CSCI 381/780 Data Analytics -- logistics

Course Description:

- Data science has been one of the fastest growing professions recently. The goal of the Data Analytics class is to prepare students with the necessary skill set and understanding to succeed in this area.
- The first part of the course will go over fundamental concepts spanning across statistics, cross validation, data visualization, data warehousing and python as data manipulation and model development platform.
- Second part of the course will cover common machine learning techniques such as linear and logistics regression, support vector machine, decisions trees and natural language processing.
- Students at the end of the course should be able to carry on to more advanced studies on machine learning and start a career in the data science areas.

CSCI 381/780 Data Analytics -- logistics

Instructor: Dr. Alex Pang

Email: chiuyan.pang@qc.cuny.edu

Lectures: Mon, Wed (8:00pm – 9:15pm)

Pre-requisites:

- CSCI 313 (Data Structures)
- Math 241 (Prob & Stat)

Teaching Assistant: None

Office hours: 9:15 to 9:45 pm after class

Course Objective:

At the end of this course students should

1. have a good overview of the data science professions and modern data analytics platforms.

2. have acquired expertise in using Python as his/her data analysis and model development platform

3. have developed a good analytical mindset in drawing insights on data and making recommendations

4. have understood some of the most common machine learning techniques and feel comfortable in pursuing more advanced skill set in machine learning areas.

CSCI 381/780 Data Analytics -- logistics

Textbook:

Data Analytics Made Accessible:

2021 edition by Anil Maheshwari

Acknowledgement:

I would like to express my special thanks to Dr. Anil Maheshwari for writing such as wonderful textbook in this exciting field as well as his generosity in sharing some of his PowerPoint slides related to his textbook, some of which have been adapted into our course materials

Optional Textbooks for 381:

- Python for Data Analysis by Wes McKinney
- An Introduction to Statistical Learning by Gareth James, Daniel Witten, et al (http://www-bcf.usc.edu/~gareth/ISL/)

Almost required textbook for 780:

- Hands-On Machine Learning with Scikit-Learn and Tensor Flow by Aurelien Geron

CSCI 381/780 Data Analytics Syllabus – Communication

Communication and Class Participation

- Communication is mainly through the following channels
  - Announcement on Blackboard
  - Discussion Forum on Blackboard
  - Direct emails

- Students are expected and highly encouraged to participate in the discussion forum on Blackboard
- Lectures will be given synchronously on Zoom and will be recorded.
- Reading Assignments will be announced and are expected to complete asynchronously

CSCI 381/780 Data Analytics -- logistics

Section 1: Core Business Intelligence and Data Analytics Concepts

1. Data Science / Data Analytics Overview
2. Probability and Statistics Review
3. Exploratory Data Analysis
4. Data Visualization
5. Machine Learning Overview, Linear Regression and Common Data Scientist's Toolbox

Section 2: Popular Data Mining Techniques

6. Classification and Naïve Bayes Algorithm
7. Classification and Logistics Regression
8. Support Vector Machine
9. Decision Trees
10. Clustering
11. Text Mining
12. Big Data

CSCI 381/780 Data Analytics -- logistics

Section 3: More Advanced Techniques and Application

13. Neural Network and Deep Learning
14. Business Intelligence Applications
15. Amazon AWS, Microsoft Azure

CSCI 381/780 Data Analytics -- logistics

Grade contribution:

- 30% Homework assignments (3 HWs)
- 20% mid-term exam
- 20% final exam
- 25% final Project
- 5% Review Quizzes & Class Participation (Blackboard)

- Some questions may be mandatory for graduate students but optional for undergraduates

Homework format:

- Python 3 Notebook

Exam format:

- Multiple choices and written short answers

Review Quizzes:

- Multiple choices and written short answers

Minimum to pass the course:

65% raw score

CSCI 381/780 Data Analytics -- logistics

Final Course grades may be curved so that the median grade is between B- and C+

CSCI 381/780 Data Analytics - Class Participation

Class Participation is important

I don’t want my class to be like this

I want my class to be like this

There is never such thing as dumb question

CSCI 381/780 Data Analytics -- logistics

Collaboration Policy:

You are allowed and encouraged to discuss homework. Discussion on Blackboard is encouraged, so everyone can benefit; however, do NOT post or share solutions or parts of solutions. Homework and final project must be done and written up independently.

