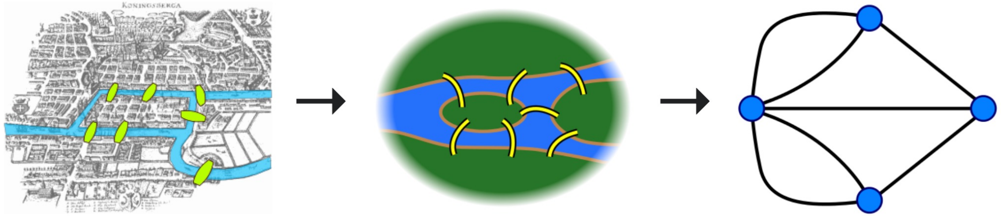
- What are graphs?
- Graphs are simple structures consisting of **nodes (or vertices)** connected by **links (or edges).**
- Mathematically, a graph is a set of vertices *V* and a set of edges *E* where each edge in *E* connects 2 vertices from *V*.
- The nodes represent entities (things) and the links represent relations/connections between the entities.

Graphs

- Graphs are sometimes referred to as networks.
- They are a simple but very powerful way of describing how things are connected.
- Graph Theory was introduced by the Swiss mathematician Leonhard Euler during the 18th century with his famous problem known as the Seven Bridges of Königsberg.



Euler used this graph to answer the question: Can you walk through the city and cross each bridge only once?



Images/Resource Credits: https://en.wikipedia.org/wiki/Seven_Bridges_of_K%C3%B6nigsberg

Knowledge Graphs (KGs)

- KGs are graphs where the nodes represent entities from a particular domain of interest, and the links represent qualified relations between these entities.

- KGs can be explored with structured queries.

- KGs can be analyzed as networks of data.

- Formal semantics can be defined for KGs by using ontologies.

https://www.ontotext.com/knowledgehub/fundamentals/what-is-a-knowledge-graph/

Ontologies (Information Science)

Ontologies define the structure/organization of knowledge in a domain.

- “An ontology is a formal description of knowledge as a set of concepts within a domain and the relationships that hold between them.”

- “an ontology encompasses a representation, formal naming, and definitions of the categories, properties, and relations between the concepts, data, or entities that pertain to one, many, or all domains of discourse”

- An ontology can be considered the data model or schema of a knowledge graph, it represents the semantics (meanings) behind the data.

- Ontology are similar to, but more comprehensive than taxonomies and controlled vocabularies.

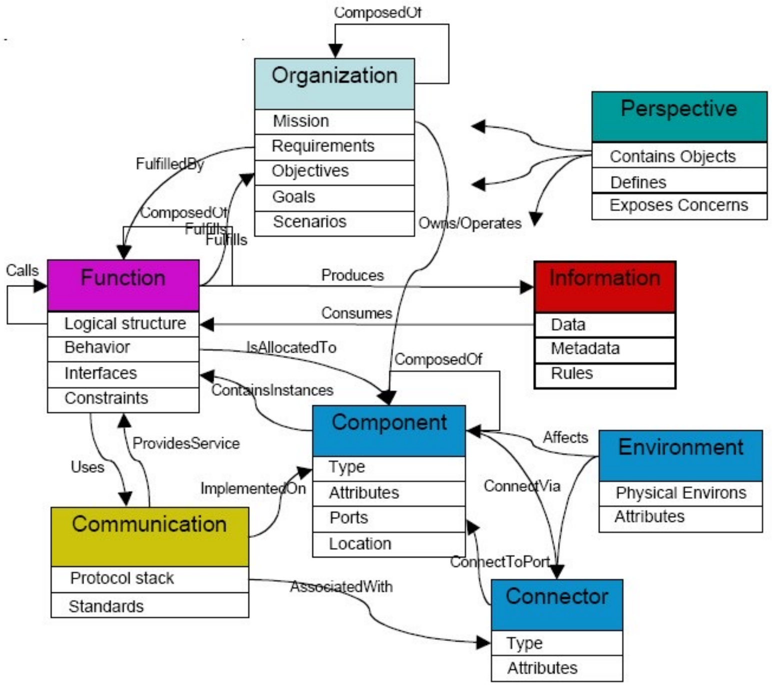
Ontologies

- An ontology contains at the least classes, properties, and relations.

- Ontologies may optionally contain restrictions, rules, and axioms which enable a person or machine (program) to draw inferences from a knowledge graph.

e.g.

- A person's age must be a positive number
- Siblings must share one or more parents
- Taking the Data Science course is necessary for success



Example Ontology - FOAF

- Friend Of A Friend (FOAF) is a “a machine-readable ontology describing persons, their activities and their relations to other people and objects.”

- “FOAF is a descriptive vocabulary expressed using the Resource Description Framework (RDF) and the Web Ontology Language (OWL).”

- Contains classes: Agent, Person, Group, Organization, Project, OnlineChatAccount, …

e.g.

```
<foaf:Person rdf:about="#danbri" xmlns:foaf="http://xmlns.com/foaf/0.1/">
  <foaf:name>Dan Brickley</foaf:name>
  <foaf:homepage rdf:resource="http://danbri.org/" />
  <foaf:openid rdf:resource="http://danbri.org/" />
  <foaf:img rdf:resource="/images/me.jpg" />
</foaf:Person>
```

FOAF Specification http://xmlns.com/foaf/spec/

The Semantic Web

- The Semantic Web is a web of data (not (just) documents)
- “It is about common formats for integration and combination of data drawn from diverse sources..”
- “It is also about language for recording how the data relates to real world objects.”
- “The goal of the Semantic Web is to make Internet data machine-readable”
- How do we encode data on the web? → RDF!

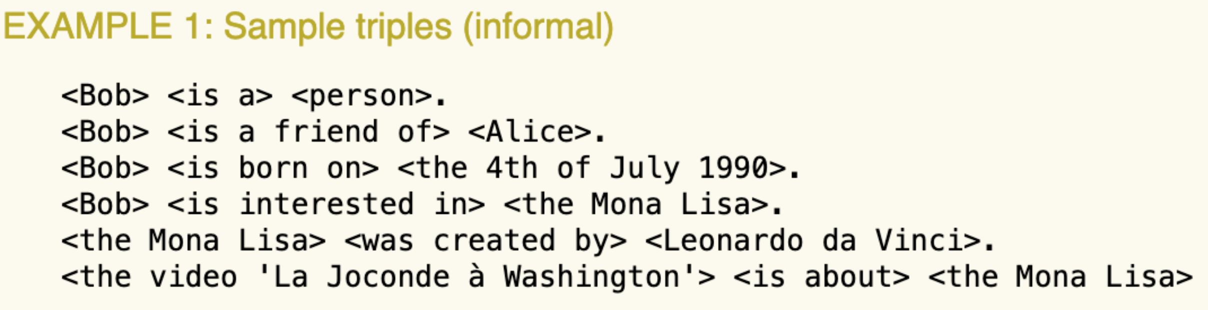
Semantic Web: https://www.w3.org/2001/sw/ Web of Data: https://www.w3.org/2013/data/

Resource Description Framework (RDF)

- A framework for expressing information about resources. Resources can be anything, including documents, people, physical objects, and abstract concepts.

- RDF encode data in the form of triples:

<subject> <predicate> <object>



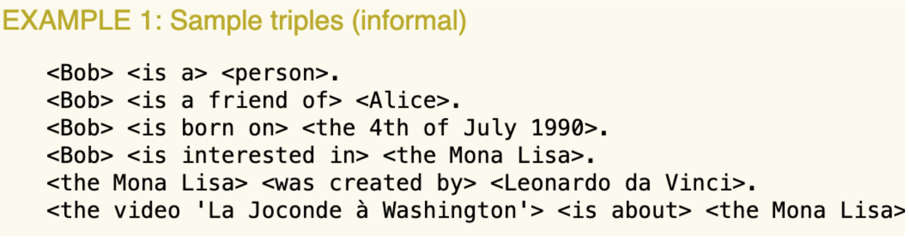
EXAMPLE 1: Sample triples (informal)

```
<Bob> <is a> <person>.
<Bob> <is a friend of> <Alice>.
<Bob> <is born on> <the 4th of July 1990>.
<Bob> <is interested in> <the Mona Lisa>.
<the Mona Lisa> <was created by> <Leonardo da Vinci>.
<the video 'La Joconde à Washington'> <is about> <the Mona Lisa>
```

