Data Science:
Data/Curation and Data Exploration Steps of the End-to-End Process

Won Kim

2022

Roadmap: End-to-End Process

- 1. Objective Setting
- 2. Data Curation
- 3. Data Inspection
- 4. Data Preparation
- 5. Data Analysis
- 6. Evaluation
- 7. Deployment

- Just as software development, steps 1-6 are not strictly sequential, and also they iterate.

(Review) Example Business Needs

- Minimize the financial loss due to use of stolen credit card
- Minimize customer churn
- Increase the effectiveness of targeted online advertising
- Predict the product failure
- Predict employee performance
- …

(Review) The Limited Role of Big Data

- Big Data can provide answers and insights in solving a business problem.
- However, usually it has a limited role in solving the business problem.
- Often business management, investment, etc. are required to solve the business problem.
- Example: to stop customer churning
  - Change pricing policy, develop a new product, acquire a company, etc.

Roadmap: End-to-End Process

- 1. Objective Setting
- 2. Data Curation
- 3. Data Inspection
- 4. Data Preparation
- 5. Data Analysis
- 6. Evaluation
- 7. Deployment

Data Curation

- Data curation tasks
  - Determine the data needed to best meet the business objective.
  - Collect the data.
  - Store the data in the computer.
- Very important but difficult and potentially time-consuming step.
- ** Note: For programming homework in this course, you will simply download available clean datasets from the Internet. That is, you will not experience real data curation, and think, wrongly, that data curation is trivial.

Difficulties in Data Curation

- It is difficult to understand the relationship between data and business needs.
- The data quality has to be acceptable.
- The amount of data has to be enough to be statistically significant. If not, the result of analysis is likely meaningless.
- Often, you do not have the exact data you need, and get it and create it.
- There are lots of data sources to investigate.

Mismatch of Data and BusinessObjective

- A cosmetics company wants to predict if young females will buy a new cosmetics product.
  - But the company only have past sales data for a similar product for old men.
- At the start of a semester, a professor wants to predict which students will receive an A in her course.
  - But she only has students' height and weight data.

Sources of Data

- Free or inexpensive sources
  - Databases and files within the organization
  - Databases and files within partner organizations (e.g., within the supply chain)
  - Internet, SNS
  - Government, public corporations
  - Associations
- Expensive sources
  - (buy data from) data vendors
    - They collect data and sell it (to businesses, individuals, and government)
  - (create new data) surveys, focus group interviews, panels

Large Data Vendors (1/2)

- Acxiom, IRI
  - consumer marketing data
- Datalogix
  - sales data for packaged consumer goods
- Corelogic
  - property and financial data (for lenders, insurers, and landlords
- Equifax, Experian, TransUnion
  - credit reporting agencies
- Ebureau
  - scoring services for fraud detection, credit risk

Large Data Vendors (2/2)

- Nielsen
  - audience data on TV, radio, mobile, online, social media
  - (400,000 households in the Nielsen panel in the US)
- DataSift, GNIP
  - social media activity data
- ID Analytics, Intelius
  - identity verification
- Recorded Future
  - real-time threat information

Data Provided by Acxiom

- demographics, such as age and gender
- home information, such as whether the consumer owns or rents
- motor vehicle information, such as make, model, and insurance renewal
- economic data, including income range and credit card use
- purchase data, types of products purchased, and frequency
- interests and indicators of interest, such as sports, arts and crafts, pet ownership, and other such categories

Data Provided by DataLogix

- thousands of consumable products in categories such as
  - food and beverages
  - clothing and shoes
  - tobacco
  - cleaning products
  - pet care items
  - cosmetics

Data Provided by DataSift, GNIP

- Type of Data
  - how often a topic is being mentioned in social media
  - who's talking, and
  - what they're saying,
- Data Sources
  - Twitter
  - Facebook
  - YouTube
  - Bitly
  - Sina Weibo
  - Intense Debate
  - Yammer…..

Written Homework

- A software company would like to use big data analysis in its hiring of new software engineers.
- It needs to develop a profile (model) of high-performing software engineers, and hire new software engineers that match the profile.
- It may use its own data, and may get data from outside (the Internet, or a data provider).
- 1. What kind of data would be necessary?
- 2. Assuming that an Employee table is one such data, list 3 features/attributes of the table that are relevant to creating the model .