Academic Integrity Policy:

Absentees are solely responsible for catching-up. Academic dishonesty, such as plagiarism or cheating - taking other people's work with or without their permission in order to get credit for yourself, will be dealt with seriously, including an "F" grade for the course and/or disciplinary action according to the University's policy on academic integrity

CSCI 381/780 Data Analytics -- logistics

Recording Consent for online session

Students who participate in this class with their camera on or use a profile image are agreeing to have their video or image recorded solely for the purpose of creating a record for students enrolled in the class to refer to, including those enrolled students who are unable to attend live. If you are unwilling to consent to have your profile or video image recorded, be sure to keep your camera off and do not use a profile image. Likewise, students who un-mute during class and participate orally are agreeing to have their voices recorded. If you are not willing to consent to have your voice recorded during class, you will need to keep your mute button activated and communicate exclusively using the "chat" feature, which allows students to type questions and comments live.

CSCI 381/780 Data Analytics – on the subject of cheating

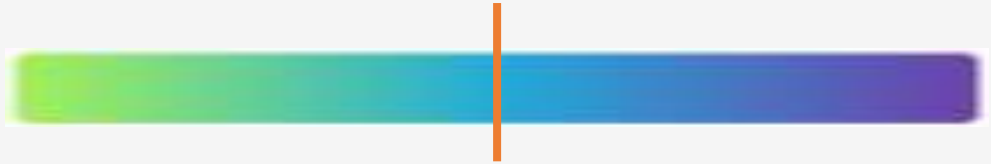
- It is not a lie if you believe it
- You are not cheating if you are not caught
- I am just trying to help my friends, no harm no foul
- I know the materials, just not have enough time for working on the HW
- …..

- You think your professors are dumb?
- When you work full-time after school, can you call your friends to see if you can copy-and-paste from him or her or look for course material from previous semester?
- What's your goal in signing up for a class? A fake GPA or learn some real skills
- You will be graded based on the whole class

CSCI 381/780 Data Analytics -- logistics

Teaching Style:

Theory Practical



No homework 20 hours per week

Dry
Only me talking

Stand-up comedy
Highly Interactive

follow textbook

No textbook

- I do NOT shy away from using various free resources from the Internet. Remember DRY principle.
- When you want to learn something, where and how do you start?

CSCI 381/780 Data Analytics -- logistics

Weekly Routine:

- Monday:
  - Theory, bi-weekly homework due (if any)
  - Finished reading assignments from the textbook

- Wednesday:
  - Application, homework description,
  - Weekly summary, next week preview, assign reading from the textbook

- Saturday
  - Informal online office hours, will announce before hand,
  - Email is the best way to communicate

- Will post the PowerPoint after class on Blackboard

CSCI 381/780 Data Analytics -- logistics

Python:

- Lectures as well as homework will be based on Python 3 notebooks
- You need to make sure you have a PC or laptop where you can run Python notebooks
- Recommended distribution and installation is Anaconda (https://www.anaconda.com/)
- Make sure you are familiar with the syntax as well as the Pandas and NumPy library for data manipulations
- The textbook has a chapter on Python !

HW and Exam Dates (tentative)

3/3: HW 1 Posted
3/10: HW 1 Due

3/24: HW 2 Posted
3/31: HW 2 Due

4/7: HW 3 Posted
4/14: HW 3 Due

4/28: Final Project Posted
5/19: Final Project Due

3/17: Mid-Term Exam

5/24: Final-Exam

Data Science Overview

Data Analytics, Data Mining

Decision Science, Business Intelligence

Artificial Intelligence, Machine Learning

Statistical Learning, Data Science

Sexiest Job of the twenty-first century

- by Harvard Rev

Data Science is new electricity

- by Andrew Ng

Data Science Overview

Why it becomes so popular now?

What are the problems does Data Analytics try to solve?

How does it work? When it will be useful and relevant?

What are the criteria that determine if it succeed in delivering its goal?