RDF Primer: https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/

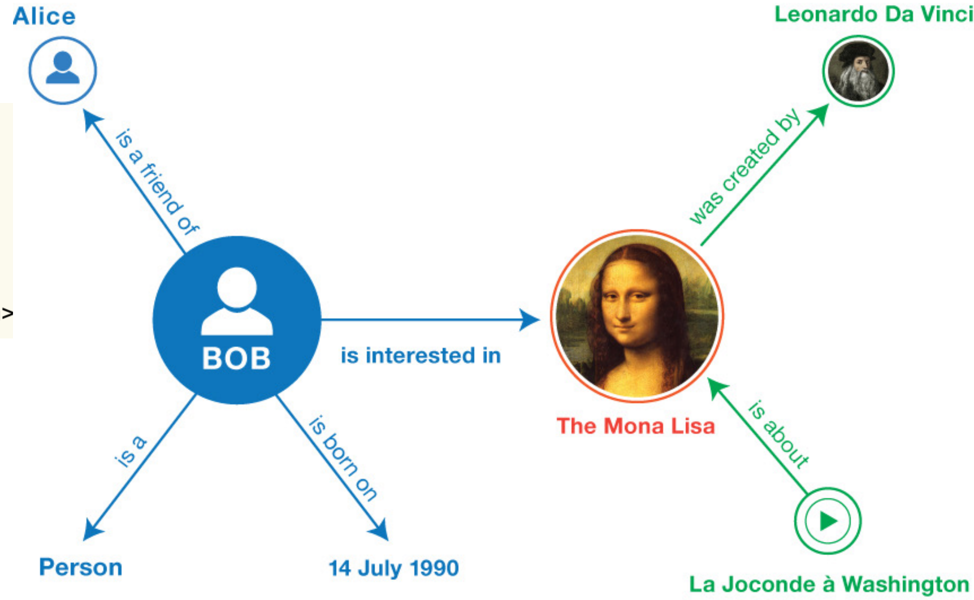
Resource Description Framework (RDF)

- RDF encode data in the form of triples



EXAMPLE 1: Sample triples (informal)

```
<Bob> <is a> <person>.
<Bob> <is a friend of> <Alice>.
<Bob> <is born on> <the 4th of July 1990>.
<Bob> <is interested in> <the Mona Lisa>.
<the Mona Lisa> <was created by> <Leonardo da Vinci>.
<the video 'La Joconde à Washington'> <is about> <the Mona Lisa>
```



RDF Primer: https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/

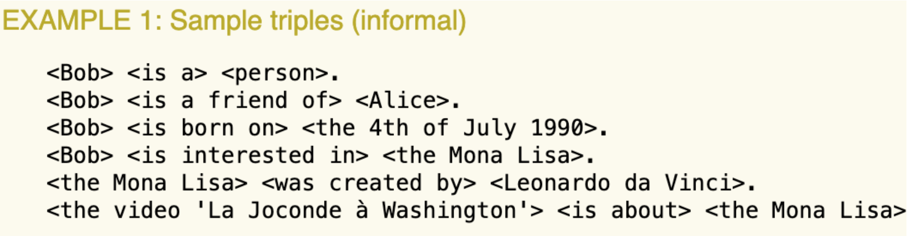
Resource Description Framework (RDF)

- In RDF everything is a resource.
- Every resource has Uniform Resource Identifier (URI).
- Classes and properties must have namespaces (e.g. FOAF).
- RDF can be serialized (written in a file) in several ways:
  - N-Triples, Turtle, JSON-LD, RDFa, RDF/XML

RDF Primer: https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/

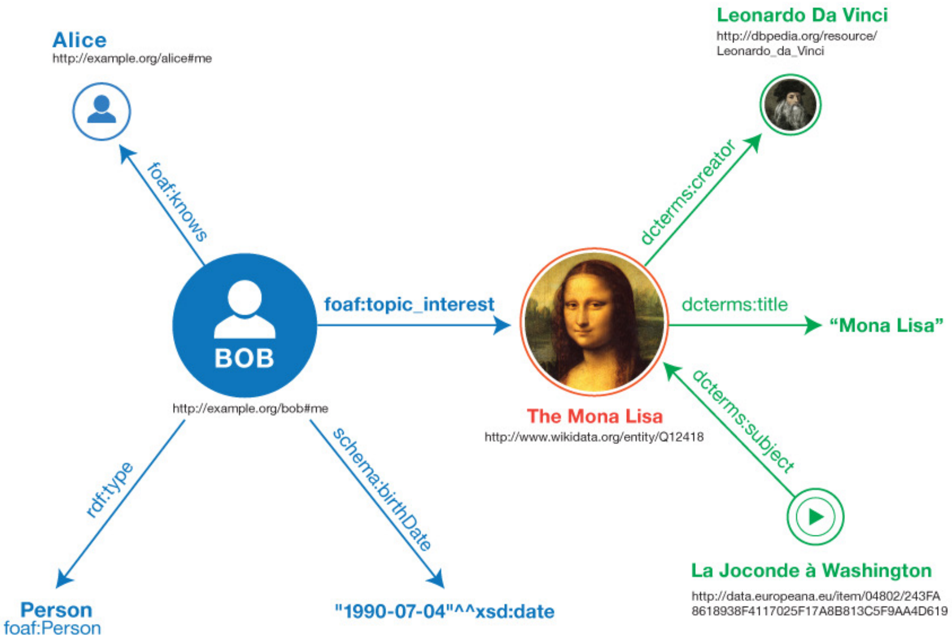
Resource Description Framework (RDF)

- RDF encode data in the form of triples



EXAMPLE 1: Sample triples (informal)

```
<Bob> <is a> <person>.
<Bob> <is a friend of> <Alice>.
<Bob> <is born on> <the 4th of July 1990>.
<Bob> <is interested in> <the Mona Lisa>.
<the Mona Lisa> <was created by> <Leonardo da Vinci>.
<the video 'La Joconde à Washington'> <is about> <the Mona Lisa>
```

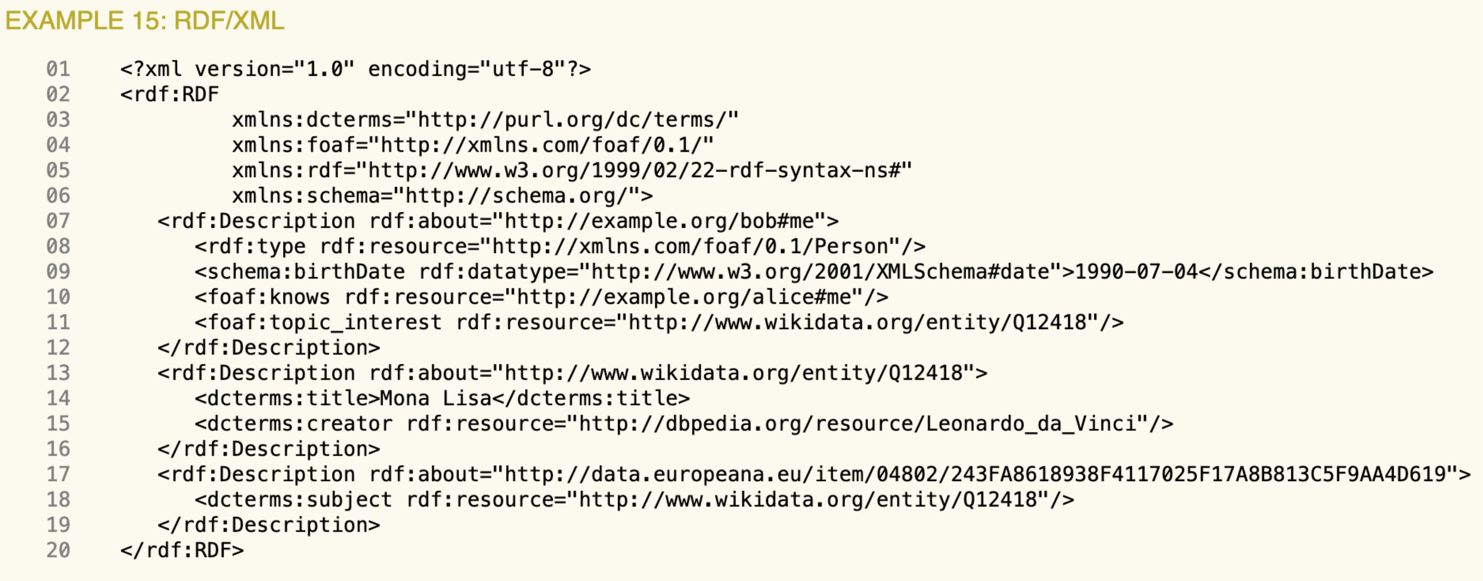


Alice
http://example.org/alice#me
Leonardo Da Vinci
http://dbpedia.org/resource/Leonardo_da_Vinci
foaf:knows
dcterms:creator
BOB
http://example.org/bob#me
foaf:topic_interest
dcterms:title
"Mona Lisa"
The Mona Lisa
http://www.wikidata.org/entity/Q12418
dcterms:subject
rdf:type
schema:birthDate
La Joconde à Washington
http://data.europeana.eu/item/04802/243FA8618938F4117025F17A8B813C5F9AA4D619
Person
foaf:Person
"1990-07-04"^^xsd:date

