Data Analysis – Part I Knowledge Graphs, Concepts for Analysis, Project Groups

Ahmed Eleish

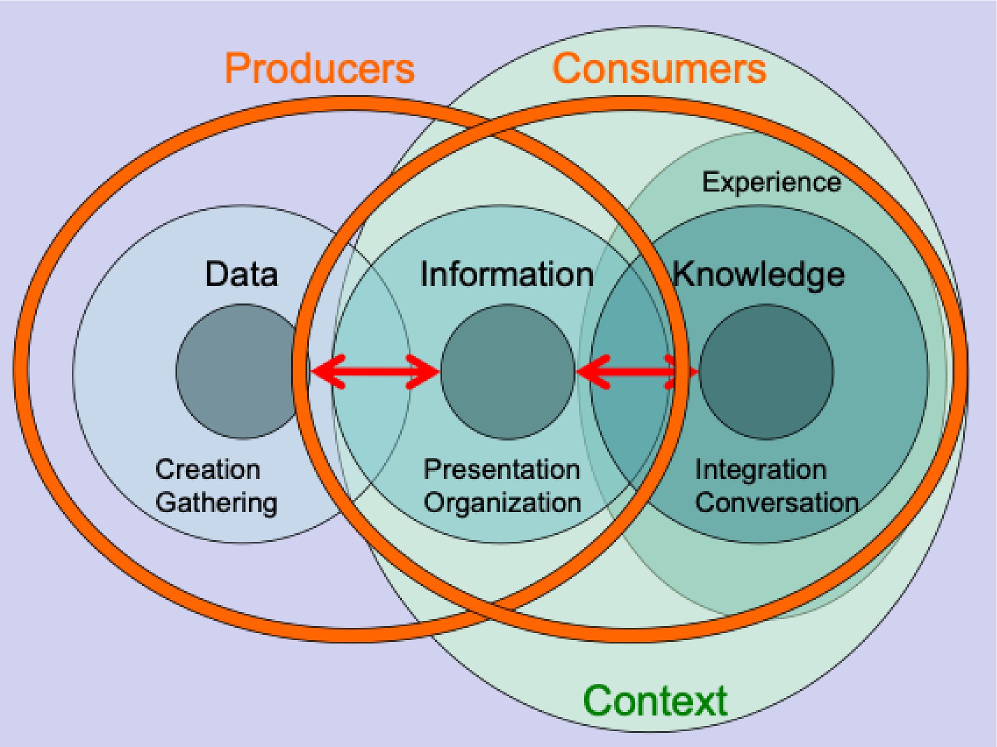
Data Science – ITWS/CSCI/ERTH-4350/6350 Module 4, September 19th, 2024

Tetherless World Constellation
Rensselaer Polytechnic Institute

Contents

- Module 2 & 3 review
- Knowledge Graphs
- Data in context
- Building Knowledge Graphs
- Data Analysis Concepts, Exploratory Analysis
- Visualizations, Distributions, Statistics, Regression

Data-Information-Knowledge Ecosystem



Modes of collecting data, information

- Observation
- Measurement
- Generation

- Driven by
  - Questions
  - Research idea
  - Exploration

Data Management

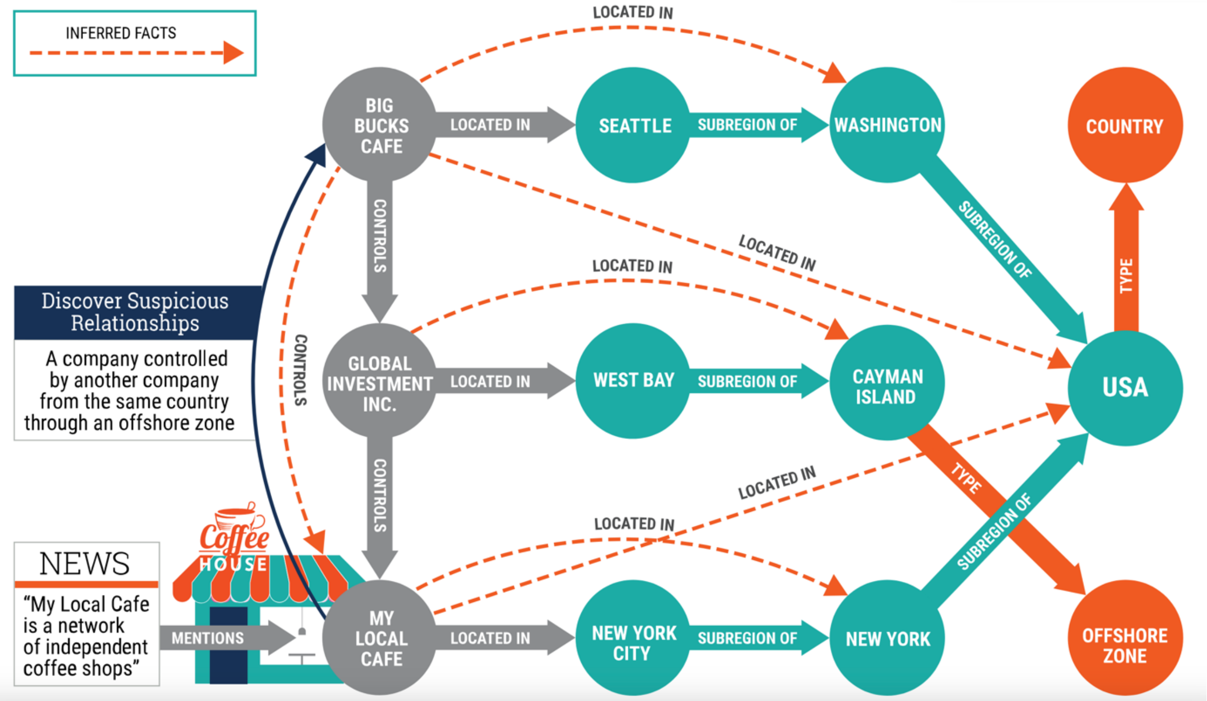
- Creation of logical collections
- Physical data handling
- Interoperability support
- Security support
- Data ownership
- Metadata collection, management and access
- Persistence
- Knowledge and information discovery
- Data distribution and publication

When thinking of provenance

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

Knowledge Graphs

What is a Knowledge Graph?