Roadmap: End-to-End Process

- 1. Objective Setting
- 2. Data Curation
- 3. Data Inspection
- 4. Data Preparation
- 5. Data Analysis
- 6. Evaluation
- 7. Deployment

Data Inspection

- Data inspection consists of two tasks.
- Data exploration
  - Explores the data collected to understand the general characteristics of the data
- Suitability check
  - Confirms suitability of the data for the business objective.
- ** Note: Often, people call this step “Data Exploration” or “Exploratory Data Analysis(EDA)” step.
- But we lump the data exploration task and suitability check task into one step and call it the “Data Inspection” step.
- It is because suitability check is important and this is the logical step to include it.

Roadmap: Data Inspection

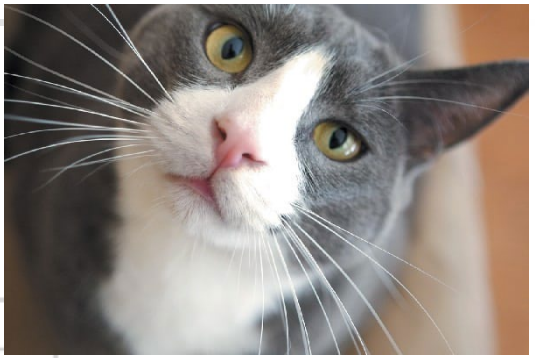
- Data exploration
- Suitability check

Data Exploration

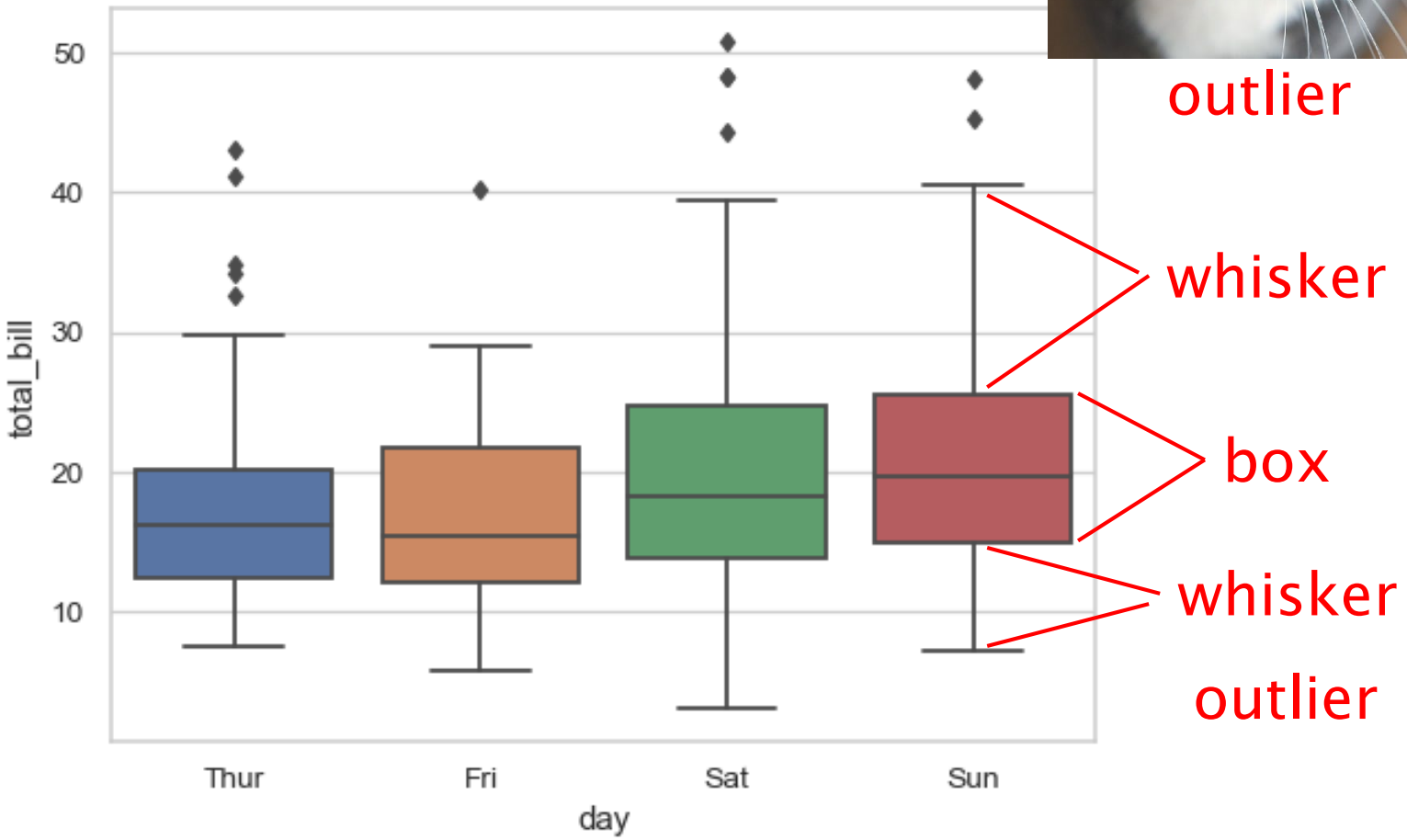
- Check central tendency and dispersion
  - mean, median, range, variance, standard deviation
- Check data distributions
- Check for outliers.
- Check for correlation among attributes.

