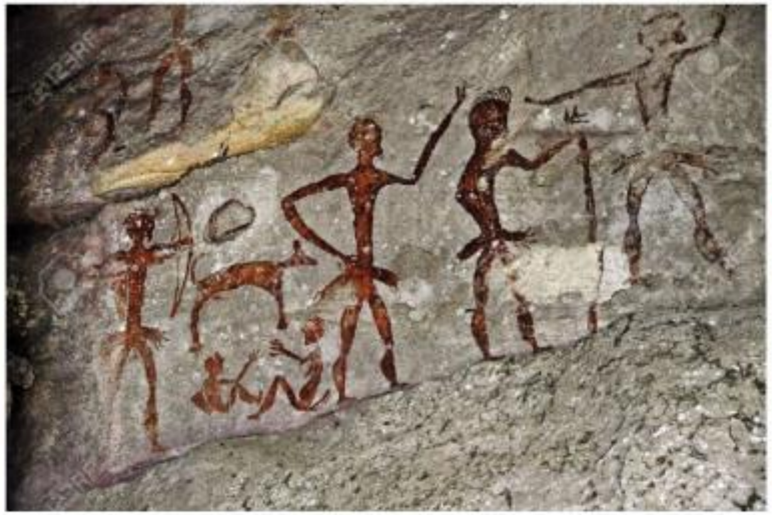
Introduction to
Data Manipulation &
Visualization

Visualization Objectives

- Record information
- Analyze data to support reasoning
- Confirm hypotheses
- Communicate ideas to others

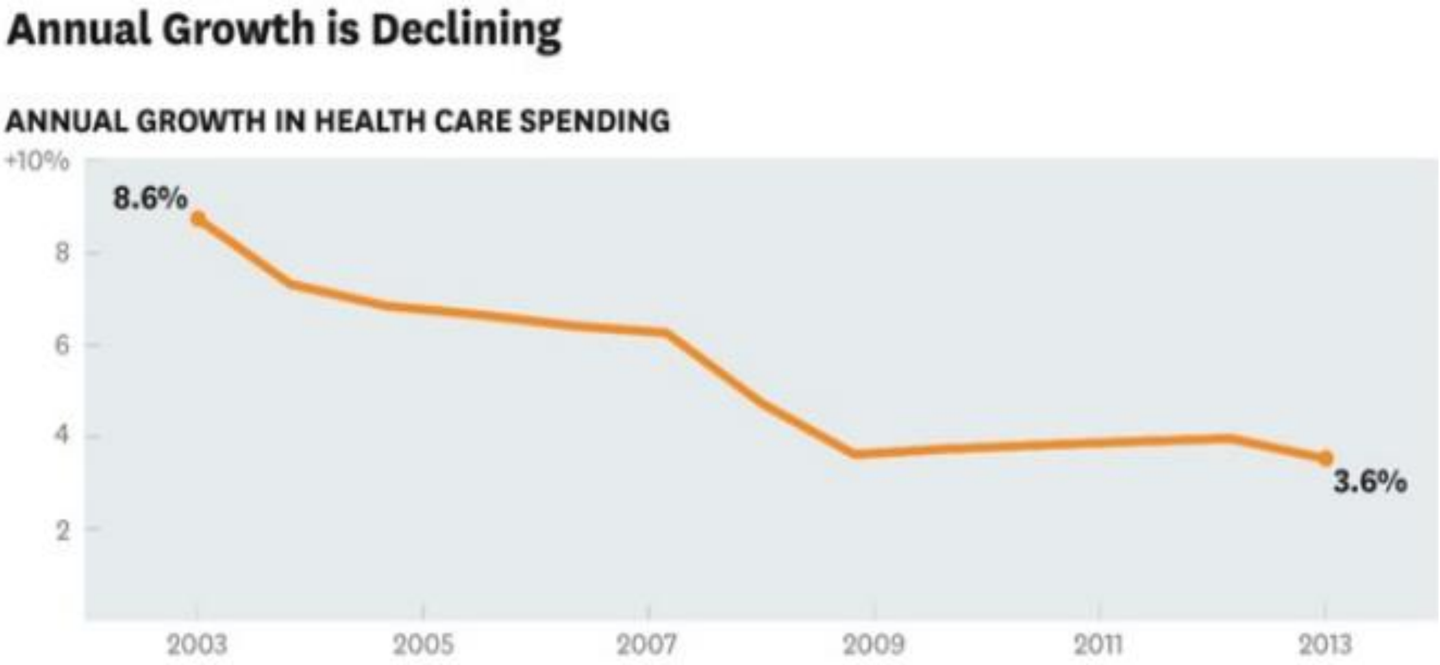
Why Visualize?

To record information



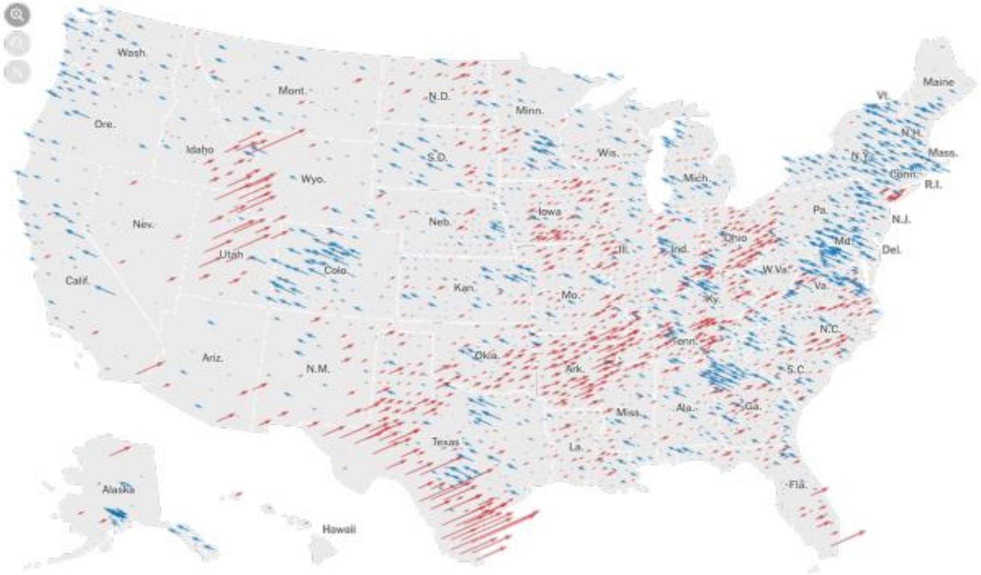
Why Visualize?

To communicate information



Why Visualize?

To analyze data



2020 US Elections (NYTimes)

Types of Plots

- Line plots
- Bar plots
- Scatter plots
- Box plots
- Histograms

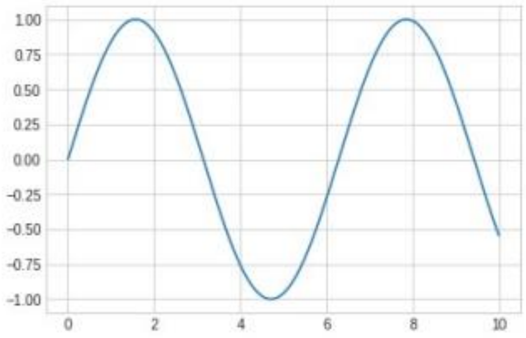
What are line plots?

Two types of relational plots:

1) **Scatter plots**
   - Each plot point is an independent observation

2) **Line plots**
   - Each plot point represents the same "thing", typically tracked over time

Line plot



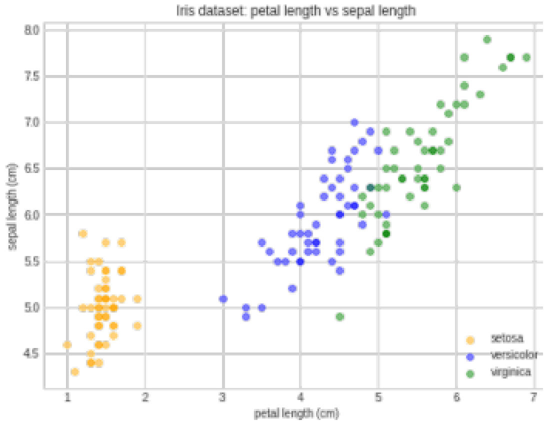
- Used for numeric data
- **Used to show trends**
- Compare two or more different variables over time
- Could be used to make predictions

View trends in data over time.

Examples: Stock price change over a five-year period or website page views during a month.