What are all these Data Warehouse, Data Mart, ETL Process, Business Intelligence, OLAP, Dashboard, Data Mining, Machine Learning terms mean?

How to become a successful Data Scientist? What are the required skillset?

Why Data Science becomes so popular

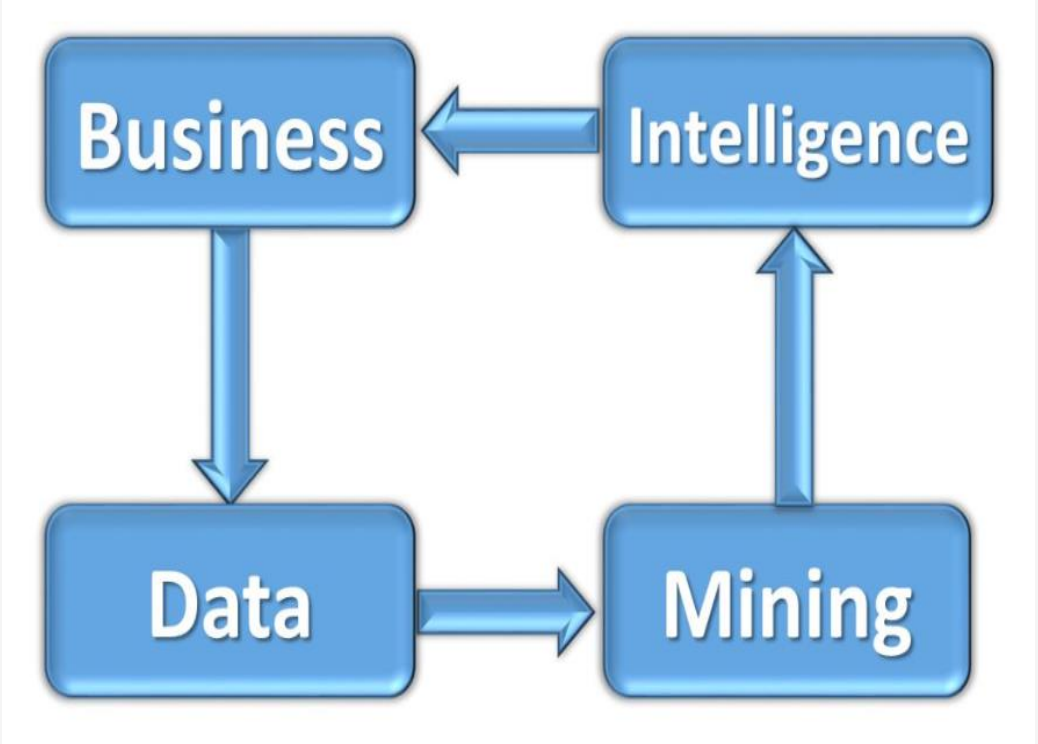
- More and more data available b/c web 2.0
- Faster and Faster Hardware
- Social Media, Availability of unstructured data/alternative data
- Deep Neural Network
- Real-life Proof: Alpha-Go, Alexa, Driverless car ….

It answers questions and solve business problems

Goal of Data Analytics

- *Answer basic strategic business questions* based on data a company has collected.
  - What is the best-selling beer brand.
  - Does reducing classroom size (teacher to students ratio) really improve student performance?

- Discover not-so-obvious *patterns* from the data



BIDM Cycle

- Business is the act of doing something productive to serve someone's needs, and thus earn a living and make the world a better place.
- Business activities are recorded on paper or using electronic media, and then these records become data. There is more data from customers' responses and on the industry as a whole.
- All this data can be analyzed and mined using special tools and techniques to generate patterns and intelligence, which reflect how the business is functioning.
- These ideas can then be fed back into the business so that it can evolve to become more effective and efficient in serving customer needs. And the cycle continues on

Real-life example of Data Analytics Solutions

- Example from the textbook: MoneyBall

- Another real-life Example:

Most Expensive State to buy a home

The Most Expensive State to Buy a Home in Isn’t New York or California

https://www.barrons.com/articles/most-expensive-state-to-buy-a-home-51567108698

https://smartasset.com/mortgage/cheapest-states-to-buy-a-home-2019?mod=article_inline