- Using statistics and data visualization tools
  - Boxplot, histogram, scatterplot,…

Boxplot (or Box and Whisker Plot) (shows min, Q1, median, Q3, max)



whisker



outlier

whisker

box

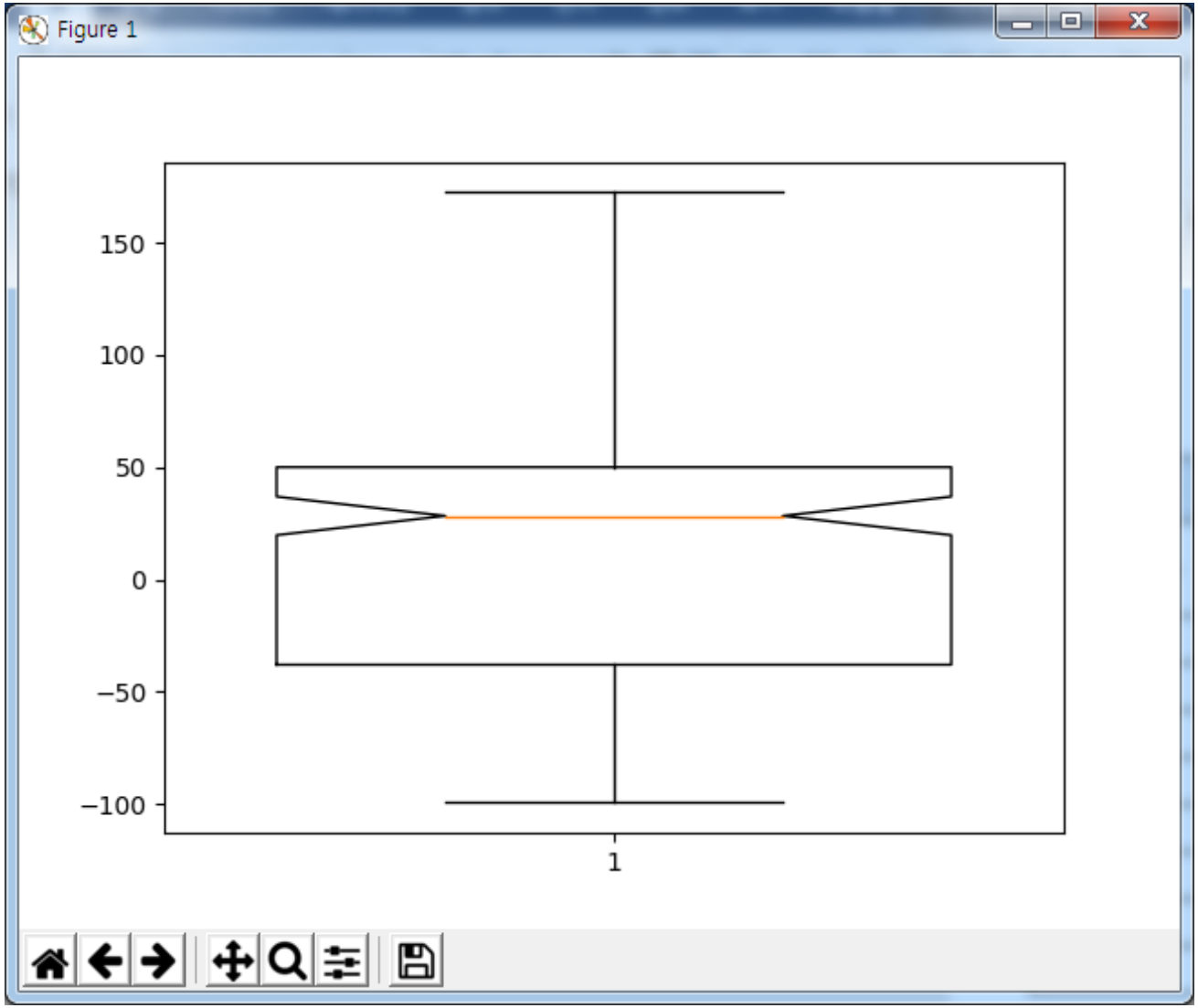
whisker

outlier

MatPlotLib: Depicting groups using boxplots

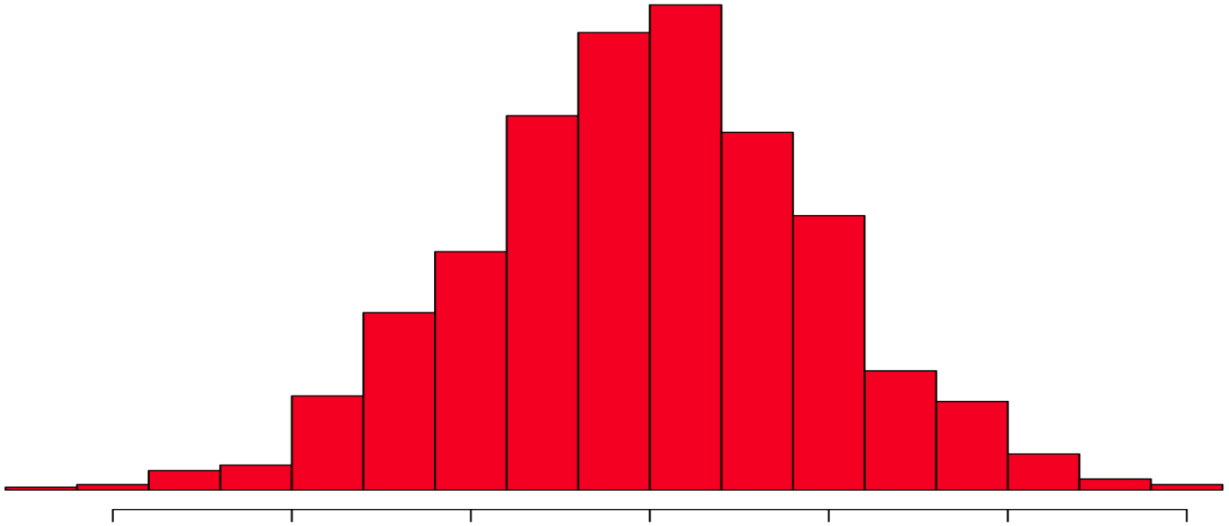
```
import numpy as np
import matplotlib.pyplot as plt

spread = 100 * np.random.rand(100)
center = np.ones(50) * 50
flier_high = 100 * np.random.rand(10) + 100
flier_low = -100 * np.random.rand(100)
data = np.concatenate ( (spread, center, flier_high, flier_low) )

plt.boxplot (data, sym='gx', widths=.75, notch=True)
plt.show()
```