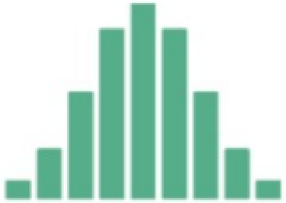
Scatter plots

- Investigate relationships between quantitative values.
  - Used to visualize relation between two numeric variables
  - Used to visualize correlation in a large data set



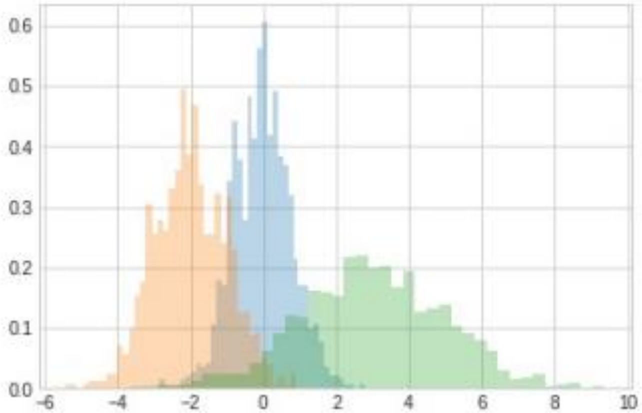
Histograms

- Understand the distribution of your data.
    - Displays the frequency distribution (shape)
- Summarize large data sets graphically
- Compare multiple distributions



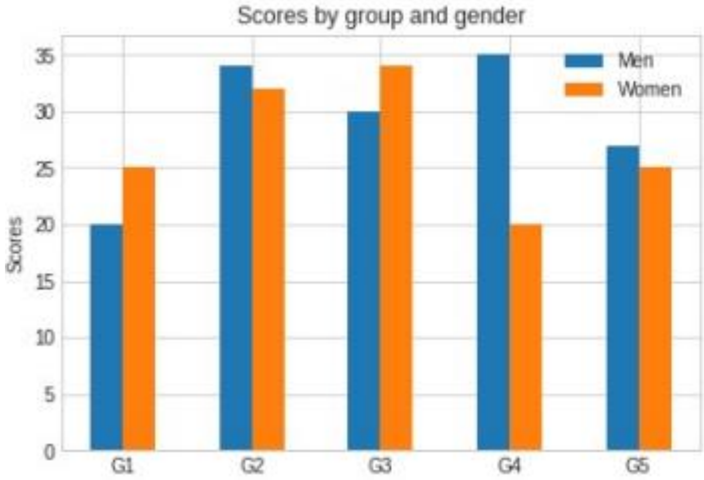
Examples:

- Number of customers by company size,
- Student performance on an exam,
- Frequency of a product defect.



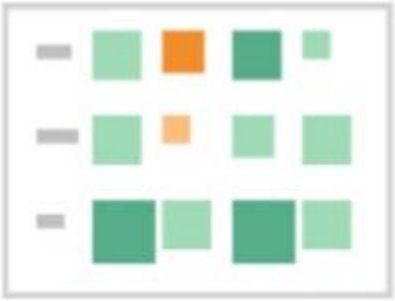
Bar plots

Compare data amongst different categories



Examples: percent of spending by department.

Heat Map



— Show the relationship between at least two factors.

Data Visualization

Part 1

Basic Visuals | Matplotlib, Seaborn

Basic Visualization Concepts, Introduction and Comparison b/t Matplotlib and Seaborn Python Libraries in Jupyter Notebook.

Part 2

Interactive Visuals | Plotly, Bokeh, Tableau, etc.

Deeper insights into more interactive and fun data visualization functions. Introduction to Plotly, Bokeh and Tableau.

Data Visualization-cont’d

What is data visualization?
Data visualization is the graphical representation of information and data.

What makes for effective data visualization?
Visualization **transforms data into images effectively and accurately** represent information about the data.

What are the advantages of data visualization?
Makes for easier **interpretation of patterns and trends** as opposed to looking at data in a tabular/spreadsheet format.

About Data Visualization

What Would You Like to Show?

- Relationships between variables
- Composition of the data over time
- Distribution of variable(s) in data
- Comparison of data with relation to time, variables, categories, etc.

Now, what you'll learn

- How do you choose an appropriate plot?
- How do you interpret common types of plots?
- What are best practices for drawing plots?

Continuous and categorical variables

- Continuous: usually numbers
  - heights, temperatures, revenues

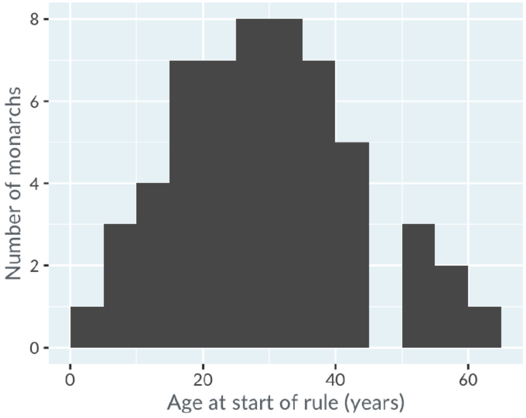
- Categorical: usually text
  - eye colors, countries, industry

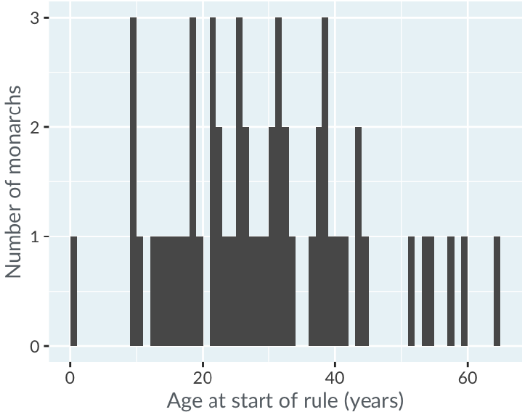
- Can be either
  - age is continuous, but age group is categorical
  - time is continuous, month of year is categorical

Histograms

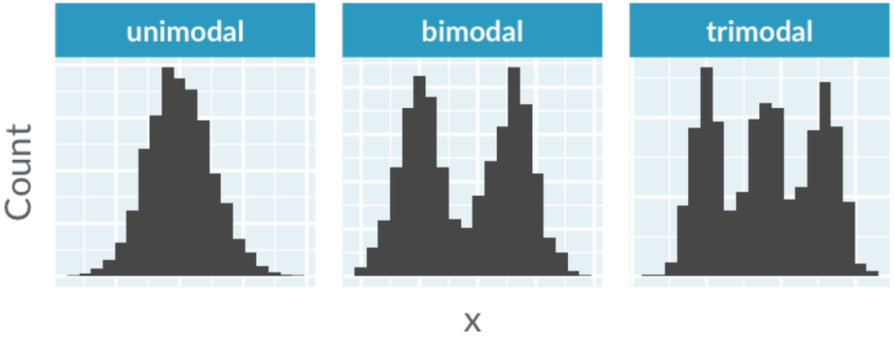
When should you use a histogram?

1) If you have continuous variable(s).
2) You want to ask questions about the shape of its distribution.

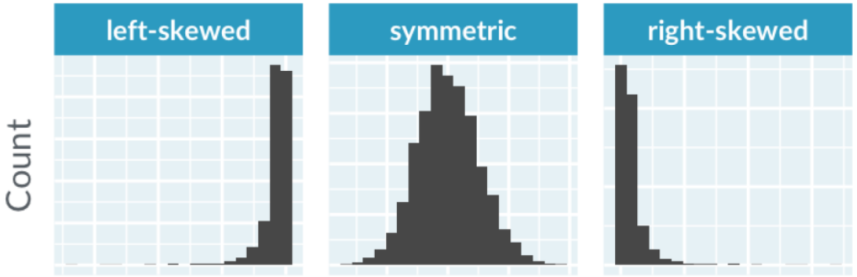




Modality: how many peaks?



Skewness: is it symmetric?

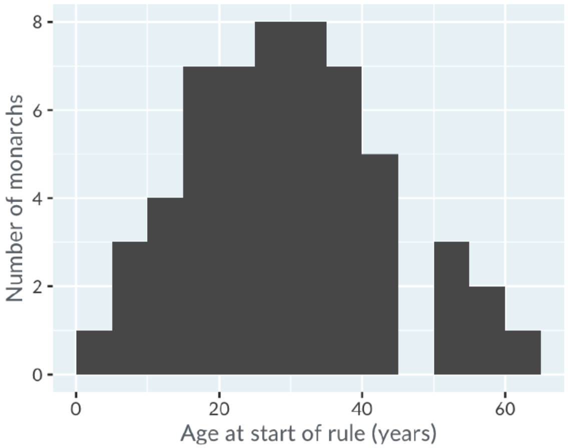


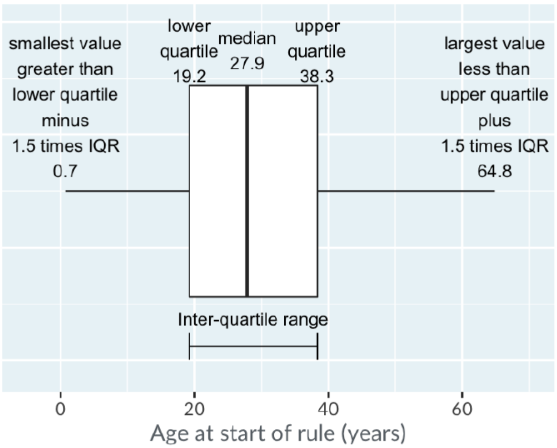
Box plots

When should you use a box plot?

1) When you have a continuous variable, split by a categorical variable.
2) When you want to compare the distributions of the continuous variable for each category.