RDF Primer: https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/

RDF/XML

- Serialized RDF contains namespaces, prefixes, URIs



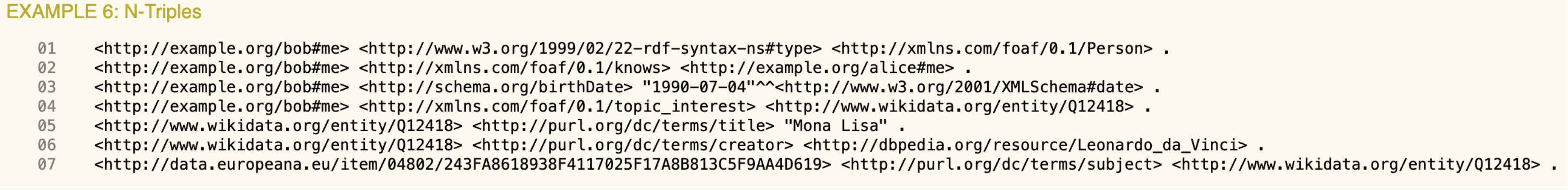
EXAMPLE 15: RDF/XML

```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF
         xmlns:dcterms="http://purl.org/dc/terms/"
         xmlns:foaf="http://xmlns.com/foaf/0.1/"
         xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:schema="http://schema.org/">
   <rdf:Description rdf:about="http://example.org/bob#me">
      <rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Person"/>
      <schema:birthDate rdf:datatype="http://www.w3.org/2001/XMLSchema#date">1990-07-04</schema:birthDate>
      <foaf:knows rdf:resource="http://example.org/alice#me"/>
      <foaf:topic_interest rdf:resource="http://www.wikidata.org/entity/Q12418"/>
   </rdf:Description>
   <rdf:Description rdf:about="http://www.wikidata.org/entity/Q12418">
      <dcterms:title>Mona Lisa</dcterms:title>
      <dcterms:creator rdf:resource="http://dbpedia.org/resource/Leonardo_da_Vinci"/>
   </rdf:Description>
   <rdf:Description rdf:about="http://data.europeana.eu/item/04802/243FA8618938F4117025F17A8B813C5F9AA4D619">
      <dcterms:subject rdf:resource="http://www.wikidata.org/entity/Q12418"/>
   </rdf:Description>
</rdf:RDF>
```

RDF Primer: https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/

N-Triples

- Serialized RDF contains namespaces, prefixes, URIs



EXAMPLE 6: N-Triples

```
<http://example.org/bob#me> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://xmlns.com/foaf/0.1/Person> .
<http://example.org/bob#me> <http://xmlns.com/foaf/0.1/knows> <http://example.org/alice#me> .
<http://example.org/bob#me> <http://schema.org/birthDate> "1990-07-04"^^<http://www.w3.org/2001/XMLSchema#date> .
<http://example.org/bob#me> <http://xmlns.com/foaf/0.1/topic_interest> <http://www.wikidata.org/entity/Q12418> .
<http://www.wikidata.org/entity/Q12418> <http://purl.org/dc/terms/title> "Mona Lisa" .
<http://www.wikidata.org/entity/Q12418> <http://purl.org/dc/terms/creator> <http://dbpedia.org/resource/Leonardo_da_Vinci> .
<http://data.europeana.eu/item/04802/243FA8618938F4117025F17A8B813C5F9AA4D619> <http://purl.org/dc/terms/subject> <http://www.wikidata.org/entity/Q12418> .
```

RDF Primer: https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/

Why Semantic Web, RDF, Ontologies, FOAF, etc.??

- Linked Open Data:

“The Semantic Web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other, related, data.”

- Linked Open Data Cloud

Linked Open Data: https://www.w3.org/DesignIssues/LinkedData.html

Data Analysis I

Definitions

- **Data** - are encodings that represent the qualitative or quantitative attributes of a variable or set of variables.

- A ***variable*** is a quantity, quality, or property that you can measure.

- A ***value*** is the state of a variable when you measure it. The value of a variable may change from measurement to measurement.

- An ***observation***, or a *case*, is a set of measurements made under similar conditions (you usually make all of the measurements in an observation at the same time and on the same object). An observation will contain several values, each associated with a different variable. **Sometimes refer to an observation as a data point**.

- **Dataset** – a set of data points (observations) collected for a particular purpose.

Data

- Types of Data:

– Qualitative (or Categorical)
– Quantitative (data like numeric values)

Data

- Qualitative:

Qualitative (or Categorical) data are descriptive, but not numeric.

Example: your eye-color, you gender, color of a vehicle, your birthplace

Data

- Quantitative:

Quantitative data take on numeric values

- Discrete data take whole numbers (0,1,2,3..), representing countable things.

e.g. The number of times an employee is late to work, number of cars in the parking garage.

- Continuous data can take infinite possibilities within a range of numeric values (fractions, decimals are included..)

e.g. height of a person, weight of an apple, number of times an employee late to work.

Courtesy: Quick Study Academic – Statistics www.quickstudy.com

Levels of Measurement

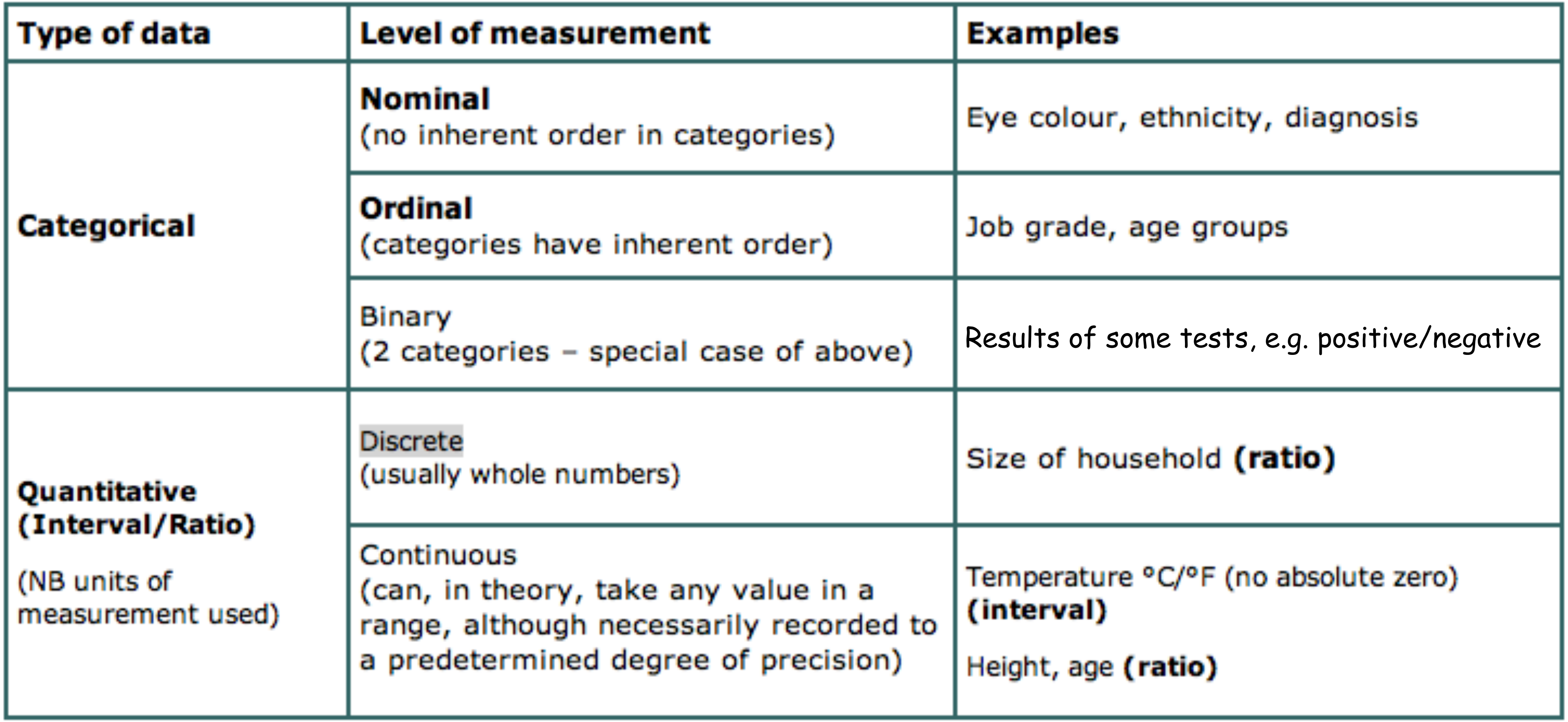
Qualitative (or Categorical) data can be measured at the:

- Nominal Level: Values are just names, without any order (example: eye-color)
- Ordinal Level: Values have some natural order, example: school class (freshman, sophomore, ..), test grade, military rank

Quantitative (data like numeric values) can be measured at the:

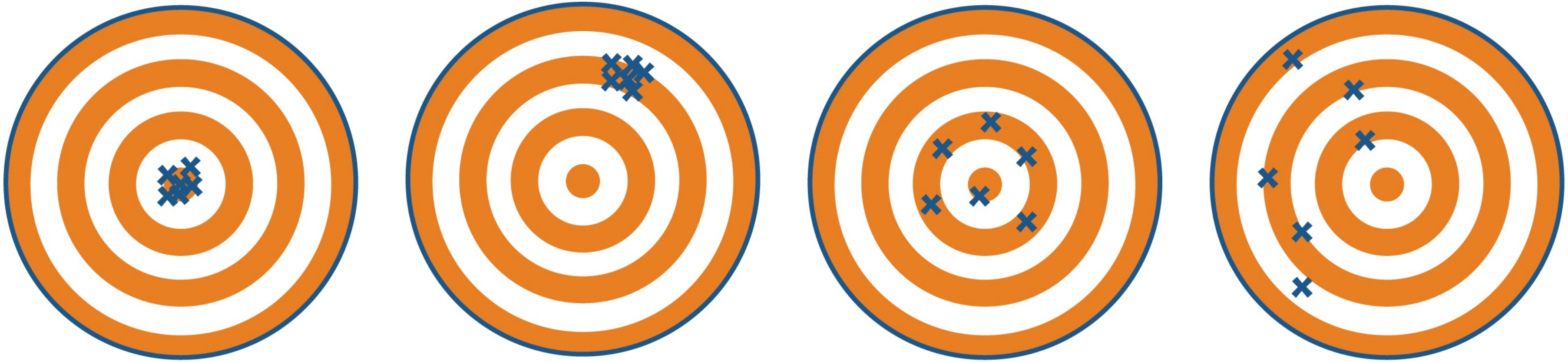
- Interval Level: Numeric data with no natural zero point; intervals (differences) are meaningful but ratios are not, can be nagative e.g. Temperature in Fahrenheit (or Celsius) degrees 80F is 20F hotter than 60F, but it is not 150% as hot.
- Ratio Level: Numeric data for which there is true zero, both intervals and ratios are meaningful, cannot be negative, e.g. weight, length, duration..

Types of Data



Type of data	Level of measurement	Examples
Categorical	**Nominal** (no inherent order in categories)	Eye colour, ethnicity, diagnosis
	Ordinal (categories have inherent order)	Job grade, age groups
	Binary (2 categories – special case of above)	Results of some tests, e.g. positive/negative
Quantitative (Interval/Ratio) (NB units of measurement used)	Discrete (usually whole numbers)	Size of household **(ratio)**
	Continuous (can, in theory, take any value in a range, although necessarily recorded to a predetermined degree of precision)	Temperature °C/°F (no absolute zero) **(interval)** Height, age **(ratio)**

Accurate vs. Precise



High Accuracy
High Precision

Low Accuracy
High Precision

High Accuracy
Low Precision

Low Accuracy
Low Precision

http://climatica.org.uk/climate-science-information/uncertainty

‘Signal to noise’

- Understanding accuracy and precision
  - Accuracy
  - Precision

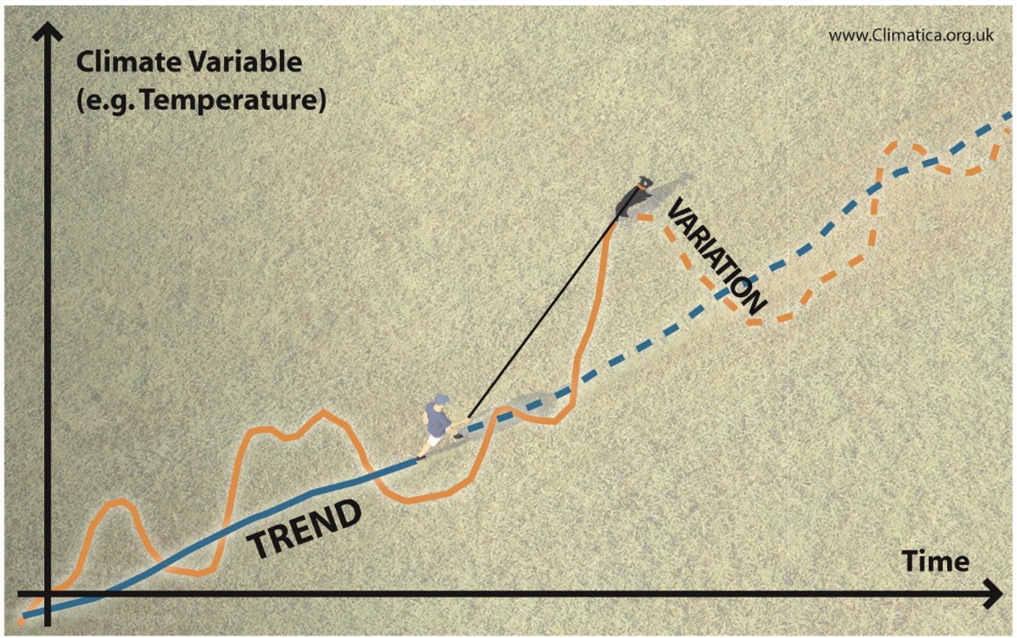
- Affects choices of analysis

- Affects interpretations

- Leads to data quality and assurance specification

- Signal and noise are context dependent

Data Analysis: Chasing Trends



Person walking a dog. The dog's path may vary as it wanders from a straight line, but both the dog and the person are headed the same way. You can predict where they will end up by looking at the trend of the person walking the dog.

http://climatica.org.uk/climate-science-information/uncertainty

Other considerations

- Continuous or discrete
- Underlying reference system
- Metadata standards and conventions

- The underlying data structures are important at this stage but there is a tendency to read in partial data/small amount of data
  - Why is this a problem? Because it can be biased
  - How to ameliorate any problems?

Outlier

- An extreme, or atypical, data value(s) in a sample.
- They should be considered carefully, before exclusion from analysis.
- For example, data values maybe recorded erroneously, and hence they may be corrected.
- However, in other cases they may just be surprisingly different, but not necessarily 'wrong'.

Special values in data

- Fill value
- Error value
- Missing value
- Not-a-number (NAN)
- Infinity
- Default
- Null

Exploratory Data Analytics (EDA)

- To use visualization and transformation to explore your data in a systematic way, a task that statisticians call exploratory data analysis, or EDA for short. EDA is an iterative cycle where you:

- Generate **questions** about your **data**.

- **Search** for **answers** by **visualizing, transforming,** and **modeling** your data.

- Use what you learn to refine your questions and/or generate **new questions**.

Exploratory Data Analytics (EDA)

- **EDA is not a formal process with a strict set of rules**. More than anything**, EDA is a state of mind**. During the initial phases of EDA you should feel free to investigate every idea that occurs to you.

- Some of these ideas will pan out, and some will be dead ends. As your exploration continues, you will hone in on a few particularly productive areas that you'll eventually write up and communicate to others.

- **EDA is an important part of any data analysis**, even if the questions are handed to you on a platter, because you always need to investigate the quality of your data.

- **Data cleaning is just one application of EDA**

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

- *There are no routine statistical questions, only questionable statistical routines.*
~ Sir David Cox

- *Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.*
~John Tukey

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

Exploratory Data Analytics (EDA)

- **Your goal during EDA is to develop an understanding of your data**.

- The easiest way to do this is to use questions as tools to guide your investigation. When you ask a question, **the question focuses your attention on a specific part of your dataset and helps you decide which graphs, models, or transformations to make.**

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

Exploratory Data Analytics (EDA)

creative process..

- **EDA is fundamentally a creative process**. And like most creative processes, the key to asking *quality* questions is to generate a large *quantity* of questions.

- At the beginning, It is difficult to ask revealing questions at the start of your analysis because you do not know what insights are contained in your dataset.

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

Exploratory Data Analytics (EDA)

No rule on which Question..

- **There is no rule about which questions you should ask to guide your research**.

- However, two types of questions will always be useful for making discoveries within your data. You can loosely word these questions as:

1) What type of variation occurs within my variables?
2) What type of covariation occurs between my variables?

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

Variation...

- ***Variation*** is the tendency of the values of a variable to change from measurement to measurement. You can see variation easily in real life; if you measure any continuous variable twice, you will get two different results. This is true even if you measure quantities that are constant, like the speed of light. Each of your measurements will include a small amount of error that varies from measurement to measurement.

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

Variation...

- **Categorical variables can also vary if you measure across different subjects (e.g. the eye colors of different people**), or different times (e.g. the energy levels of an electron at different moments).

- Every variable has its own pattern of variation, which can reveal interesting information. **The best way to understand the pattern is to visualize the distribution of variables' values**.

Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

Visualizing Distributions

- How you visualize the distribution of a variable will depend on whether the variable is categorical or continuous.

A variable is *categorical* if it can only take one of a small set of values.