MatPlotLib: boxplots result



Histograms

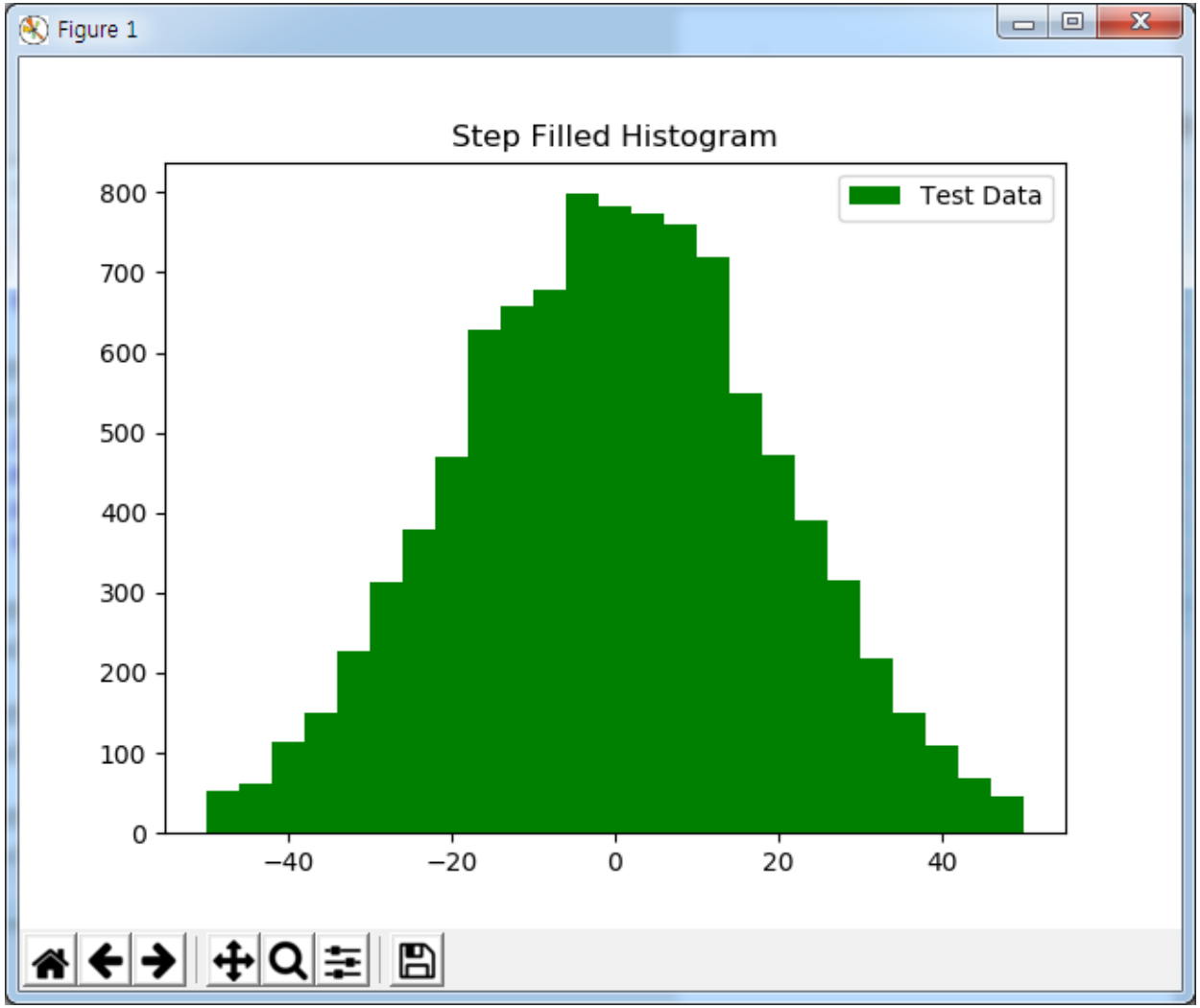
- Represent the *distribution* of data
- Used for continuous data
- Similar to bar graph, but group numbers into ranges (e.g. numbers 1 to 100 into 10 ranges)
- Allow you to visualize the *mean, median, mode, variance,* and *skew* at once!