Discover Suspicious Relationships

A company controlled by another company from the same country through an offshore zone

NEWS

"My Local Cafe is a network of independent coffee shops"

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
- Data Integrations
- Answering Questions

What are Knowledge Graphs

- 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/

How do we create KGs?

Let’s focus on the following questions as we move along with this lecture:

- How do we create Knowledge Graphs?
- How do we use Knowledge Graphs with modern AI algorithms?
- What are open research questions in the field of Knowledge Graphs?
- What are some practical applications with Knowledge Graphs?

Knowledge Graphs (KGs)

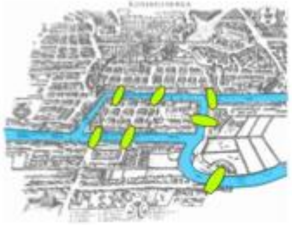
What are “Graphs”?

- **Knowledge Graphs are a type of graph.**
- Therefore, it is important to understand the basics of “graphs” before we move on to “Knowledge Graphs”.
- **Graphs are simple structures**, where we use:

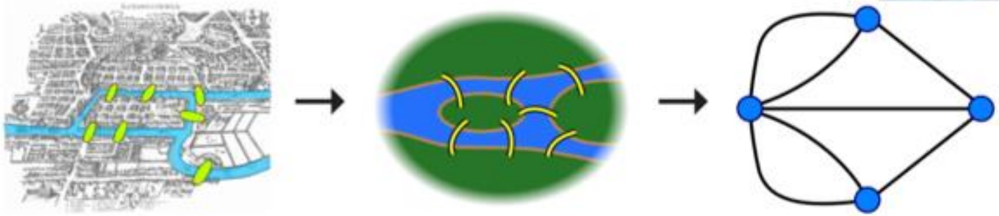
Nodes (or Vertices) connected by Relationships (or Edges)

Graphs

- Graphs we talk about during this lecture, sometimes referred to as networks.
- They are a simple but very powerful way of describing how things are connected.
- Graphs are not a new thing. 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

Nodes and Relationships

“Nodes” (or Vertices) connected by “Links/Relationships” (or Edges)

Taxonomy

- To support “*x is a kind of y*” reasoning, we need a more hierarchical view called “taxonomy”.
- **“Taxonomy” is a classification scheme that organizes categories** in a broader-narrower hierarchy.
- Items that share similar qualities are grouped into the same category and the **taxonomy provides a global organization by relating categories to one another**.

Resource/Reference: Knowledge Graphs: Data in Context for Responsive Businesses

Taxonomy

- The hierarchy is constructed with category nodes connected by *sub-category_of* relationships.
- The beauty of the Knowledge Graph(KG) is that we can choose to use multiple hierarchical organizations simultaneously to provide even more insights.
- Classification is completely dynamic in a KG. The new categories and their associativity as well as the linage to the categories are just additional nodes and relationships in the KG.

Data Analysis

Induction or deduction?

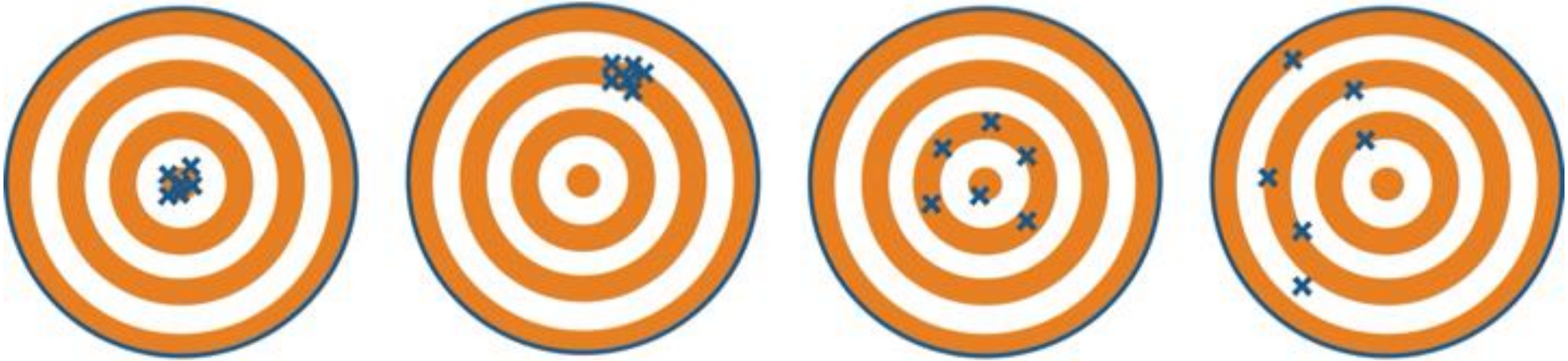
- Induction: The development of theories from observation