Histogram vs. box plot

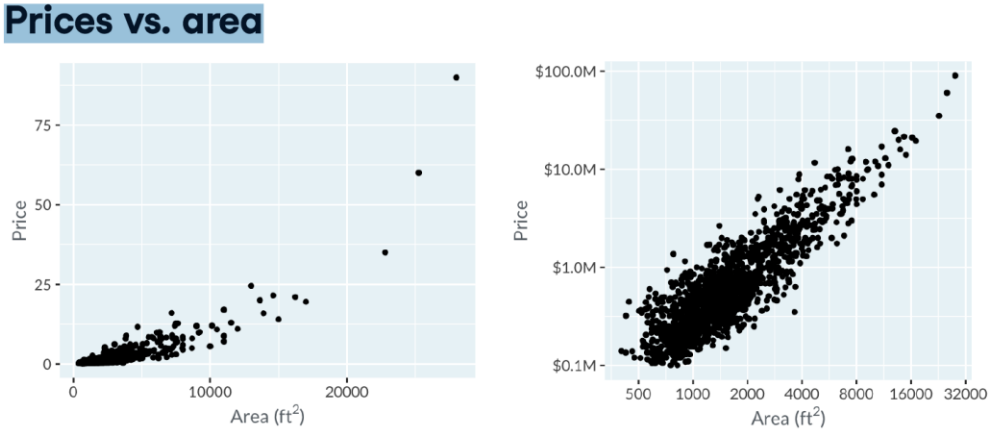




Scatter plots

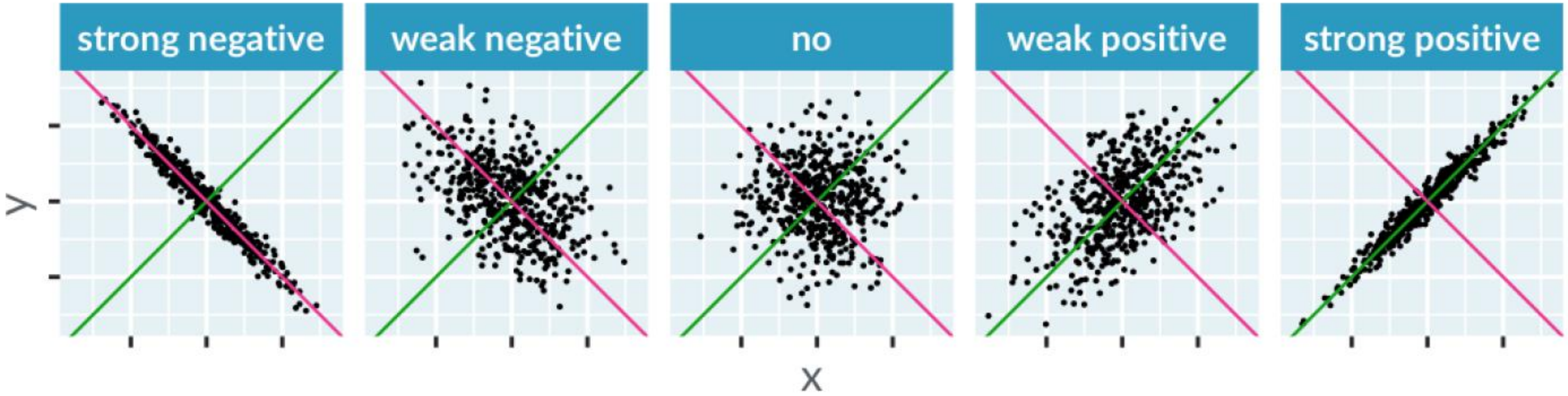
When should you use a scatter plot?

1) You have two continuous variables.
2) You want to answer questions about the relationship between the two variables.

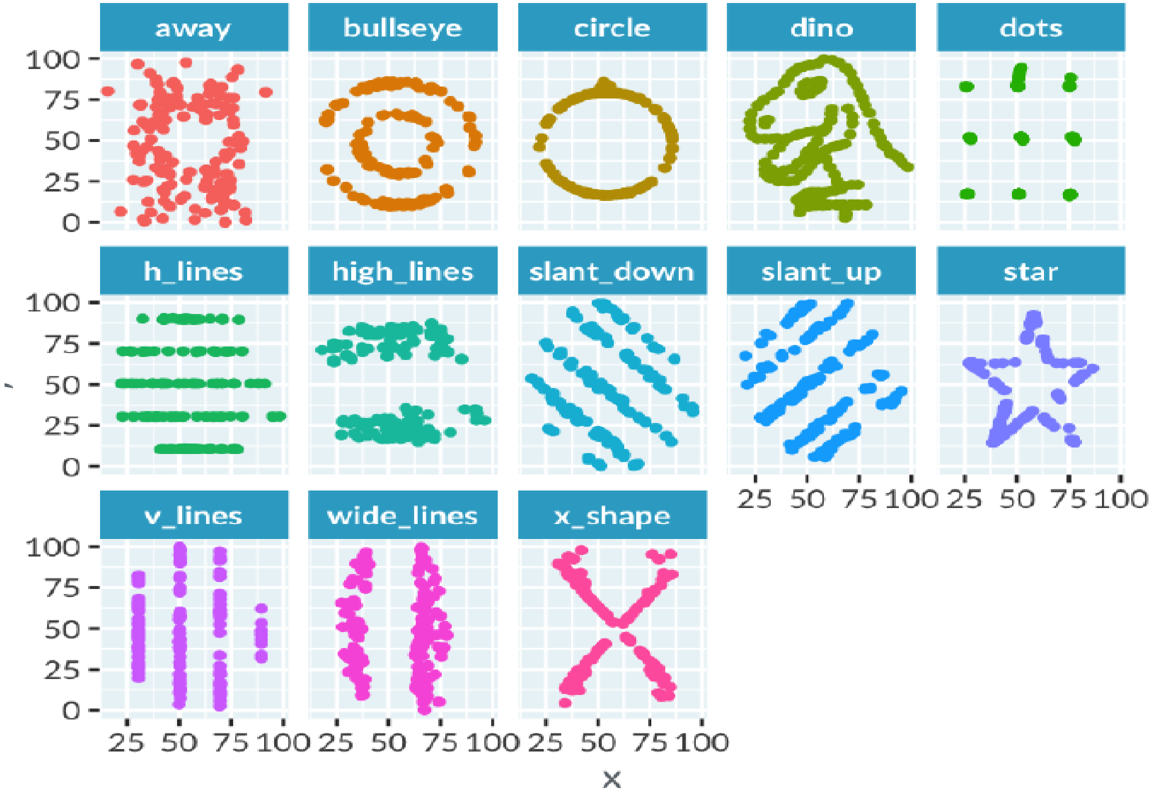


Correlation

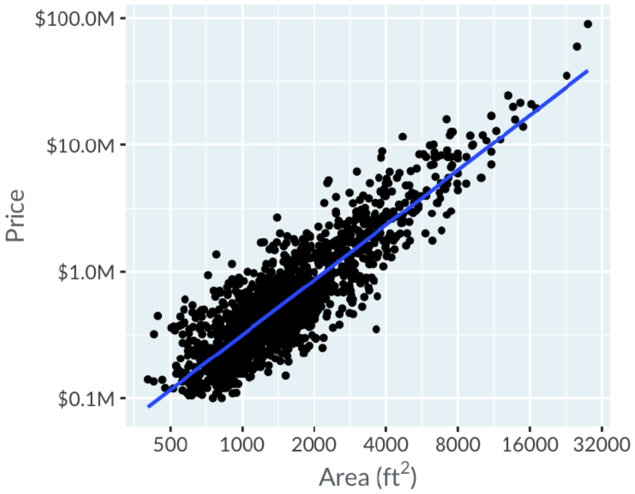
Straight line through the points?



Sometimes correlation isn't helpful



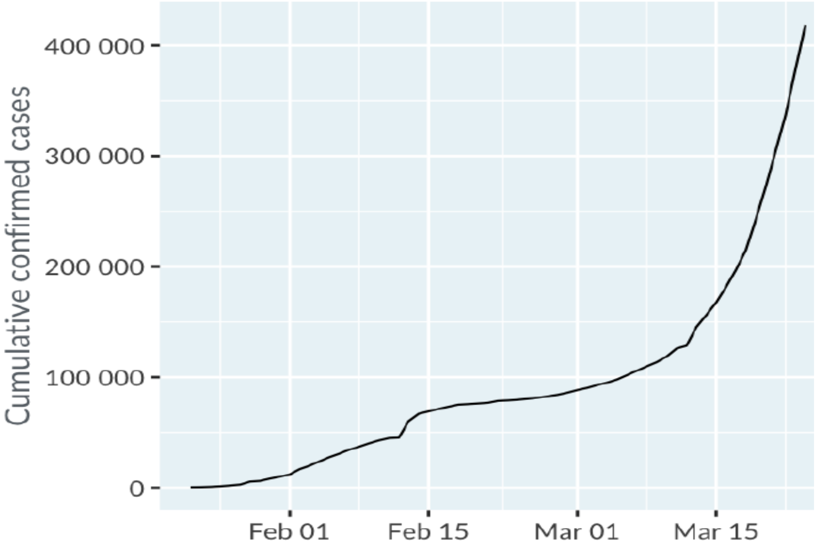
Adding trend lines



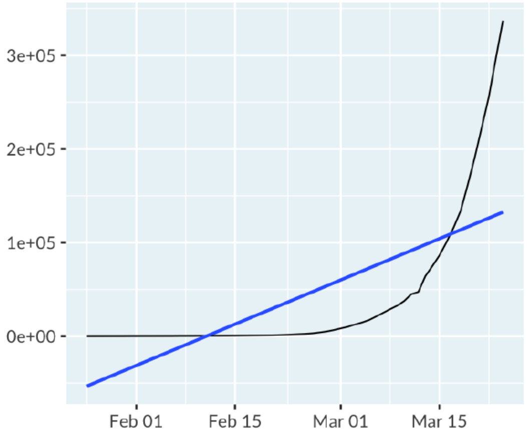
Line plots

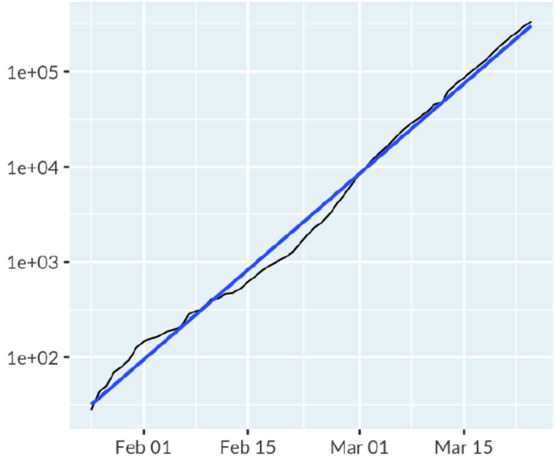
When should you use a line plot?