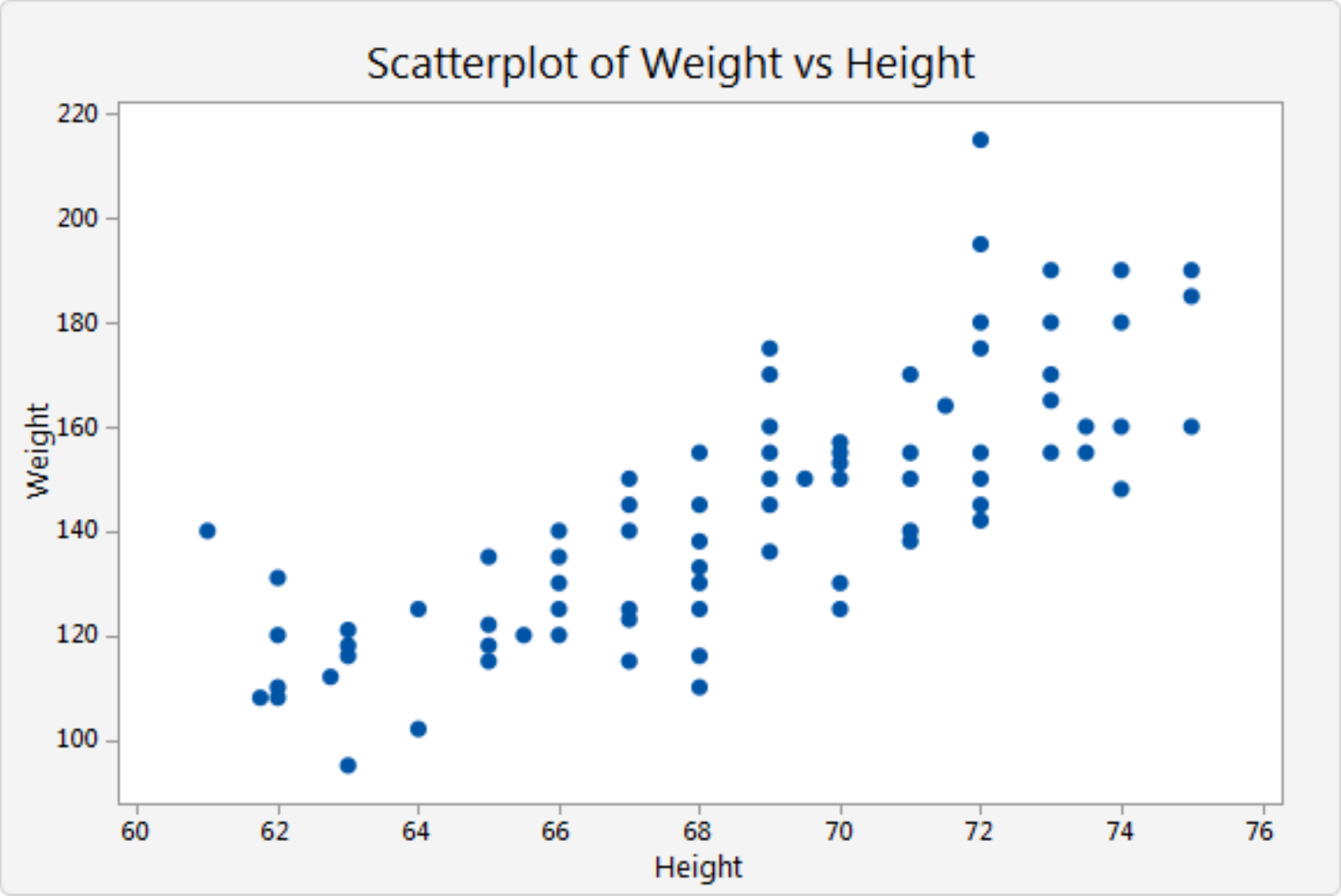
MatPlotLib: Showing distributions using histograms

```
import numpy as np
import matplotlib.pyplot as plt

x = 20 * np.random.randn(10000)

plt.hist (x, 25, range=(-50, 50), histtype='stepfilled',align='mid',
color='g', label='Test Data')
plt.legend()
plt.title('Step Filled Histogram')
plt.show()
```