– Qualitative – usually information-based

- Deduction: The testing/application of theories

– Quantitative – usually numeric, data-based

Accurate vs. Precise



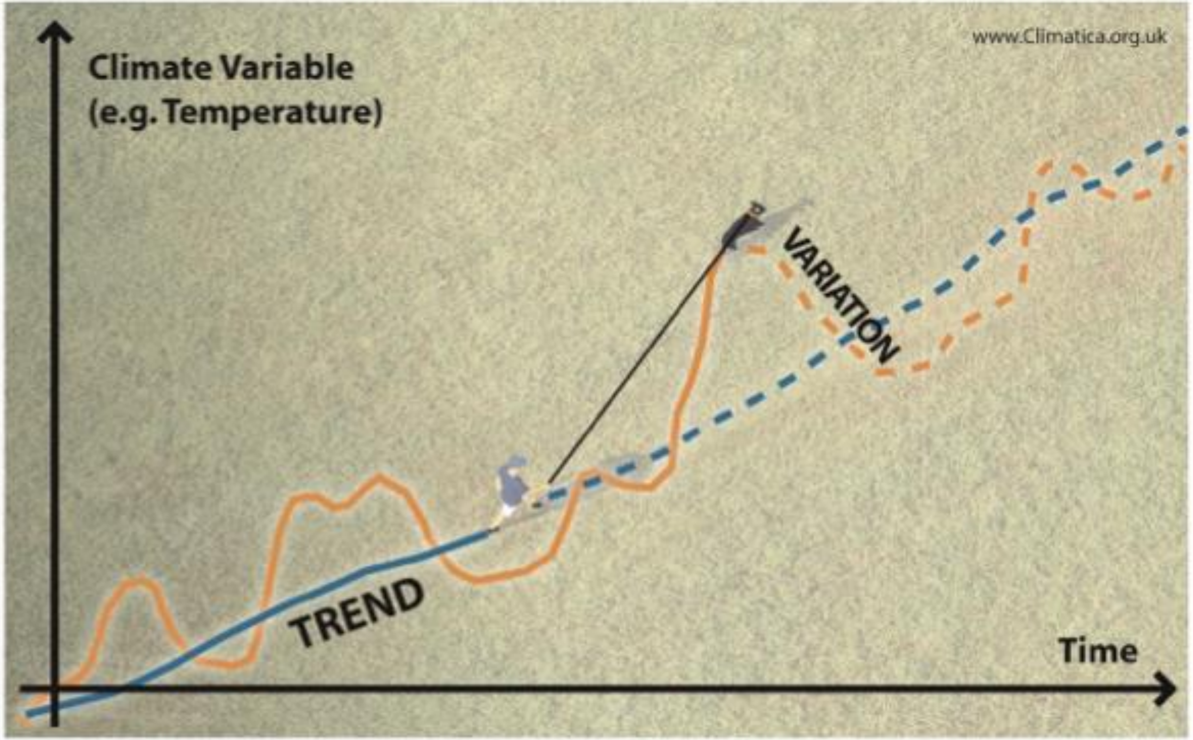
High Accuracy
High Precision

Low Accuracy
High Precision

High Accuracy
Low Precision

Low Accuracy
Low Precision

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



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

‘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

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
- Rational numbers

Exploratory Data Analytics (EDA)

- How 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.

You will:

- **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.

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

- **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.

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

- **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/

- **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/

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/

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/

Variable, Value, Observation

- 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**.

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.

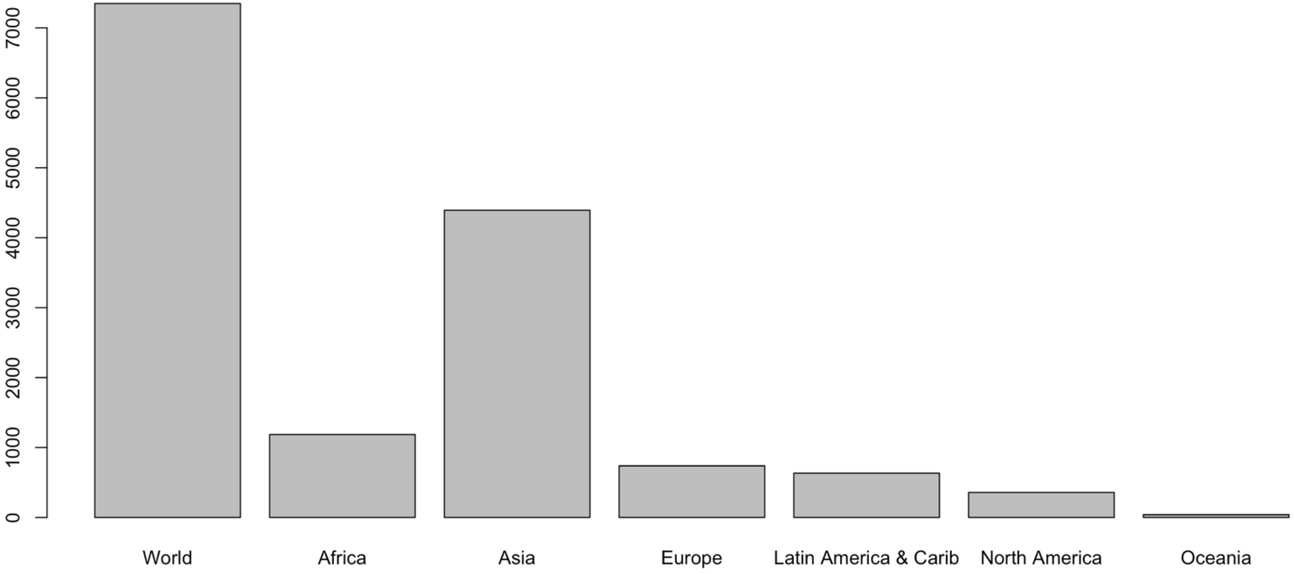
- To examine the distribution of a categorical variable, use a bar chart.