1. You have two continuous variables.
2. Consecutive observations are connected somehow. Usually, but not always, the x-axis is dates or times.



Trend lines + log scale



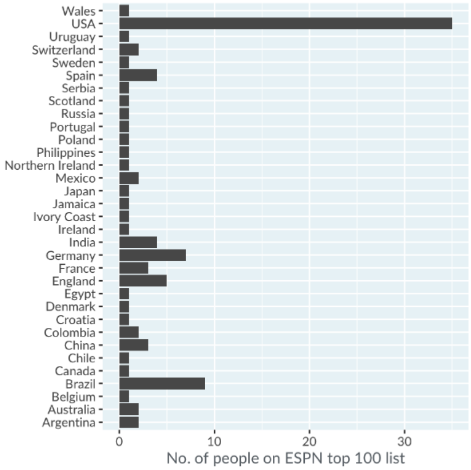


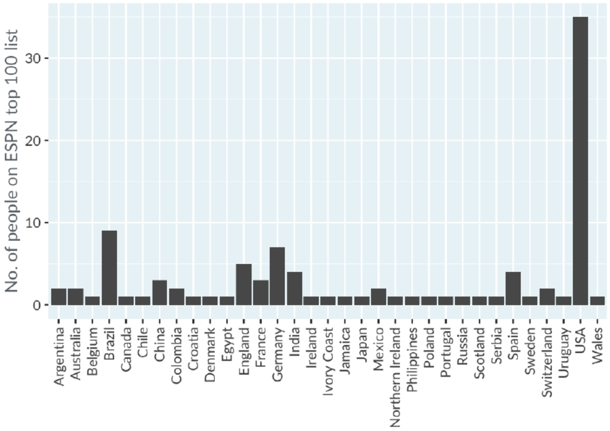
Bar plots:

When should you use a bar plot?

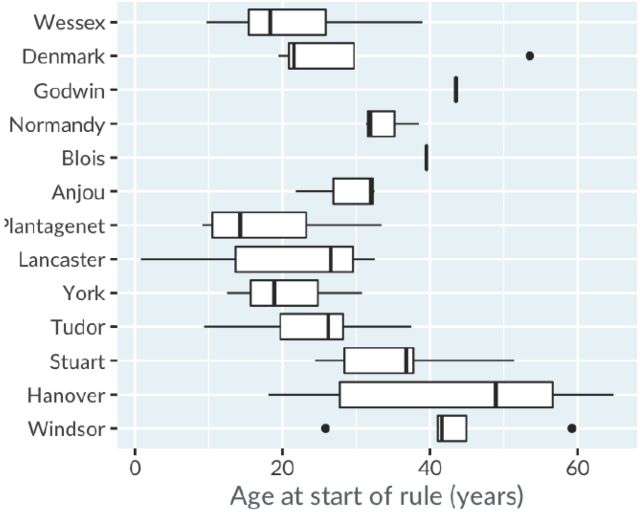
Most common cases:

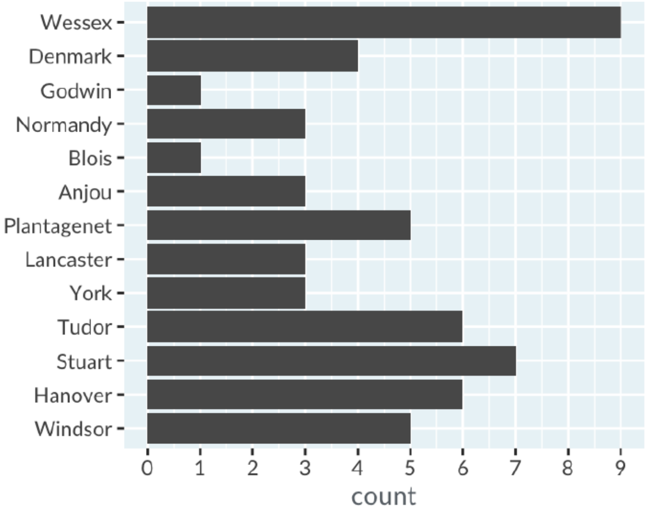
1) You have a categorical variable.
2) You want counts or percentages for each category.





Bar plots vs. box plots

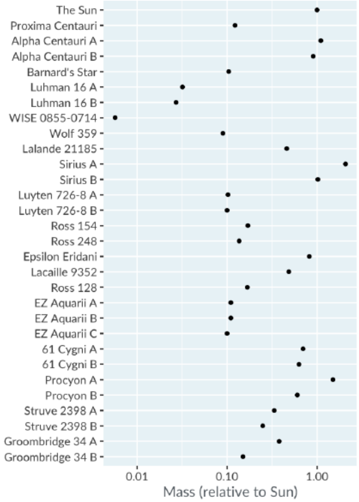




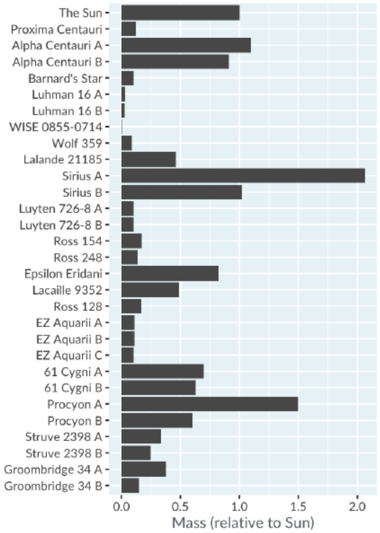
Dot plots

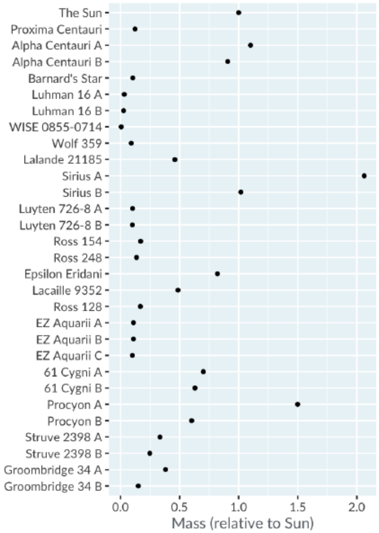
When should you use a dot plot?

1) You have a categorical variable.
2) You want to display numeric scores for each category on a log scale, or
3) You want to display multiple numeric scores for each category

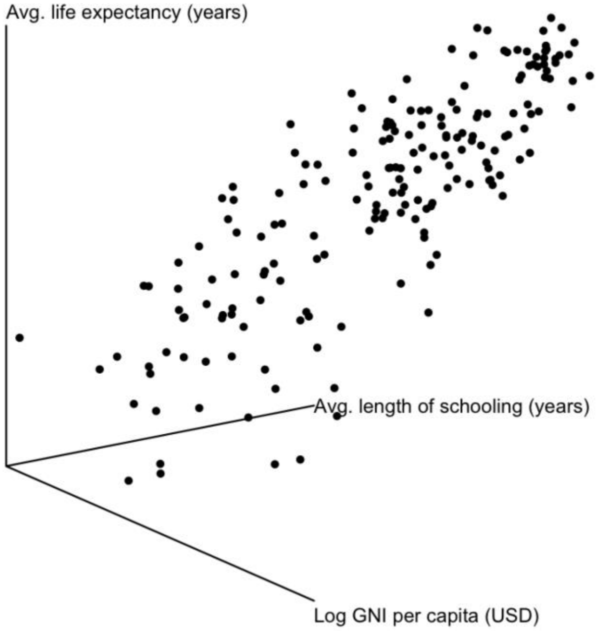


Bar plot vs. dot plot





Higher dimensions: 3D

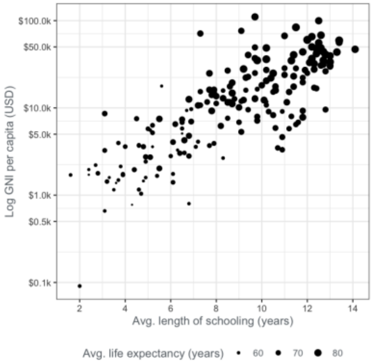


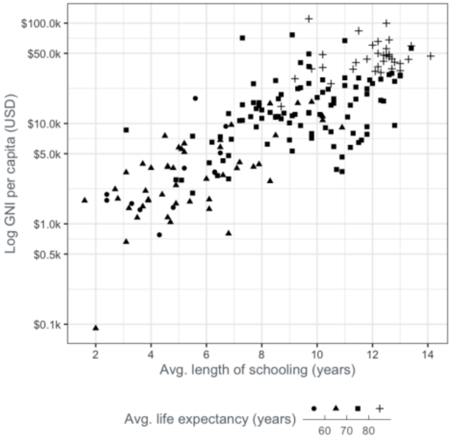
3D scatter plots

x and y are not the only dimensions

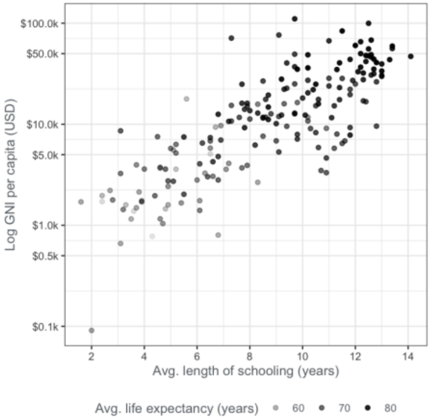
- Points also have these dimensions
  - color
  - size
  - transparency
  - shape