MatPlotLib: histograms result



Scatterplot



(Review) Normal (Gaussian) Distribution

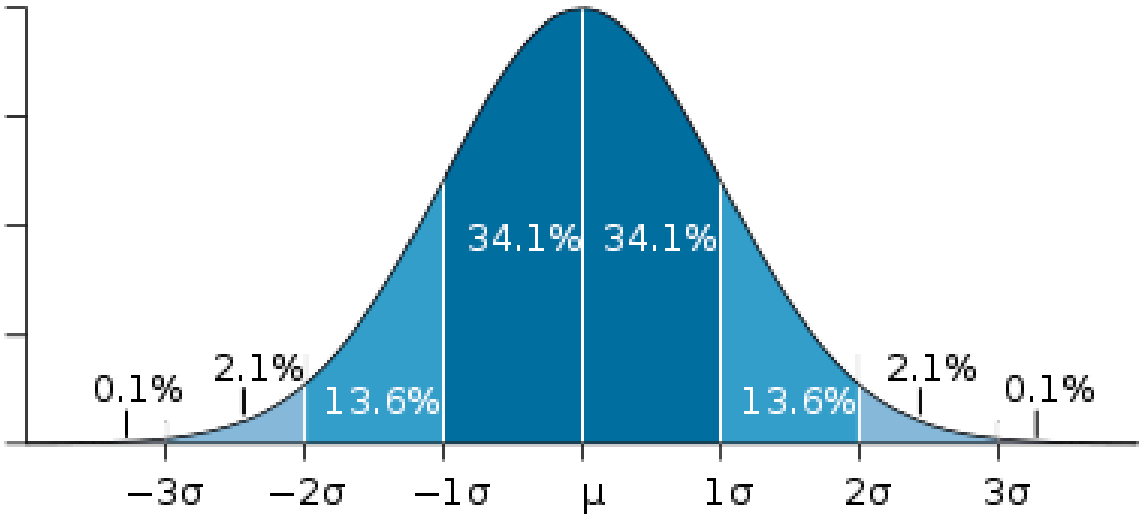
- Symmetric

  Mean = Median = Mode

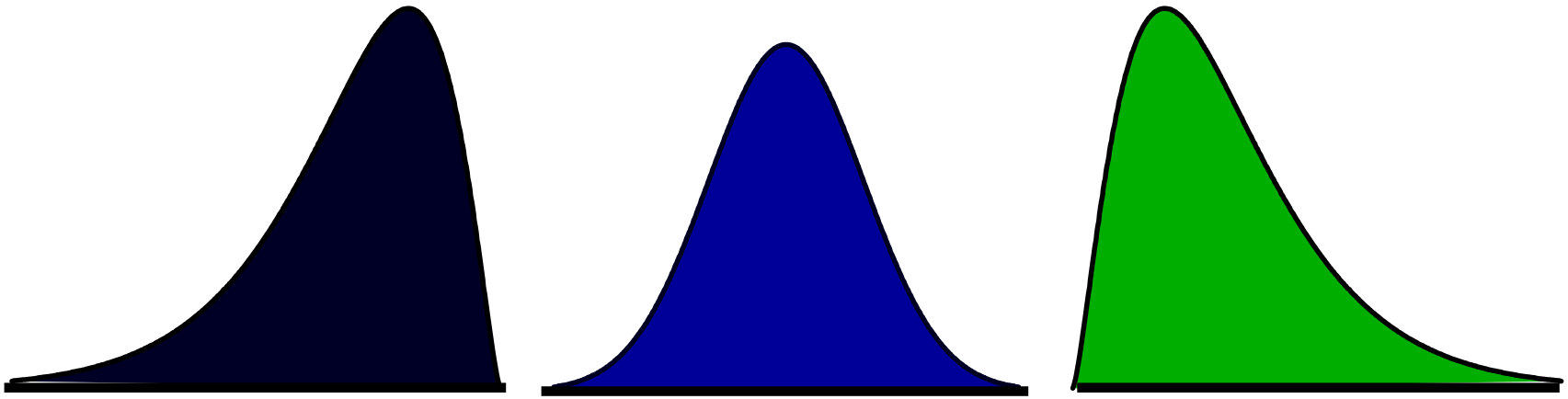
- Theoretical percentiles can be computed exactly

  ~68% of data are within 1 standard deviation of the mean

  >99% within 3 standard deviations (“skinny tails”)



(Review) Skewed Distributions



Negatively Skewed

Symmetric (Not Skewed)

Positively Skewed

Roadmap: Data Inspection

- Data exploration
- Suitability check

Suitability Check

- This is the second phase of data curation.
- Requires business domain experts.
- Requires a metadata (data dictionary) and visualization software tools for browsing the metadata.
- Tasks needed
  - Study the names and meanings of the tables and features, and relationships between the features (attributes).
  - Determine that the existing features seem relevant to data analysis, and that important features do not seem missing (* This is the first phase of feature engineering).

Metadata, Data Dictionary

- Metadata
  - “Data About Data”
  - names of entities and attributes; relationships between attributes; data type of each attribute, constraints on attributes, indexes on attributes; access rights on entities and attributes
  - RDBs store it in a “system catalog”
- Data Dictionary
  - metadata +
  - additional description of metadata +
  - logical data design (ER diagram)