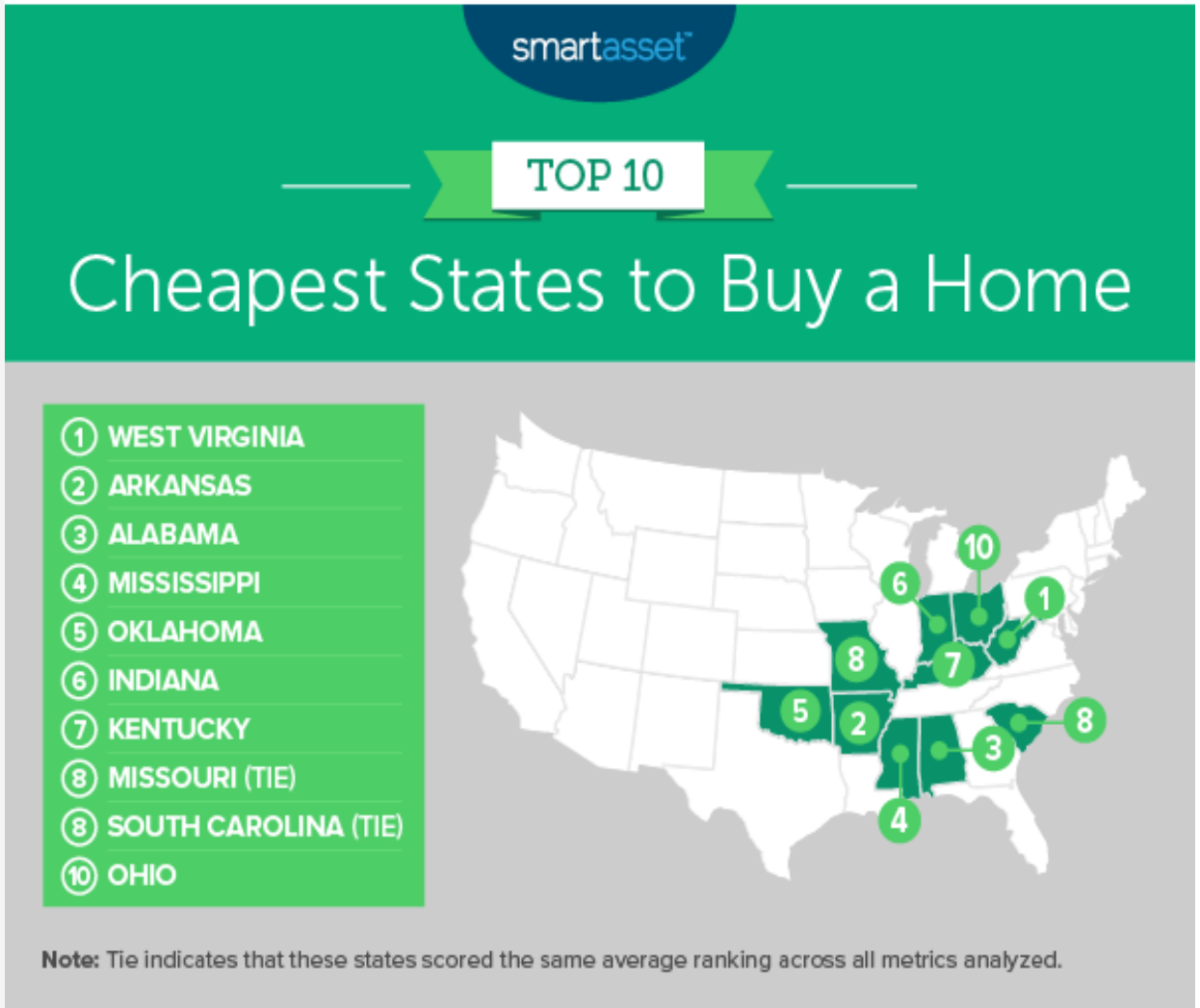
Data Science Real Life Example

Photograph by Breno Assis

Massachusetts ranked as the most expensive state to buy a home in, according to a report from personal-finance website SmartAsset. The analysis assessed 48 states and Washington, D.C., based on the following metrics (Delaware and Louisiana were not included due to insufficient data):

- Effective property tax rate, based on U.S. Census Bureau data
- Median listing price and price per square foot, according to Zillow ZG
- Median value for homes in the bottom third of the market
- Average closing costs, according to SmartAsset's own closing cost calculator

All the states were ranked for each of these metrics, and then researchers calculated their average ranking. This then determined the index value they were assigned on a scale from 0 to 100, with 100 being the cheapest. Massachusetts' index came in at zero, while West Virginia was the most affordable state with a value of 100.