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

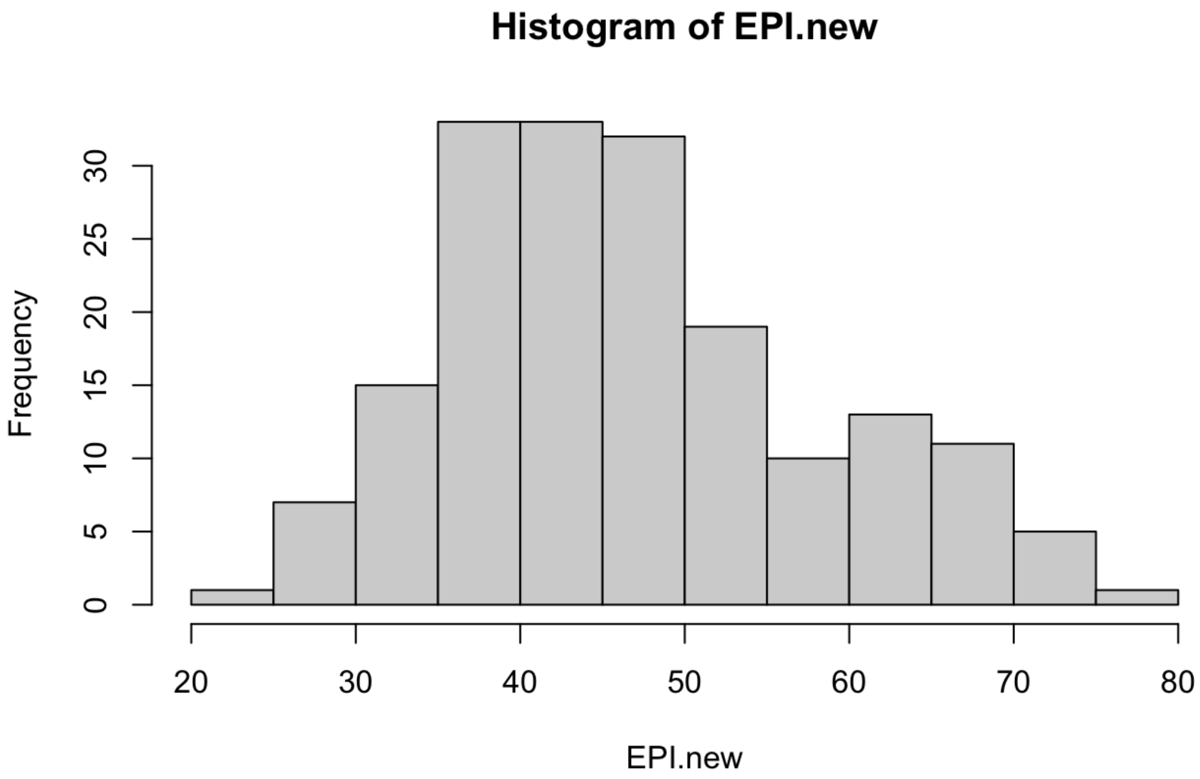
- To examine the distribution of a categorical variable, use a histogram.