Size

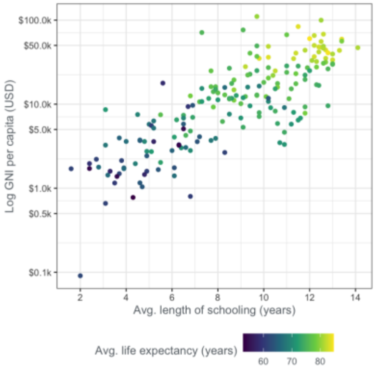




Shape



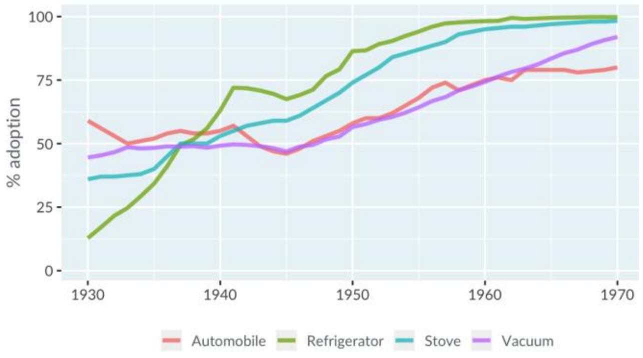
Transparency



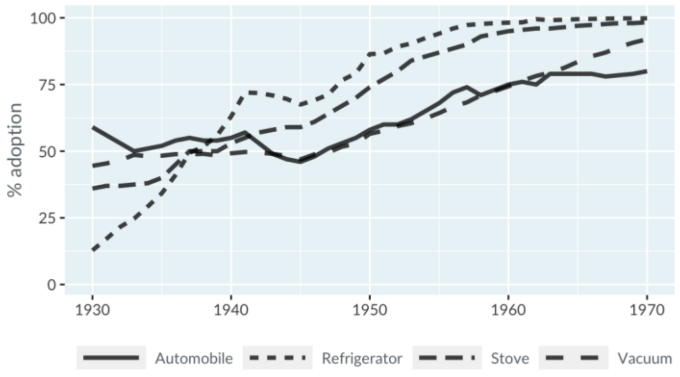
Color

Other dimensions for line plots

- color
- thickness
- Transparency
- Line type (solid, dashes, dots)



Color

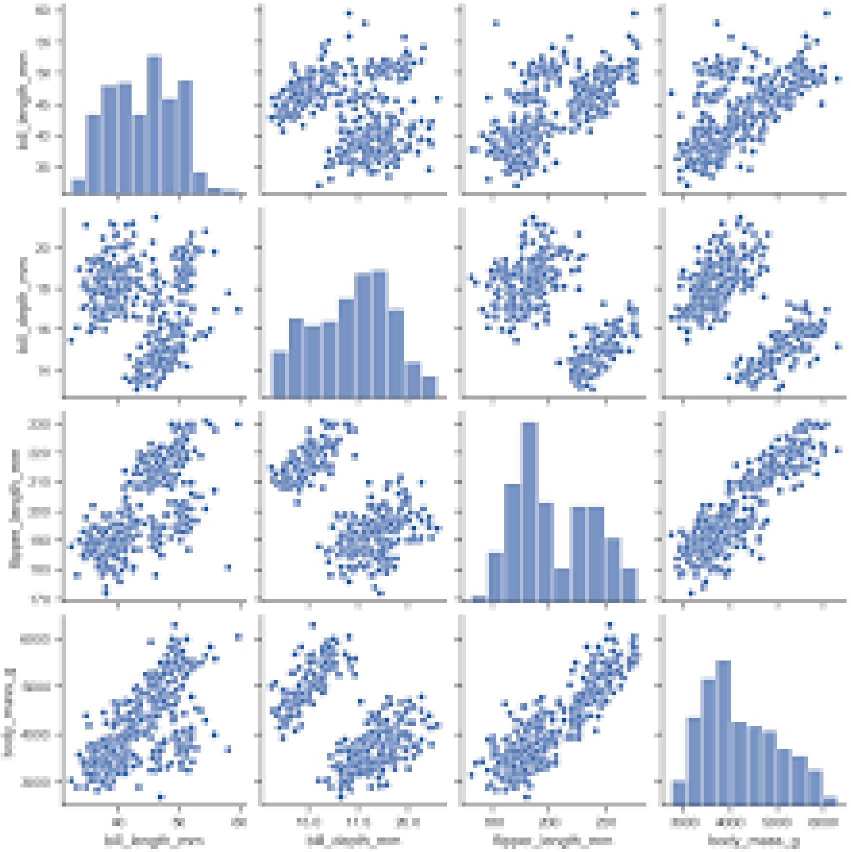


Linetype

Plotting many variables at once

When should you use a pair plot?

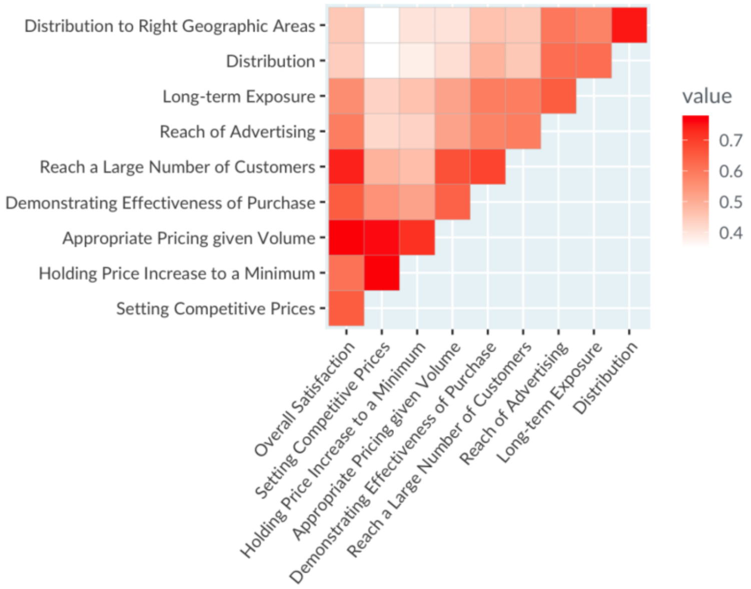
- You have a set of variables (either continuous, categorical, or a mix).
- You want to see the distribution for each variable.
- You want to see the relationship between each pair of variables.



Correlation heatmap

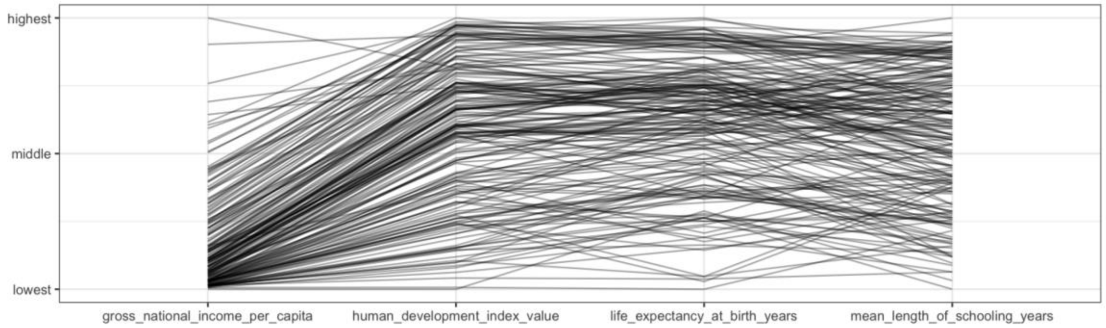
When should you use a correlation heatmap?