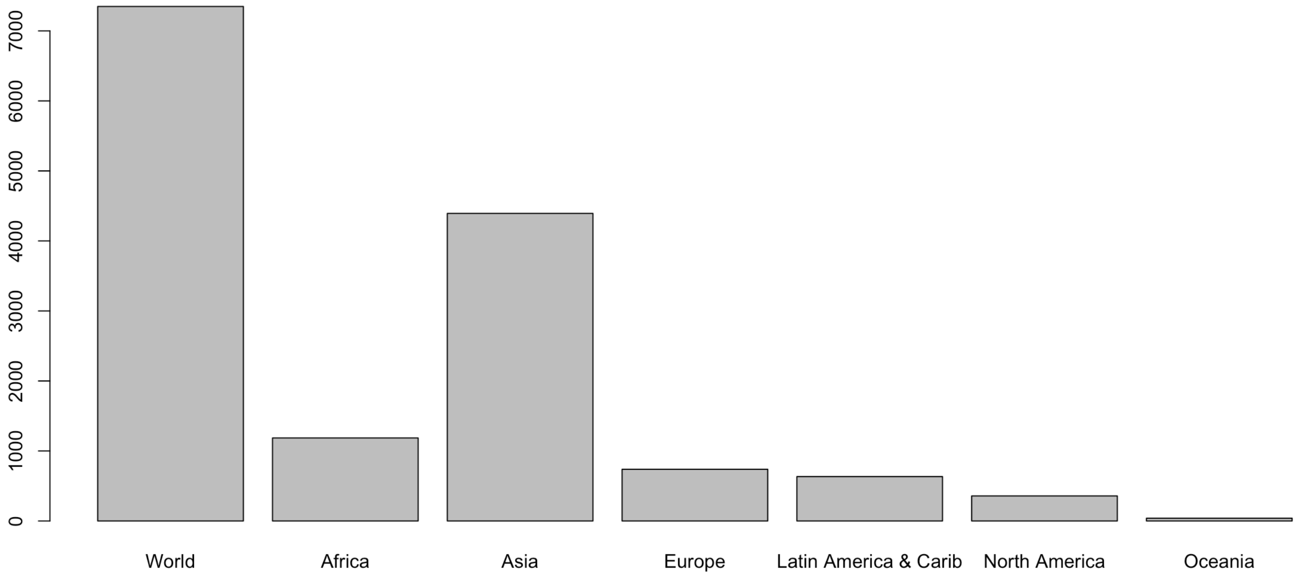
- To examine the distribution of a categorical variable, use a bar chart.

A variable is *continuous* if it can take any numeric value within an interval.

- To examine the distribution of a categorical variable, use a histogram.

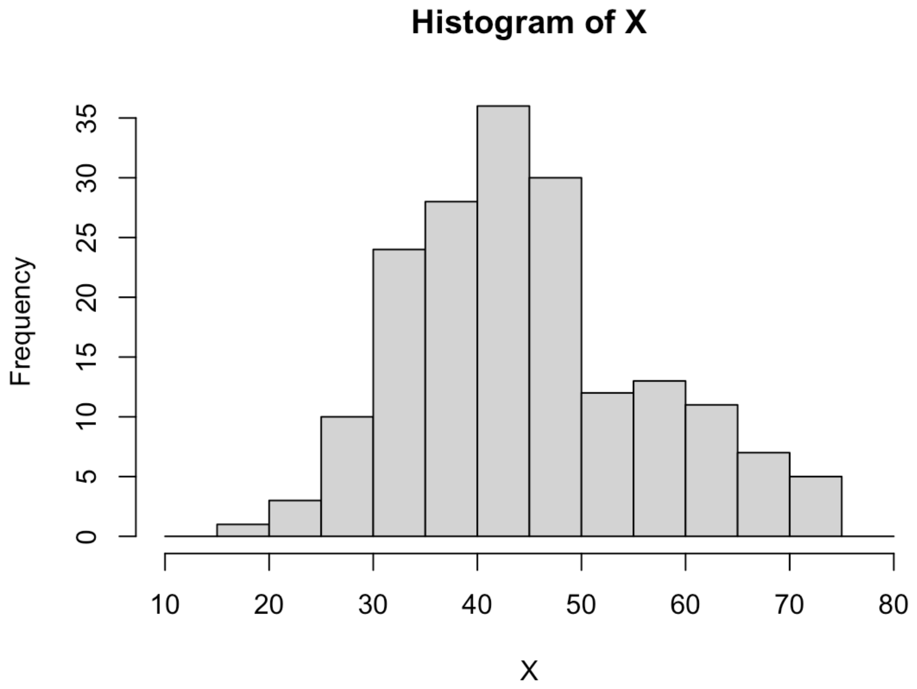
Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham,* Chapter 5, https://r4ds.had.co.nz/

Bar plot



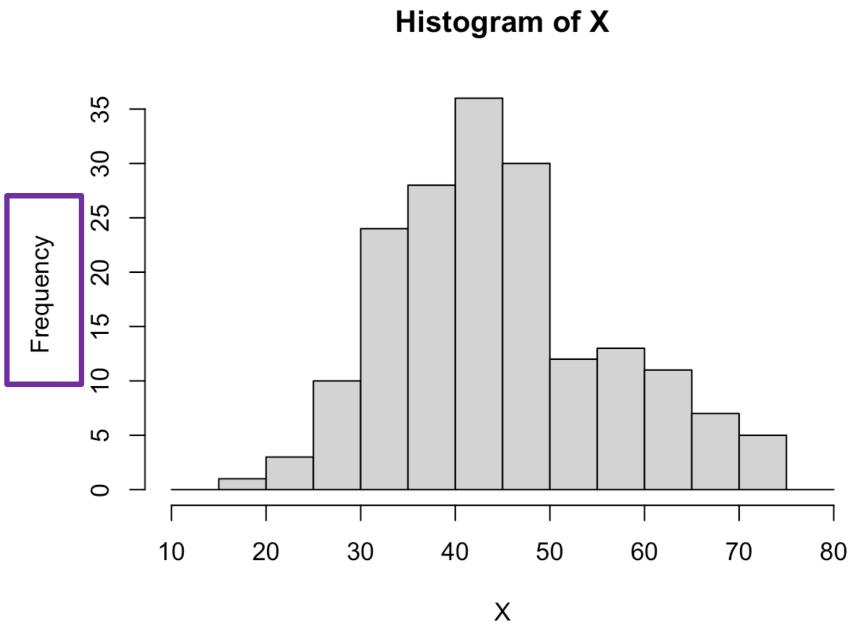
0
1000
2000
3000
4000
5000
6000
7000
World
Africa
Asia
Europe
Latin America & Carib
North America
Oceania

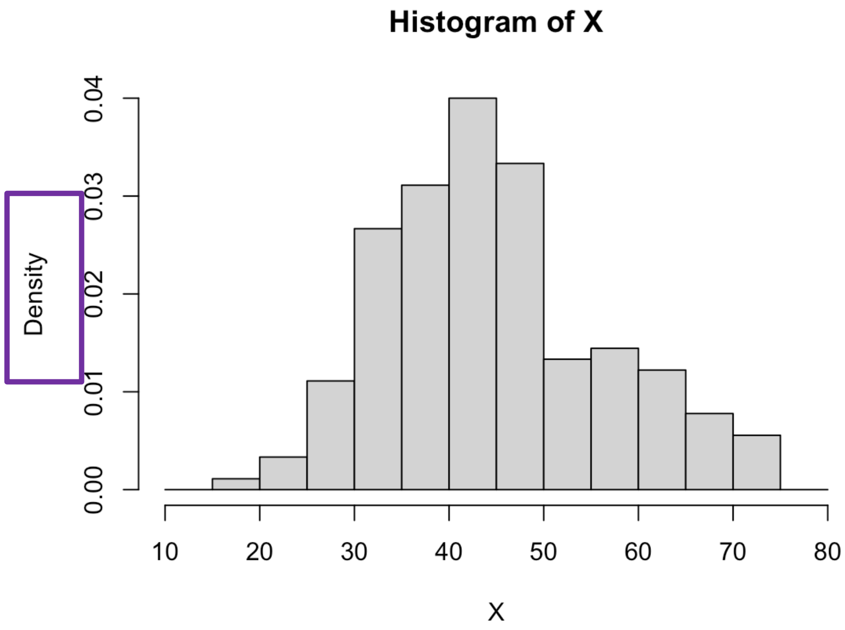
Grouped Frequency Distribution aka binning



Bin size = 5

Frequency vs. Density





- 36 observations where 40 < x < 45
- 36/180 (total) = 0.2 or 20%

- density at bar = 0.04 where 40 < x < 45
- area of bar = 0.04 * 5 (width of bar) = 0.2 or 20%