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

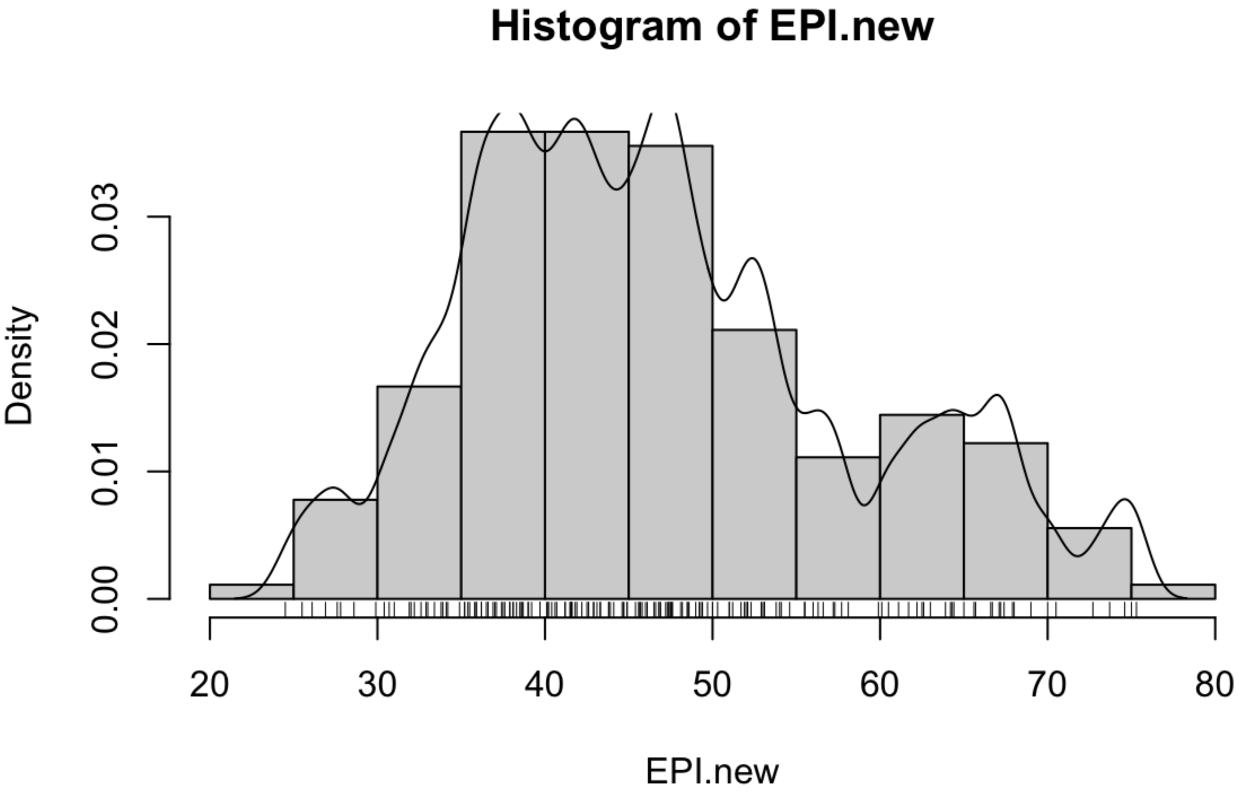
Bar plot


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

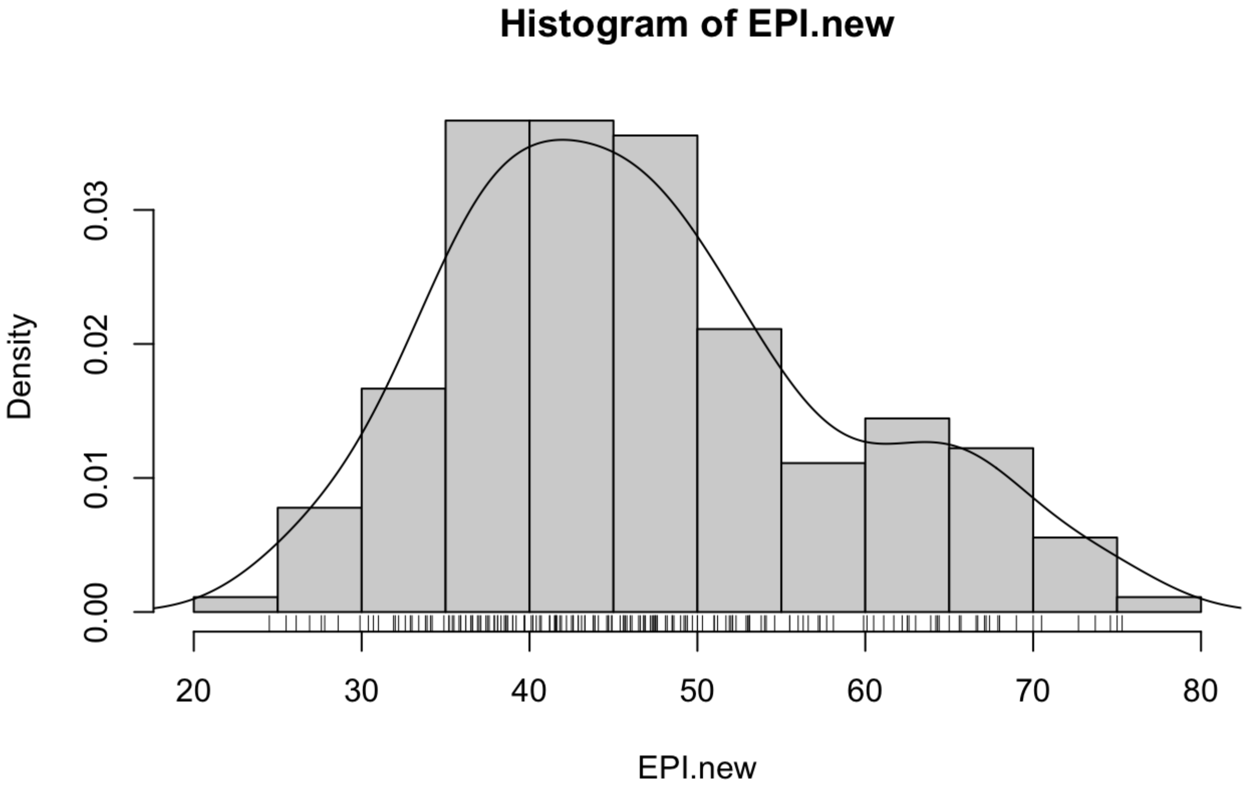
Histogram – type of bar plot



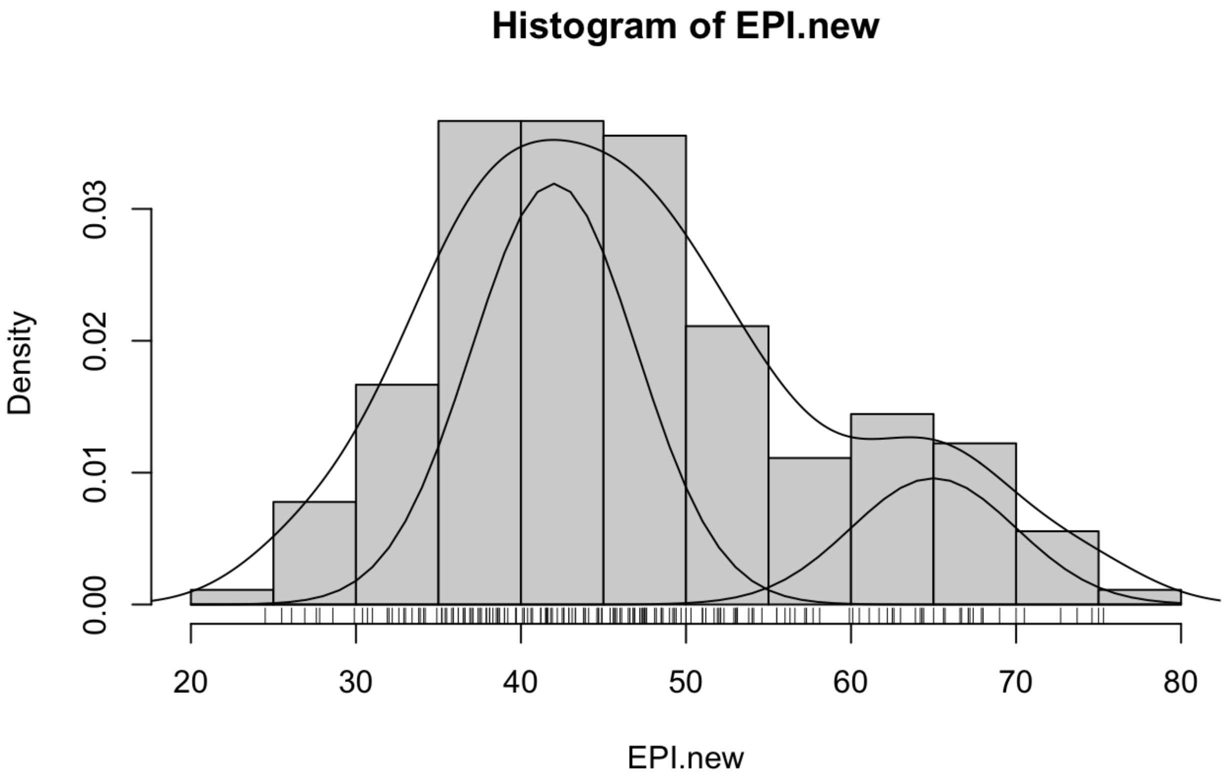
Histogram – type of bar plot



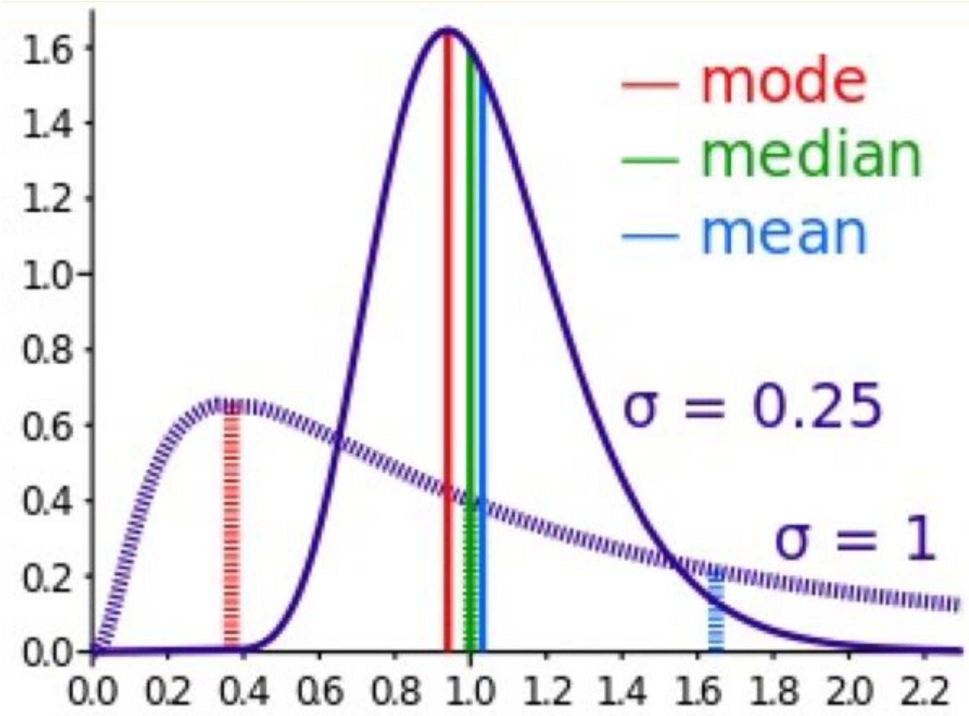
Histogram – type of bar plot



Histogram – type of bar plot



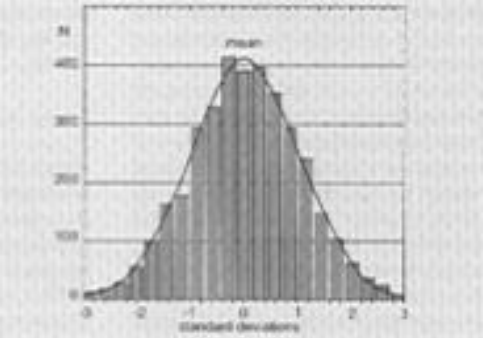
Statistics - central tendency – median, mean, mode



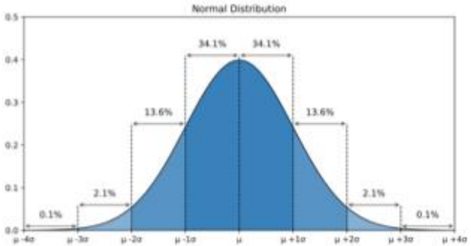
Statistics

- We will most often use a Gaussian distribution (*aka* normal distribution, or bell-curve) to describe the statistical properties of a group of measurements.

- The variation in the measurements taken over a finite spatial region may be caused by intrinsic spatial variation in the measurement, by uncertainties in the measuring method or equipment, by operator error, ...