- You have lots of **continuous variables**.
- You want to a simple overview of how each pair of variables is related

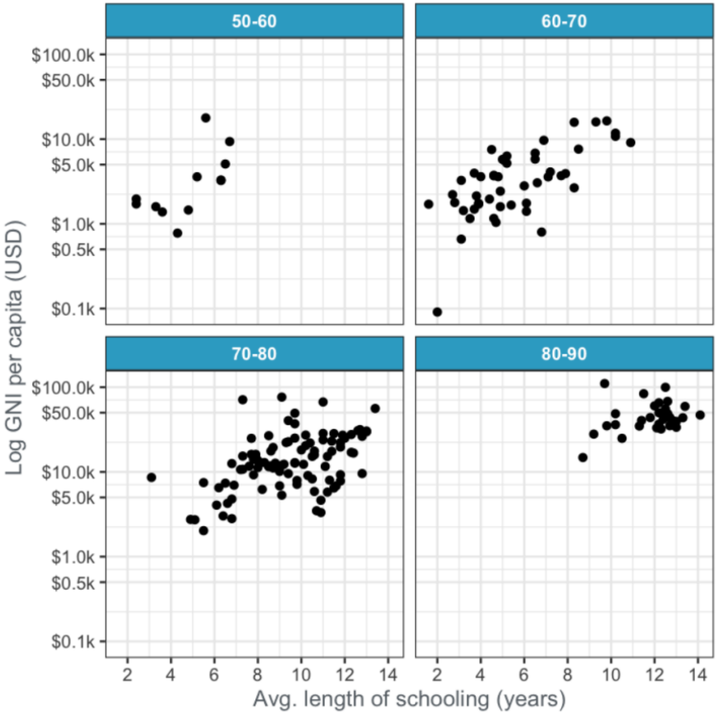


When should you use a parallel coordinates plot?

- You have lots of continuous variables.
- You want to find patterns across these variables, or
- You want to visualize clusters of observations.



Lots of panels

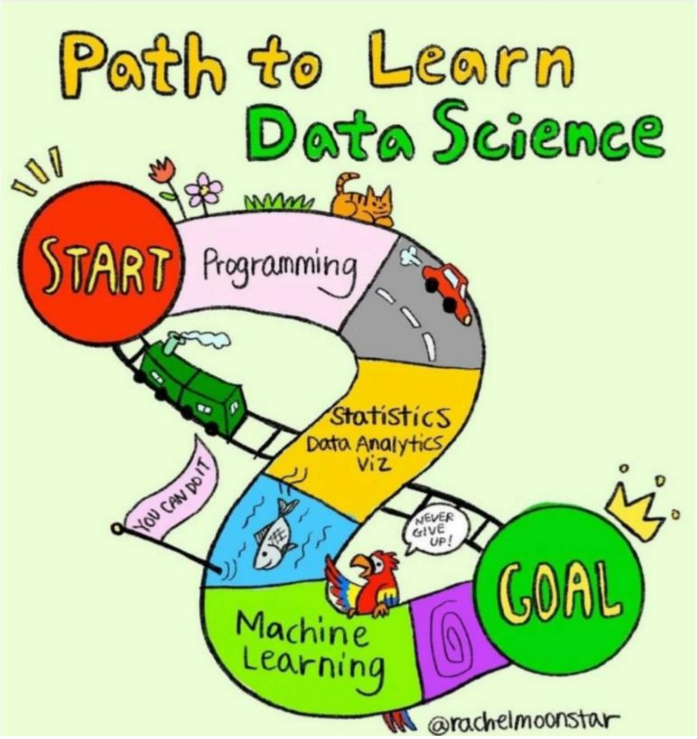


Summary

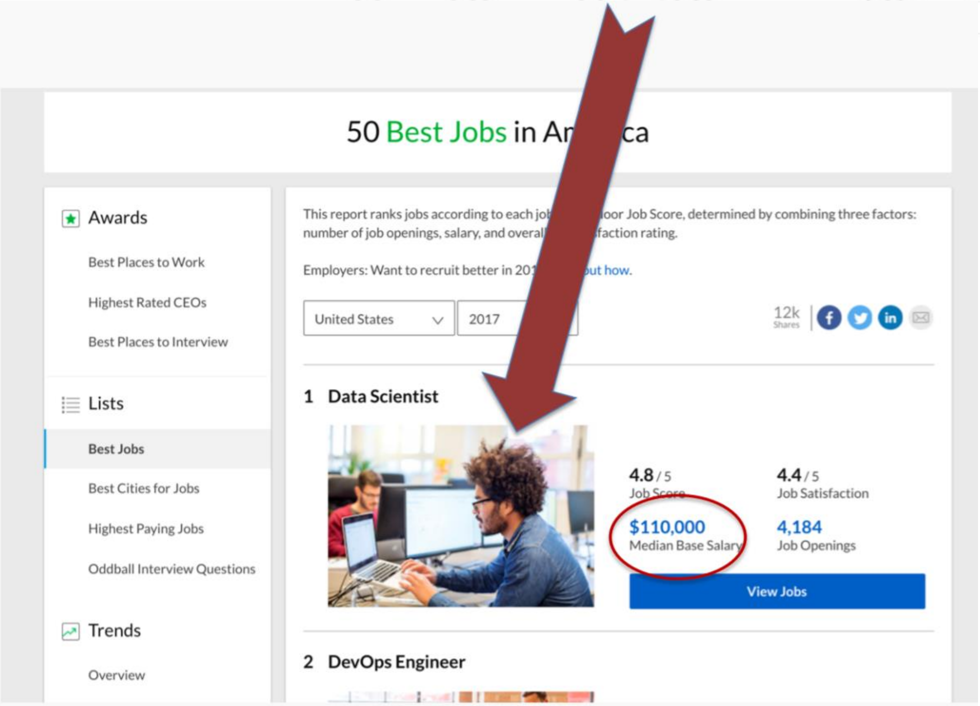
- Histograms: show a distribution
- Scatter plots: compare two numeric variables
- Line plots: show trends over time
- Bar plots: show counts by category
- Pair plot: compare many variables
- Correlation heatmap: show related variables
- Parallel coordinates plot: find patterns across variables

DATA MANIPULATION & VISUALIZATION

Data Science



Jobs!



50 Best Jobs in Am ca

This report ranks jobs according to each job oor Job Score, determined by combining three factors: number of job openings, salary, and overal faction rating.

Employers: Want to recruit better in 201 out how.

United States ⌵ 2017

12k Shares

1 Data Scientist

4.8 / 5 Job Score

4.4 / 5 Job Satisfaction

$110,000 Median Base Salary

4,184 Job Openings

View Jobs

2 DevOps Engineer

Data versus Information

- Data: Any observation collected in respect of any characteristic or event is called data.

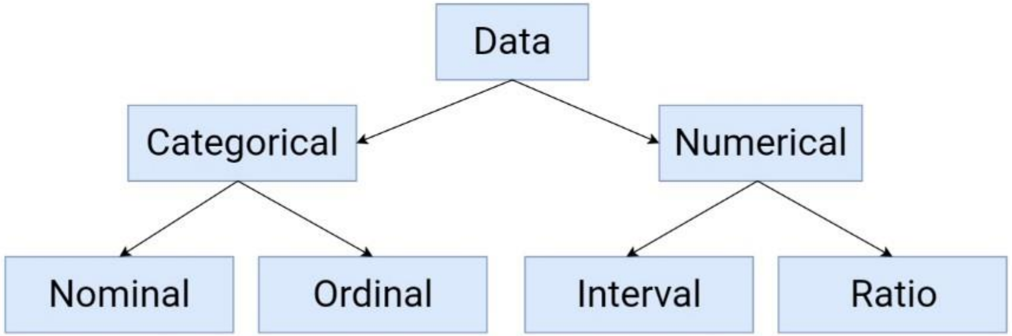
- Information
    - Raw data carry/convey little meaning, when it is considered alone.
    - The data is minimized: processed/analyzed and then presented systematically.
        - It is converted into Information.

- Data, that is not converted into information is of little value for evaluation and planning and cannot be used by those who are involved in decision making.

Data Classification

Classification data can be divided into two types

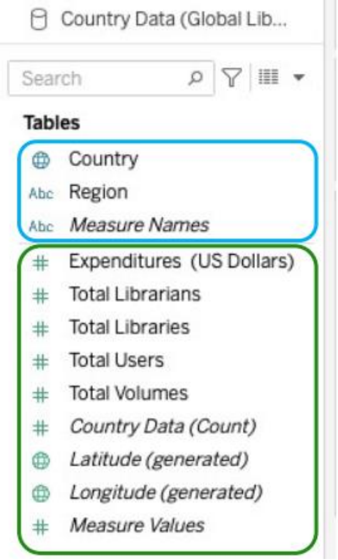
- Quantitative data (numerical);
- Qualitative data (descriptive, categorical/frequency count).



Data Field

A field, also known as a column, is a single piece of information from a record in a data set.

- Qualitative Field (Dimensions)
  - Describes or Categorizes Data
    - What, when or who
- Quantitative Field (Measures)
  - Numerical Data
  - Provides measurement for qualitative category
  - Can be used in calculations



Country Data (Global Lib...

Tables

Country
Abc Region
Abc *Measure Names*

Expenditures (US Dollars)
Total Librarians
Total Libraries
Total Users
Total Volumes
Country Data (Count)
Latitude (generated)
Longitude (generated)
Measure Values

Dimensions

Measures

Quantitative Data vs Qualitative Data Field

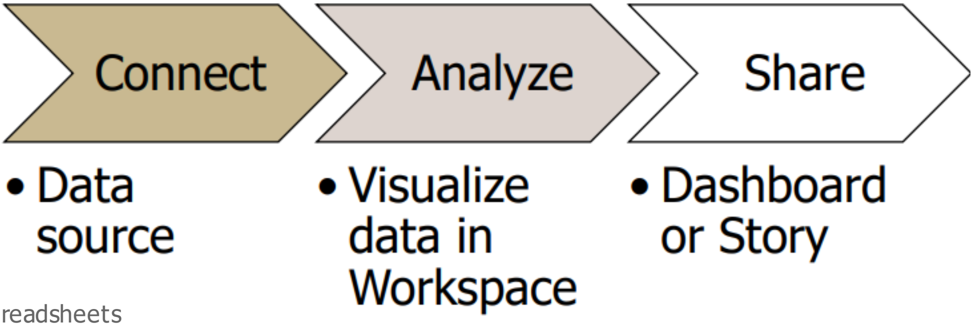
Quantitative Data has two types:

- (a) **Discrete**: Discrete variables can take only certain values.
- (b) **Continuous**: Continuous variables may take any value (typically between certain limits).