Visualizations

Visualizations

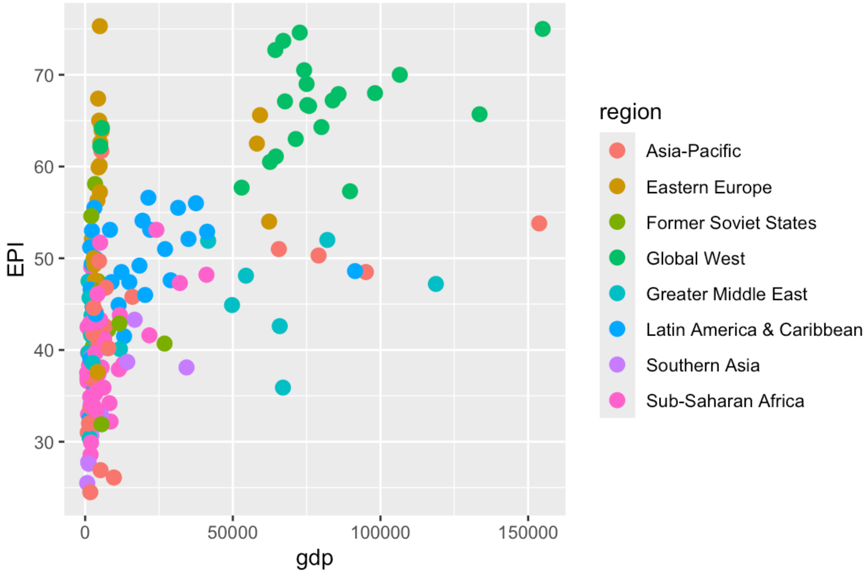
- Scatter Plot – Paired data (x,y)

- 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

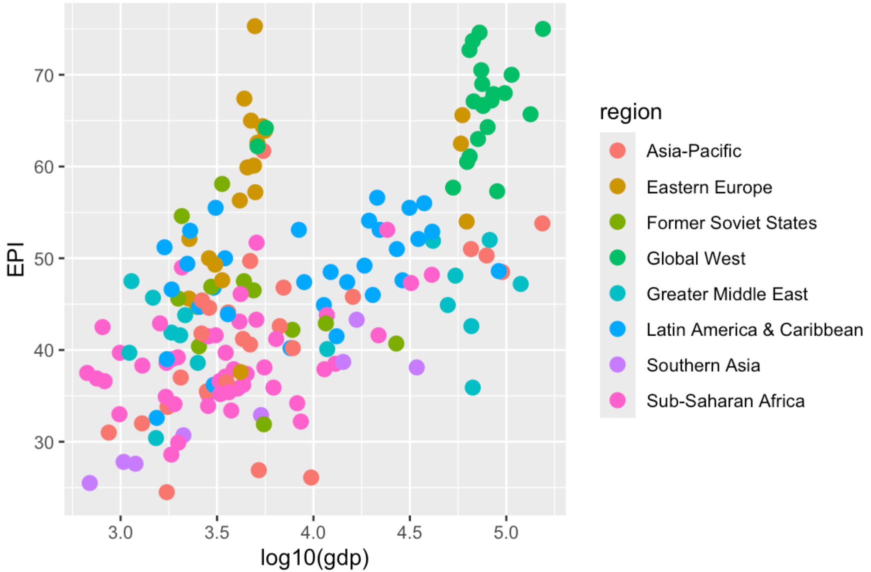
Visualizations

- Scatter Plot – Paired data (x,y)

GDP vs EPI by Region



log10(GDP) vs EPI by Region



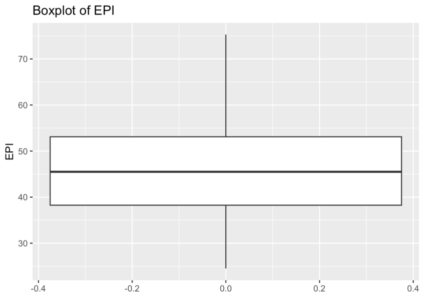
Visualizations

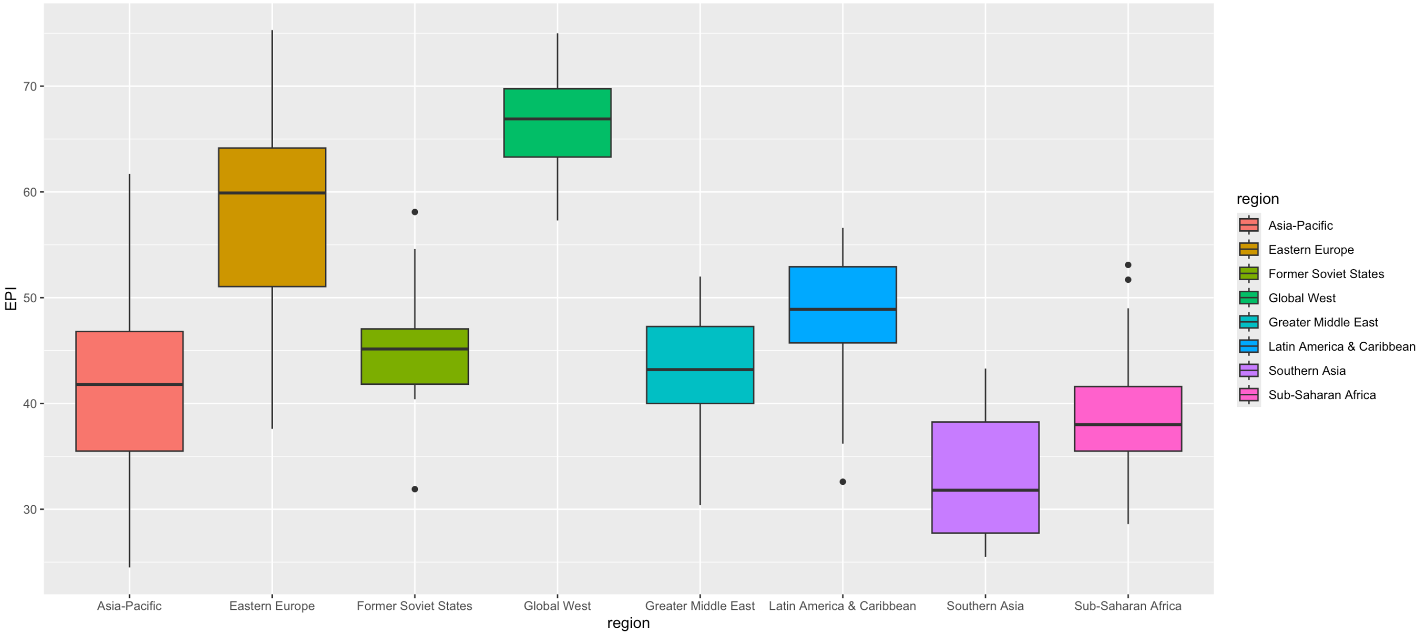
- Boxplot - box-and-whisker of a variable x

- 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

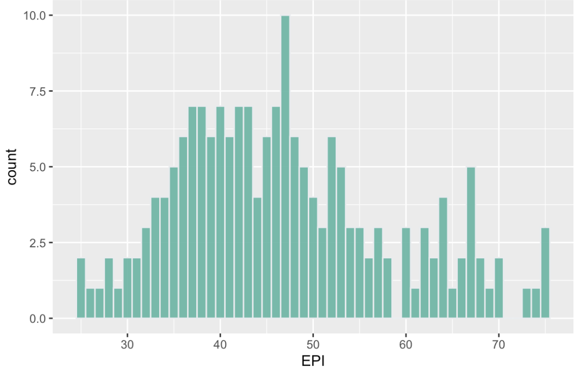
- Histogram – Distribution of variable X

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

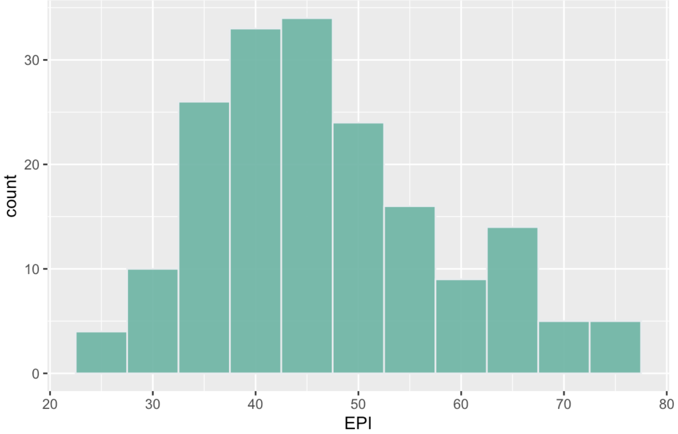
Visualizations

- Histogram – Distribution of variable X

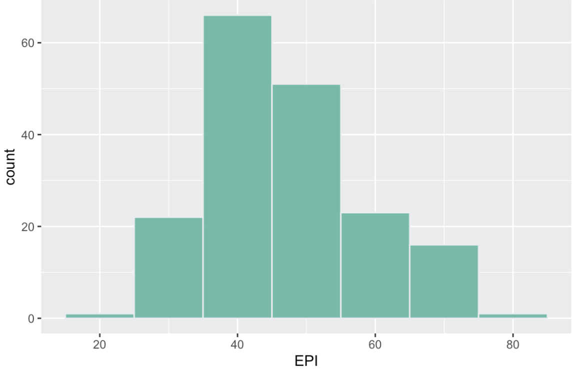
Histogram of EPI - Bin size = 1



Histogram of EPI - Bin size = 5



Histogram of EPI - Bin size = 10



Statistical Concepts

Populations and samples

- Population : The complete set of actual or potential elements about which inferences are made

- Sample : A subset of the population selected using some sampling methods.