The standardized normal curve fitted to experimental data



- Roughly 68.3% of the data is within 1 standard deviation of the average (from μ-1σ to μ+1σ)
- Roughly 95.5% of the data is within 2 standard deviations of the average (from μ-2σ to μ+2σ)
- Roughly 99.7% of the data is within 3 standard deviations of the average (from μ-3σ to μ+3σ)

Image Credit: W3C school:
https://www.w3schools.com/statistics/statistics_normal_distribution.php

Mean and standard deviation

•The mean, *m*, of *n* values of the measurement of a property *z* (the average).

m=[SUM{i=1,n}z_i]/n

•The standard deviation *s* of the measurements is an indication of the amount of spread in the measurements with respect to the mean.

s^2 =[SUM{i=1,n}(z_i -m)$_2$]/n

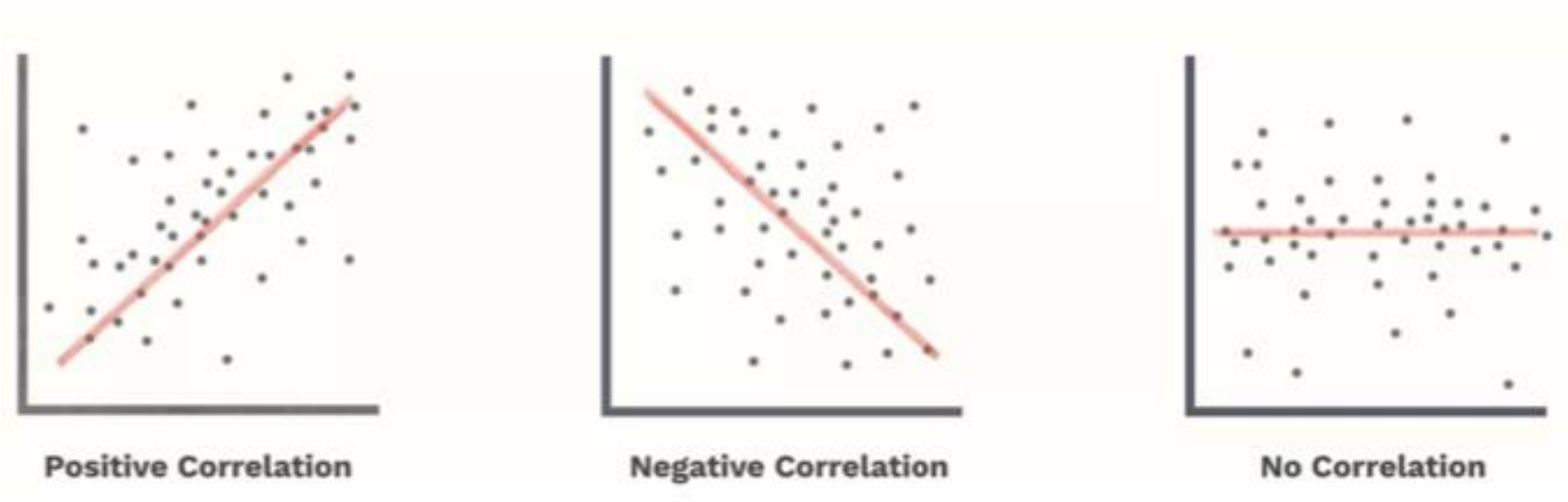
•The quantity s^2 is known as the variance of the measurements.

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. The 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

Input/Output

- Input: input go by different names,

input: *features, predictors*, *independent variables*, sometimes just variables

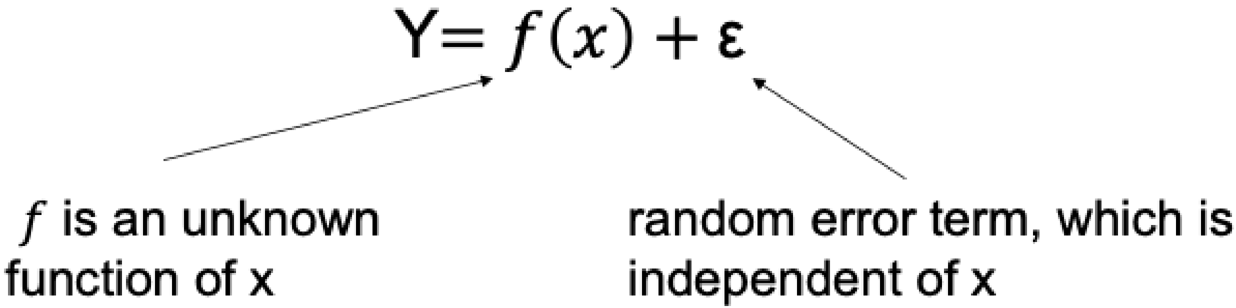
$$X = (x_1, x_2, \ldots, x_n)$$

- **Output**: The output variable called *response* or *dependent variable*, typically denoted by *Y*

- Suppose that we observe quantitative response Y with *p* different predictor variables,

$x_1, x_2,\ldots x_p$.

- We assume some relationship between Y and X =($x_1, x_2,\ldots x_p$) , which can be written as:

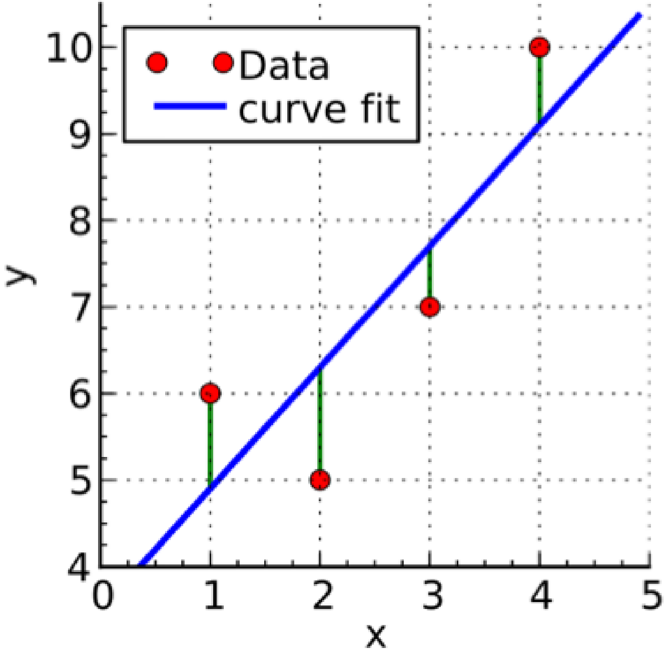


$$Y = f(x) + \varepsilon$$

f is an unknown function of x

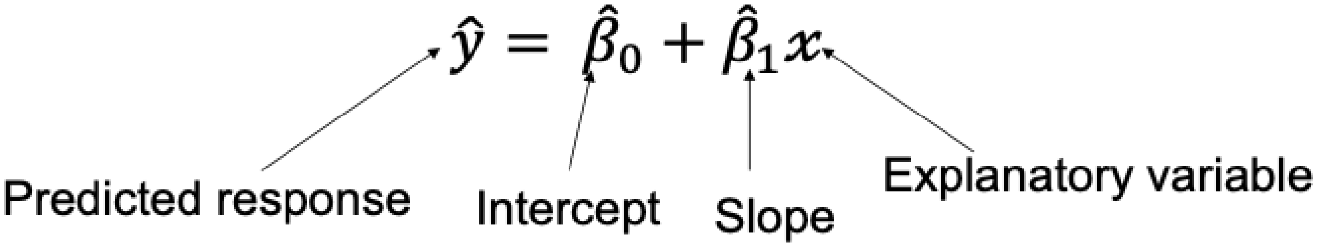
random error term, which is independent of x

Regression



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



$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Predicted response

Intercept

Slope

Explanatory variable

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

Residuals ...

- The residual is defined as the difference between the observed value and the predicted value.(Difference between the observed value and the predicted value of the response variable for a given data point).

$e_i = y_i - \hat{y}_i$ represents the *ith* residual,

this is the difference between the *ith* observed response value and the *ith* response value that is predicted by the linear model.

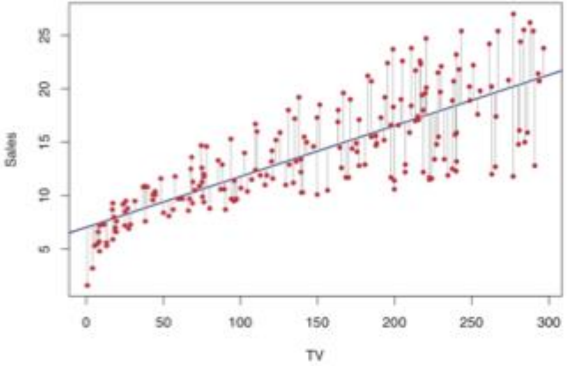


Image Credit: Introduction to Statistical Learning with Applications in R, 7th Edition, Chapter 3 – Linear Regression

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

- How do we measure the best line?
- There are two options:

<u>Option 1:</u> Minimize the sum of magnitudes (absolute values) of the residuals

$$|e_1| + |e_2| + |e_3| + \ldots + |e_n|$$

OR

<u>Option 2:</u> Minimize the sum of squared residuals

$$\text{RSS} = e_1 + e_2 + e_3 + \ldots + e_n$$

$ei = yi - \hat{y}i$ represents the *ith* residual

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

Outliers in Regression

- How does an outlier influence the least square line?
- In general, outliers are the points that fall away from the cloud of points.
- Two types –

– Leverage Points：Outliers that fall horizontally away from the center of the cloud of points but don't influence the slope of the regression line are called leverage points.

– Influential Points：Outliers that actually influence the slope of the regression line are called influential points.

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 classes so that similar objects are in the same class and dissimilar objects are in different classes
  - Discover overall distribution patterns and relationships between attributes

- Association Rule Mining

– Initially developed for market basket analysis

– Goal is to discover relationships between attributes

– Uses include decision support, classification and clustering

- Other Types of Mining

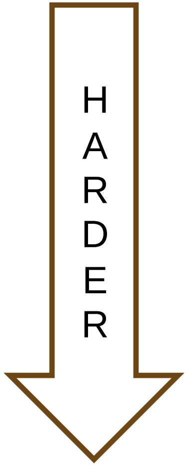
– Outlier Analysis

– 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



HARDER

Analysis – i.e. Science question

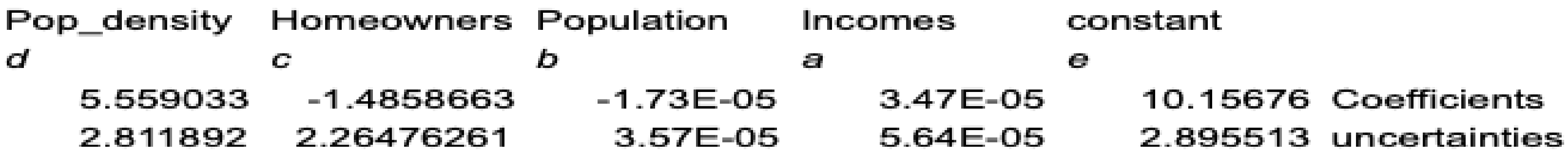
- We want to see if there is a correlation between the percent of the college-educated population and the mean Income, the overall population, the percentage of people who own their own homes, and the population density.

Model:

- %_college = a x Income + b x Population + c x Homeowners/Population + d x Population/area + e

- We solve for for the coefficients *a* through *e*.

- This can be done with Excel with the LINEST function, giving the result:



Pop_density	Homeowners	Population	Incomes	constant	
d	*c*	*b*	*a*	*e*	
5.559033	-1.4858663	-1.73E-05	3.47E-05	10.15676	Coefficients
2.811892	2.26476261	3.57E-05	5.64E-05	2.895513	uncertainties

  - Revealing that population density correlates with college-educated percentage at a significant level.

  - => college-educated people prefer to live in densely populated cities.

Thanks!

Work on your data collection!!