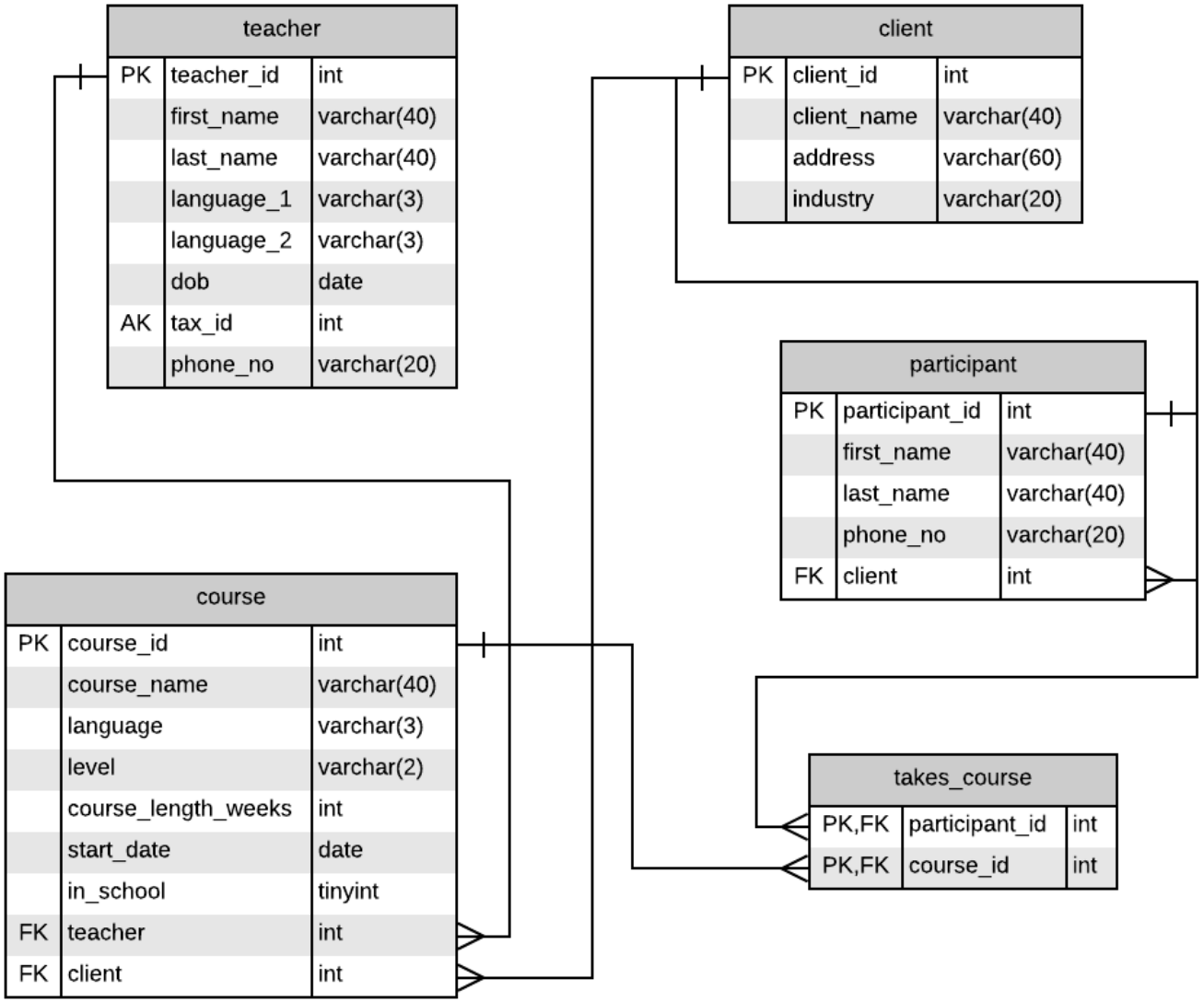
3 Types of Metadata

- Technical Metadata
  - about all entities, attributes, relationships, etc.
- Business (Descriptive) Metadata
  - (non-tech) friendly descriptions of technical metadata
  - sources of data, etc.
- Process Metadata
  - result of every major operation
  - start/end time, disk reads, CPU time, records processed, etc.

RDB System Catalog

- A table of all tables in a database
- A table of all columns for each table
- Integrity constraints for each table
  - primary key, unique, foreign key, check
- Integrity constraints for each column
  - NULL allowed, NULL not allowed
  - data type
- Indexes and hash table created for columns

ER Diagram: Example



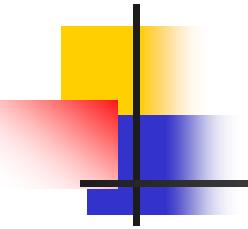
Data Dictionary (see if you understand the meanings of the attributes and data values)

Table	Column	Data type	PK	Nullable	Ref	Description
contacts_tab	contact_id	int	Y	not null		Contact row ID
contacts_tab	first_name	varchar(100)		not null		Contact first name
contacts_tab	last_name	varchar(100)		not null		Contact last name
contacts_tab	dob	date		null		Date of birth
contacts_tab	home_address_id	int		null	address_tab	Home address
contacts_tab	cor_address_id	int		null	address_tab	Correspondence address
contacts_tab	work_address_id	int		null	address_tab	Work address
contacts_tab	marital_stat	char(1)		null		Marital status 'S' - Single, 'M' - Married, 'D' - Divorced, 'W' - Widowed
contacts_tab	employer_name	varchar(100)		null		Employer name
contacts_tab	employer_id	int		null	companies	Employer company from a database, if exists
contacts_tab	position_text	varchar(100)		null		Job title
address_tab	address_id	int	Y	not null		Address row ID
address_tab	address	varchar(100)		null		Address line 1
address_tab	address_2	varchar(100)		null		Address line 2
address_tab	postal_code	varchar(10)		null		Postal code
address_tab	city	varchar(100)		null		City name
address_tab	country	char(3)		null	countries	Country ISO code

Analogy: Metadata and Function Parameters

- Understanding metadata is similar to understanding function parameters.
- If you don't understand the meanings of the parameters, you cannot code.
- If you don't understand the meanings of the metadata, you cannot prepare data for data analysis.

```
import numpy as np
import matplotlib.pyplot as plt

spread = 100 * np.random.rand(100)
center = np.ones(50) * 50
flier_high = 100 * np.random.rand(10) + 100
flier_low = -100 * np.random.rand(100)
data = np.concatenate ( (spread, center, flier_high, flier_low) )

plt.boxplot (data, sym='gx', widths=.75, notch=True)
plt.show()
```



End of Class