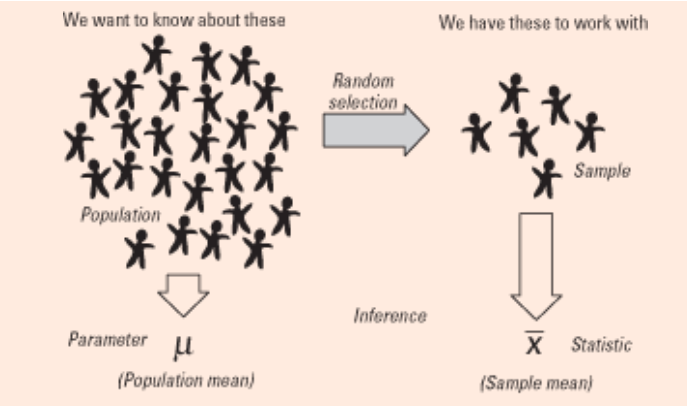


Image Credit/Reference: Quick Study Statistics

Sampling Types

- **Random Sampling**

– Sampling in which the data is collected using chance methods or random numbers.

- **Systematic Sampling**

– Sampling in which data is obtained by selecting every *k*th object.

- **Convenience Sampling**

– Sampling in which data that is readily available is used.

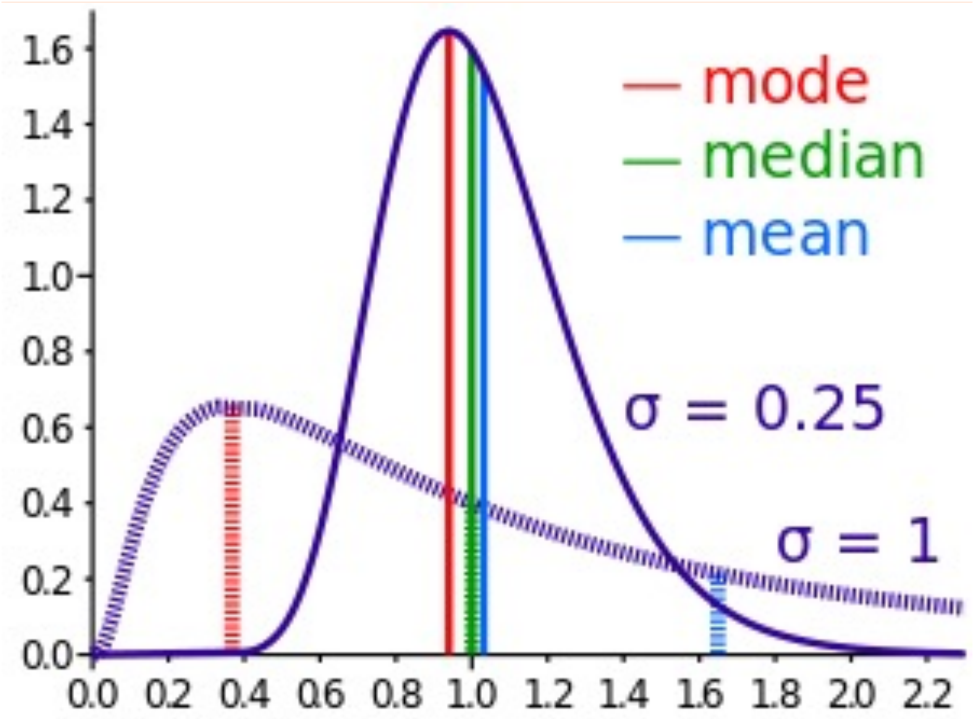
- **Stratified Sampling**

– Sampling in which the population is divided into groups (called strata) according to some characteristic. Each of these strata is then sampled using one of the other sampling techniques.

- **Cluster Sampling**

– Sampling in which the population is divided into groups (usually geographically). Some of these groups are randomly selected, and then all of the elements in those groups are selected.

Statistics - central tendency – mean, median, mode

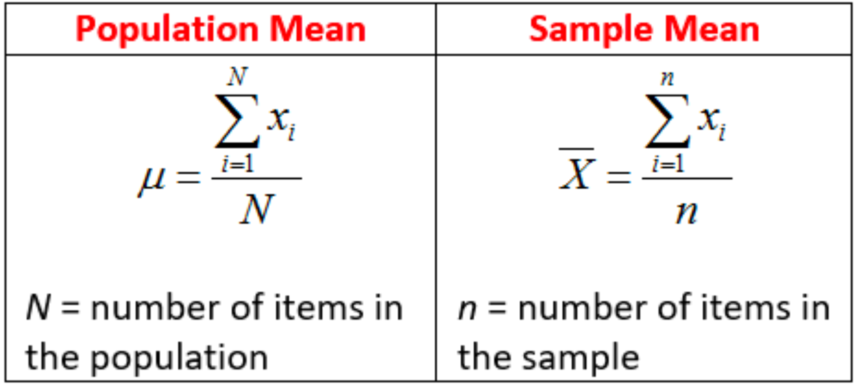


Measures of Central Tendency

- Mean: The most commonly used measure of central tendency, commonly referred to as “Average”, sensitive to extreme values (sensitive to outliers)

  – Population Mean

  – Sample Mean



Population Mean	Sample Mean
$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
N = number of items in the population	*n* = number of items in the sample

Image Resource: https://www.onlinemathlearning.com/population-mean.html Courtesy: Quick Study Academic – Statistics www.quickstudy.com

Measures of Central Tendency

- **Median**: Value that divides the set in 2 so the same number of observations lie on each side of it.

- ***The median is less sensitive to extreme values***

- For an even number, it is the average of middle two values.

1, 3, 3, **6**, 7, 8, 9

Median = $\underline{\underline{6}}$

1, 2, 3, **4**, **5**, 6, 8, 9

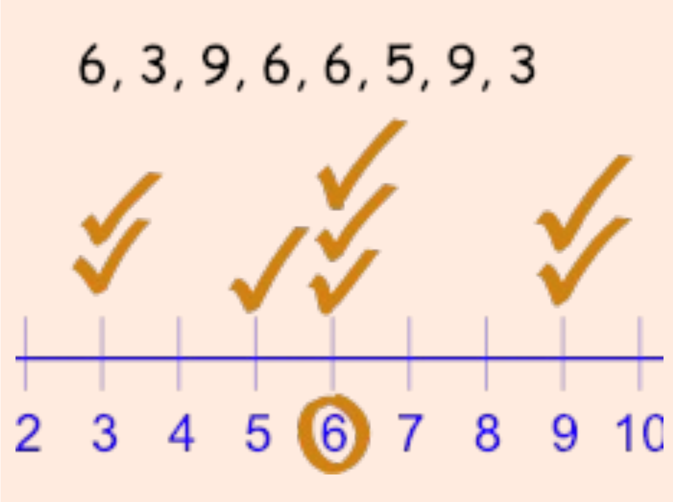
Median = $(4 + 5) \div 2$

= $\underline{\underline{4.5}}$

Courtesy: Quick Study Academic – Statistics www.quickstudy.com Image Resource: https://en.wikipedia.org/wiki/Median

Measures of Central Tendency

- **Mode**: Observation that occurs with the greatest frequency.



Courtesy: Quick Study Academic – Statistics www.quickstudy.com Image Resource: https://www.mathsisfun.com/definitions/mode.html

Correlation

- One measure of the strength of the association between two numerical variables is correlation.

- Correlation describes the strength of the linear association between two variables.