Qualitative Data is also called descriptive/ categorical data/ frequency count:

- When the data are arranged in categories on the basis of their quality and there is gap between two values,
- Qualitative data is initially expressed in non-numerical forms.

Data management



Connect
- Data source

Analyze
- Visualize data in Workspace

Share
- Dashboard or Story

- Spreadsheets
  - Excel or csv file
- Relational Databases
  - MySQL or Oracle
- Cloud Data
  - • AWS or Microsoft Azure
- Other Sources

Data quality

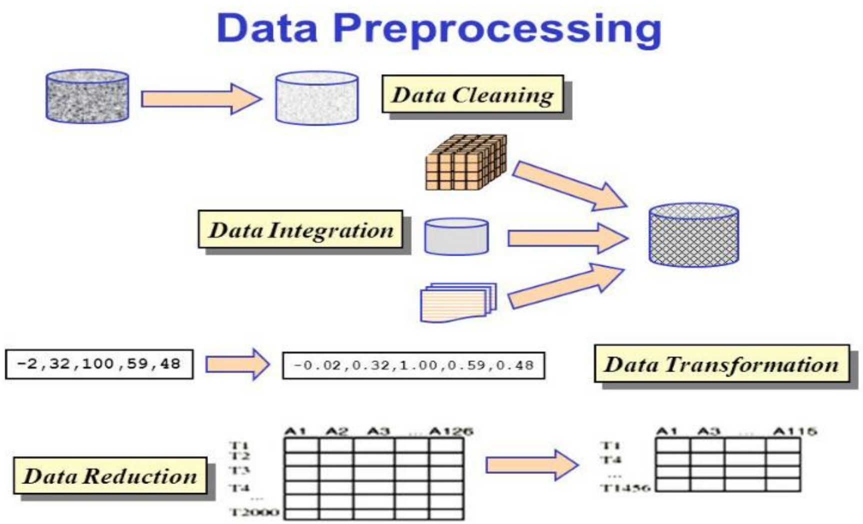
Data in the real world is dirty

- **Incomplete**: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data,
- **Noisy**: containing errors or outliers, e.g., Salary="-10"
- **Inconsistent**: containing discrepancies in codes or names
  - e.g., Age="42" Birthday="03/07/1997"
  - e.g., Was rating "1,2,3", now rating "A, B, C"
  - e.g., discrepancy between duplicate records

Why Data Pre-processing?

Data Preprocessing is a technique uses to convert the raw data into a clean data set.

- Steps are executed to convert the data into a clean data set.
- Data is gathered from different sources (collected in raw format and it is not feasible for the analysis).
- This technique is performed before the execution of Iterative Analysis.
- Steps of data-preprocessing:
  - Data Cleaning
  - Data Integration
  - Data Transformation
  - Data Reduction



Data Cleaning

Data quality is a main issue and occurs anywhere in information systems.

These problems can be solved by Data Cleaning:

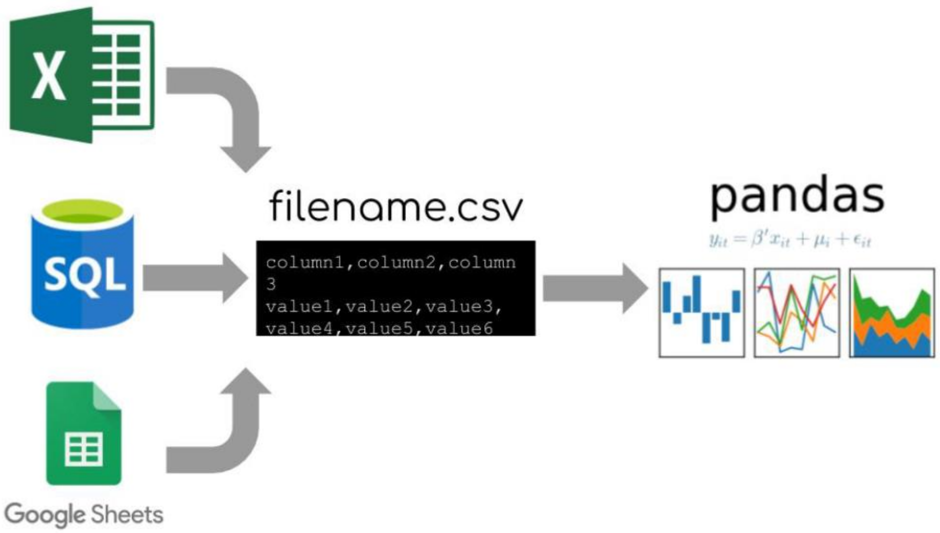
- is a process used to determine inaccurate, incomplete or unreasonable data
- and then improve the quality through correcting of detected errors
- => reduces errors and improves the data quality.

Data Cleaning can be a time consuming and tedious process, but it cannot be ignored.

Data quality criteria :

- accuracy, integrity, completeness, validity, consistency, uniqueness.

CSV files

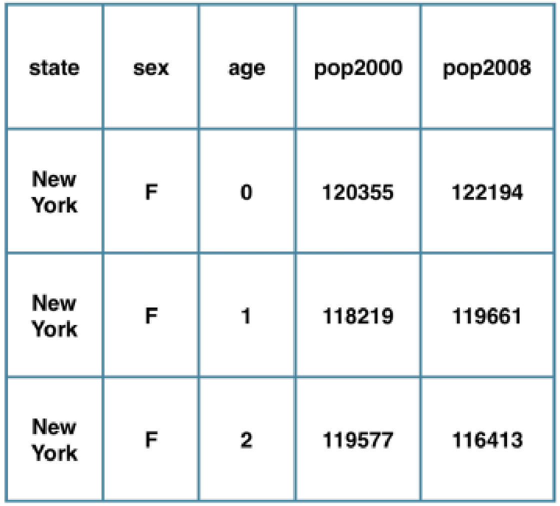


```
column1,column2,column
3
value1,value2,value3,
value4,value5,value6
```

Introduction to Databases

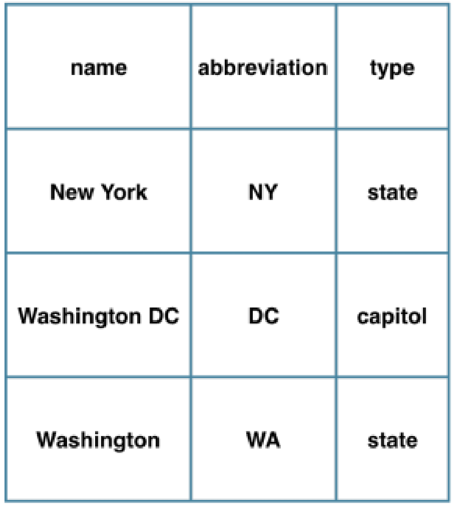
A database consists of tables

Census



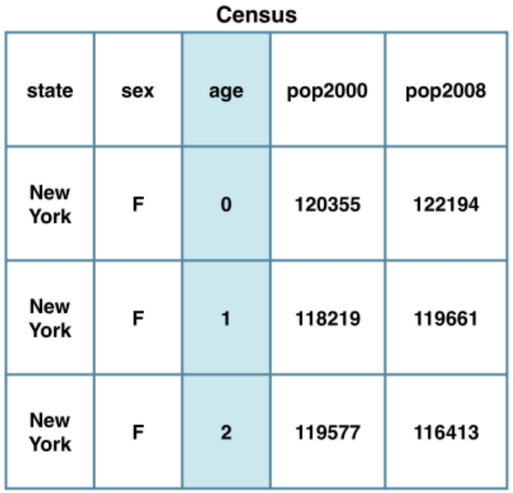
state	sex	age	pop2000	pop2008
New York	F	0	120355	122194
New York	F	1	118219	119661
New York	F	2	119577	116413

State_Fact



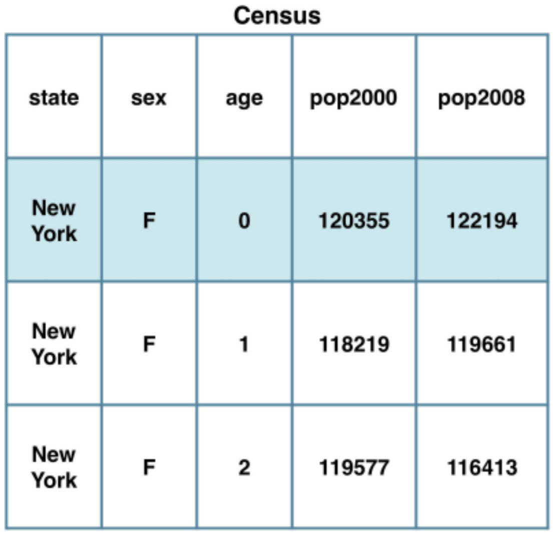
name	abbreviation	type
New York	NY	state
Washington DC	DC	capitol
Washington	WA	state