- Correlation coefficient is between -1 and +1

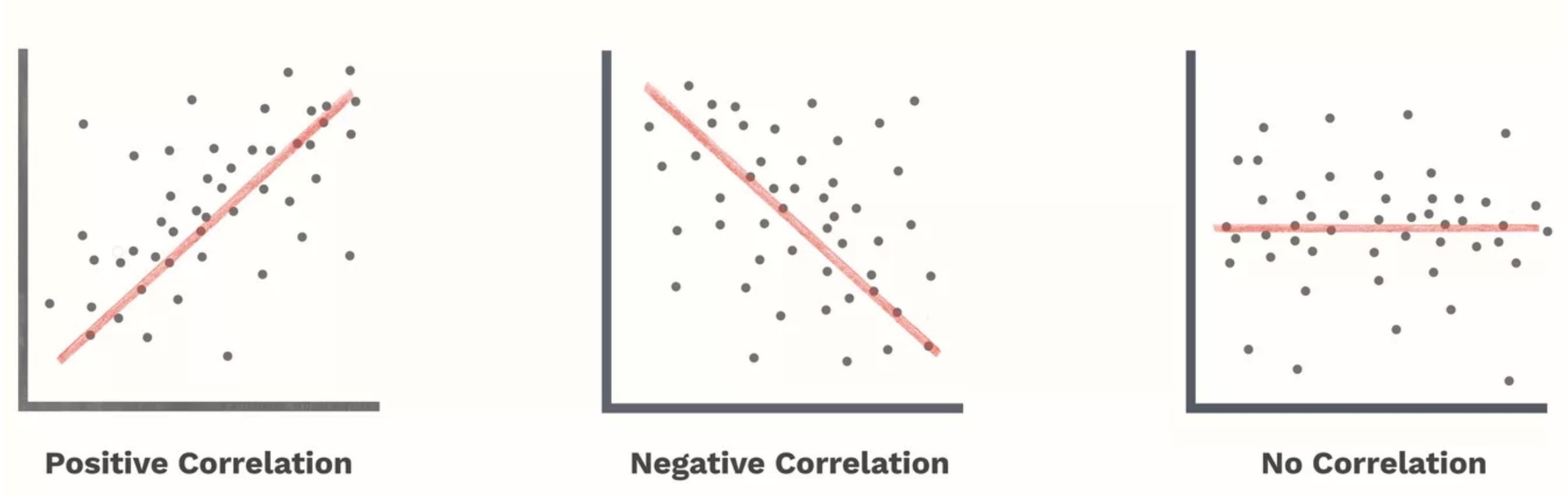
-1 indicates a perfect negative linear association and +1 indicates a perfect positive linear association. A correlation coefficient of 0, indicates that there is no linear relationship in the two variables. -0.1 and +0.1, indicates no linear relationship *or a very weak* linear relationship

- Correlation coefficient is sensitive to outliers.

- Correlation coefficient is unitless.

Reference(s): https://www.investopedia.com/terms/c/correlationcoefficient.asp
https://www.investopedia.com/ask/answers/032515/what-does-it-mean-if-correlation-coefficient-positive-negative-or-zero.asp

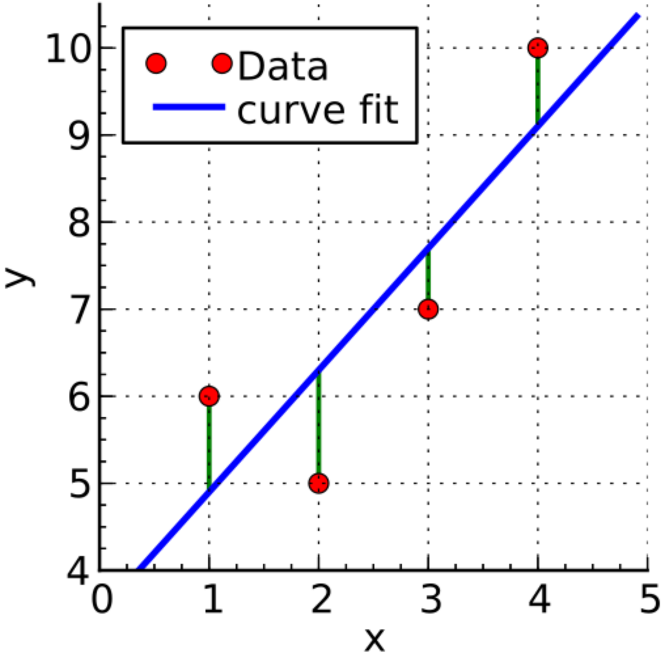
Correlation...



Image/Photo Credit: https://www.investopedia.com/ask/answers/032515/what-does-it-mean-if-correlation-coefficient-positive-negative-or-zero.asp

Linear Regression

Linear Regression



Regression

Linear Regression: In regression, fitting covariate and response data to a line is referred to as linear regression.

Covariate: A variable that is possibly predictive of the outcome under study control variable, ***explanatory variable, independent variable, predictor***

Response: dependent variable

Intercept: The expected value of the response variable when the value of the predictor variable is 0.

Slope: the average increase in Y associated with a one-unit increase in X

Reference/Resources:
The Elements of Statistical Learning. Hastie • Tibshirani • Friedman, 2nd Edition.
Introduction to Probability and Statistics, 4th Edition by Beaver.
Introduction to Statistical Learning with R, 7th Edition (ISLR).

Simple Linear Regression

- Let’s take a look at the Least Squares Method for a single covariate (single regression).

- Utilizing the statistical notion of estimating parameters from data points, we find the estimates (coefficients) using the least squares method.

- We will look later at evaluating linear models.

Simple Linear Regression

- Most commonly used approach is the *Least Squares*
- Least Squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS

$$\underbrace{\hat{y}}_{\text{Predicted response}} = \underbrace{\hat{\beta}_0}_{\text{Intercept}} + \underbrace{\hat{\beta}_1}_{\text{Slope}} \underbrace{x}_{\text{Explanatory variable}}$$

- $\hat{y}$ = Predicted value of the response variable
- x = Explanatory variable (x)
- $\hat{\beta}_0$ = Intercept
- $\hat{\beta}_1$= Slope

Least Squares Method

Equation of line: $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x$

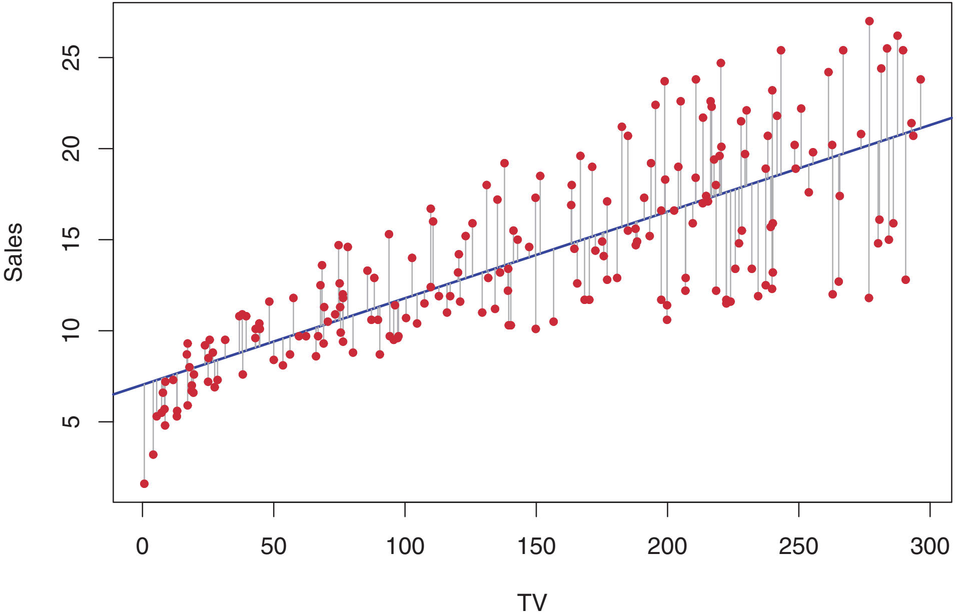
Let *n* be a positive integer. For a given data *(x1,y1), ..., (x$_n$,y$_n$)* ∊ℝ×ℝ,

- we obtain the intercept β0 and slope β1 using the least squares method.
- Residual Sum of Squares (RSS), the *i*th residual $e_i = y_i - \hat{y}_i$

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$

Evaluating Linear Models

- Sales vs. TV ad spending
- Sales in 1000s of units
- TV ad spending in 1000s of $

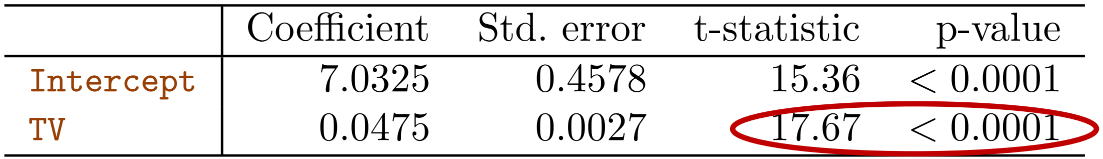


Evaluating Linear Models

Values of coefficients >> their Std. errors

High t-statistic

Very low p-value



	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

Hypothesis (more TV ads → more sales)

H0 : There is no relationship between X and Y

Ha : There is some relationship between X and Y

$$t_{\widehat{\beta}} = \frac{\widehat{\beta} - \beta_0}{\mathbf{SE}(\widehat{\beta})}$$

Reject the null hypothesis!

Residual Standard Error

- Mean sales $\approx$ 14,000 units

RSE = 3.26 = 3,260 units

good/bad?

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

R^2

- measures the proportion of the variability in *Y* that can be explained using *X*
- has a value between 0,1

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

$$\mathrm{TSS} = \sum (y_i - \bar{y})^2$$

Assignment 1 due tonight!

Assignment 2 instructions out..

Please email me the names of your team members, let me know if you don’t have a team.

No readings today!

Thanks!

Work on your data collection!!