Table consist of columns and rows



Census

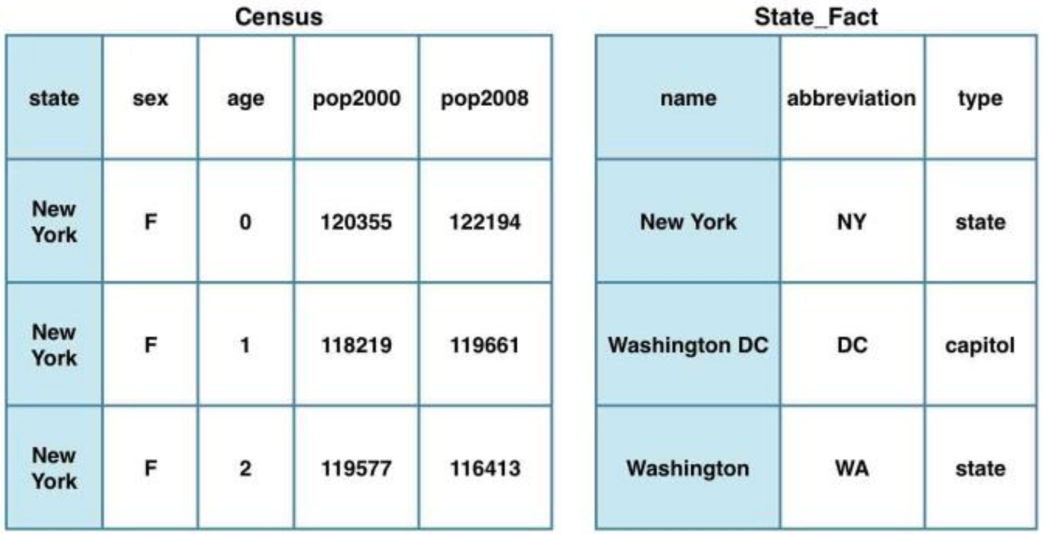
state	sex	age	pop2000	pop2008
New York	F	0	120355	122194
New York	F	1	118219	119661
New York	F	2	119577	116413



Census

state	sex	age	pop2000	pop2008
New York	F	0	120355	122194
New York	F	1	118219	119661
New York	F	2	119577	116413

Tables can be related



Census

state	sex	age	pop2000	pop2008
New York	F	0	120355	122194
New York	F	1	118219	119661
New York	F	2	119577	116413

State_Fact

name	abbreviation	type
New York	NY	state
Washington DC	DC	capitol
Washington	WA	state

Useful Python Libraries for Data visualization

NumPy pandas matplotlib seaborn bokeh

Matplotlib:

- Provides the **building blocks** for seaborn’s and pandas visualizations
- It can also be used on its own to plot data

Pandas

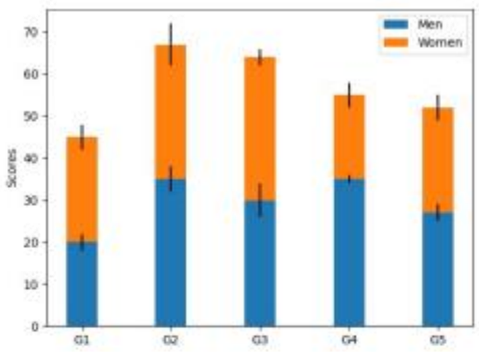
- It is a foundational library for **analyzing data**
- It also supports **basic plotting** capability

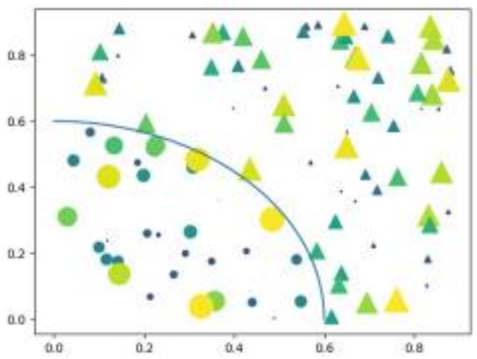
Seaborn

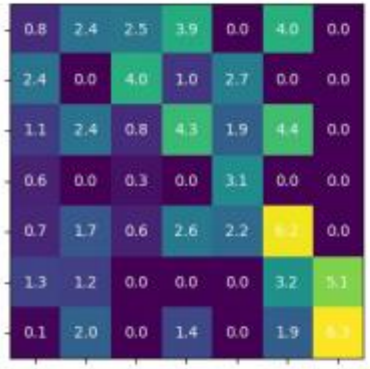
- Seaborn supports **complex visualizations** of data
- It is built on matplotlib and works best with pandas' dataframes

Matplotlib

- Used for basic plotting
- Highly customizable
- Works with NumPy and pandas







About Matplotlib:

- Matplotlib is a comprehensive library for creating static visualizations in Python.
- Usage: Matplotlib/Pandas is mostly used for quick plotting of Pandas DataFrames and time series analysis.

Advantages of Matplotlib:

- Easy to setup and use.
- Very customizable.

Llimitations of Matplotlib:

- Visual presentation tends to be simple compared to other tools.

About Seaborn:

- Seaborn is a Python data visualization library based on Matplotlib.
- It provides a **high-level interface for drawing attractive and informative statistical graphics**.
- Usage: Those who want to create **amplified data visuals, especially in color**.

Seaborn’s Pros and Cons:

- Pro: Includes higher level interfaces and settings than does Matplotlib
- Pro: Relatively simple to use, just like Matplotlib.
- Pro: Easier to use when working with Dataframes.

- Con: Like Matplotlib, data visualization seems to be simpler than other tools.

Bokeh

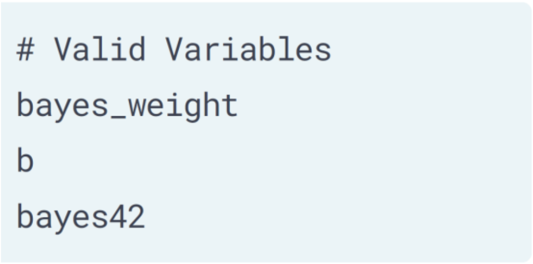
- Bokeh is an interactive visualization Python library.
- Provides elegant and concise construction of versatile graphics.

- Usage: Can be used in Jupyter Notebooks and can provide high-performance interactive charts and plots.

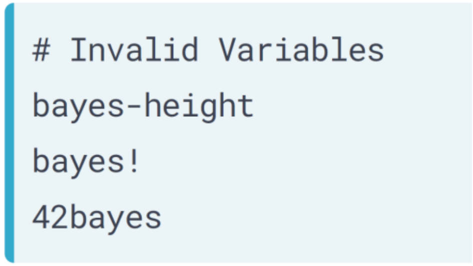
Rules for variable names in Python

Rules for variable names

- Must start with a letter(usually lowercase)
- After first letter, can use letters/numbers/underscores
- No spaces or special characters
- Case sensitive (my_var is different from MY_VAR)



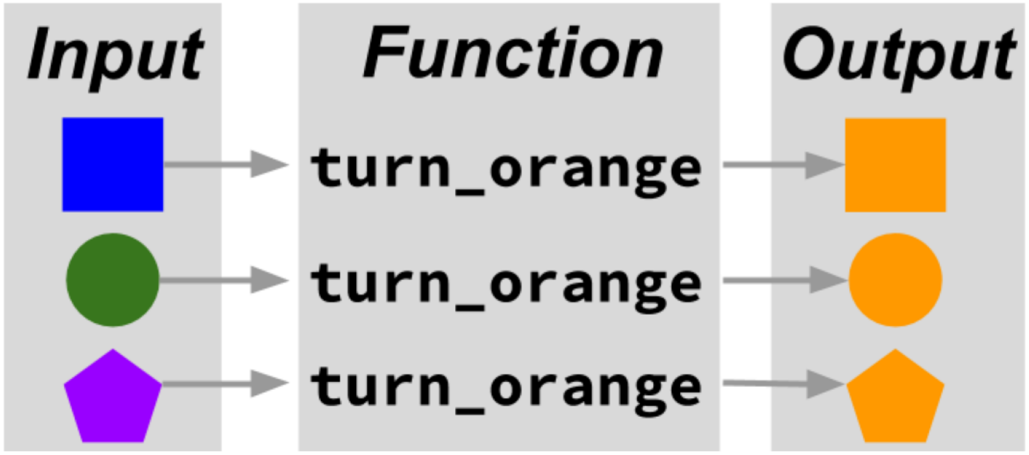
```
# Valid Variables
bayes_weight
b
bayes42
```



```
# Invalid Variables
bayes-height
bayes!
42bayes
```

Function

A function is an action



Input

Function

Output

turn_orange

turn_orange

turn_orange

- Function name is always followed by parentheses ()

NumPy

- Fundamental package for scientific computing
- Exceptionally fast – written in C
- Main data structure:
  - ndarray : n-dimensional arrays of homogeneous data types
- Data manipulation ≈ NumPy array manipulation
- Used in other libraries - Matplotlib, pandas, scikit- learn

Pandas

- **Fundamental tool for handling and analyzing input data**
- Particularly **suited for tabular data**
- Implements powerful data operations
  - Easily read datasets from csv, txt, and other types of files
  - Datasets take the form of DataFrame objects
- Main data structures:
  - DataFrame: A table with rows and columns
  - Series: A single column

pandas Philosophy

There should be one -- and preferably only one -- obvious way to do it.

What's the point of pandas?

pandas is built on NumPy and Matplotlib