Extra Credits (Class Participation)

Find a similar news article where it clearly showcases how Data Analytics can solve or answer interesting questions.

I will start a discussion thread on this. No explanation is needed. But you are more than welcome to describe how the article does help you realize how data analytics is realized in real-life

"Data Science" consists of two words:

Data and Science

Let's start with Data

Different Type of Data

Structured Data

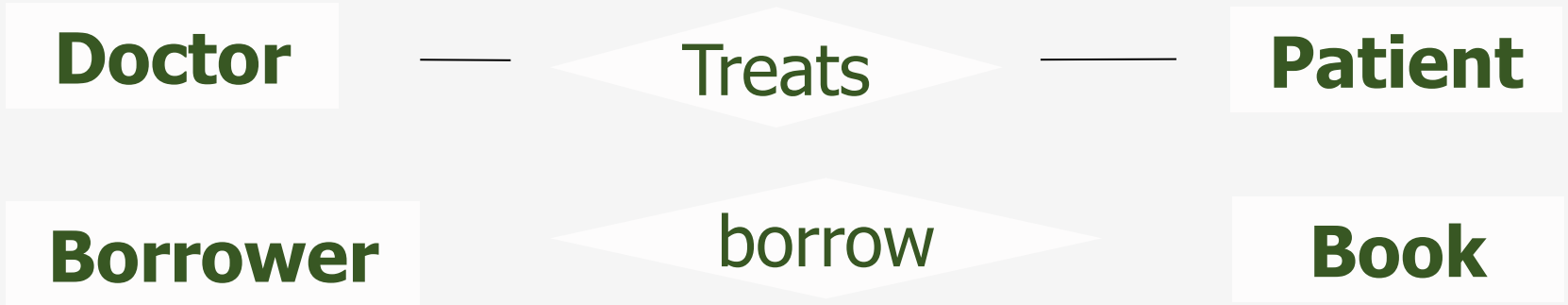
- Tabular Data, Typically stored in a relational database table
  - Example: Think in terms of managing computer system in keeping track of books in a library
    - A Borrower Table with borrower ID, name, birth-date, address as columns
    - A Book table with author, name of the book, published date, publisher as columns

- Storage format
  - CSV file, an excel file or a JSON file, but more commonly in a SQL database

- Type of data: Numerical fields, categorical field, date-time

Structured Data (database and data modeling)

- Data is organized in tables
  - Tables relate to entities and relationships among entities
  - Entities are nouns like Doctor, Patient
  - Relationships are verbs like ‘Treats’
- Data tables can be managed using SQL language in simple declarative mode



Different Type of Data

Unstructured Data

- Blob (binary large objects)
- Satellite Data
- Graph to links friends on Facebook.
- Image Data
- Video with annotation
- Twitter feeds
- Time-Series data

Storage format:

- XML, JSON

Type of Databases

- Relational databases
  - MSSQL
  - MySQL
  - Postgres
- NoSQL databases
  - MongoDB
- Time Series Databases

Common Databases for storing data

SQL Database:

- Microsoft SQL Server (https://www.microsoft.com/en-us/sql-server/sql-server-editions-express)
- MySQL (https://www.mysql.com/)
- Postgres SQL (https://www.postgresql.org/)

NoSQL Database:

- MongoDB (https://www.mongodb.com/)

Data Science is the new electricity

Data is a new set of asset class

How to collect and manage Data ?

Expectation vs Reality

Expectation:

Your CSCI-381 /780 instructor

Expectation vs Reality

Reality:

Your CSCI-381 /780 instructor

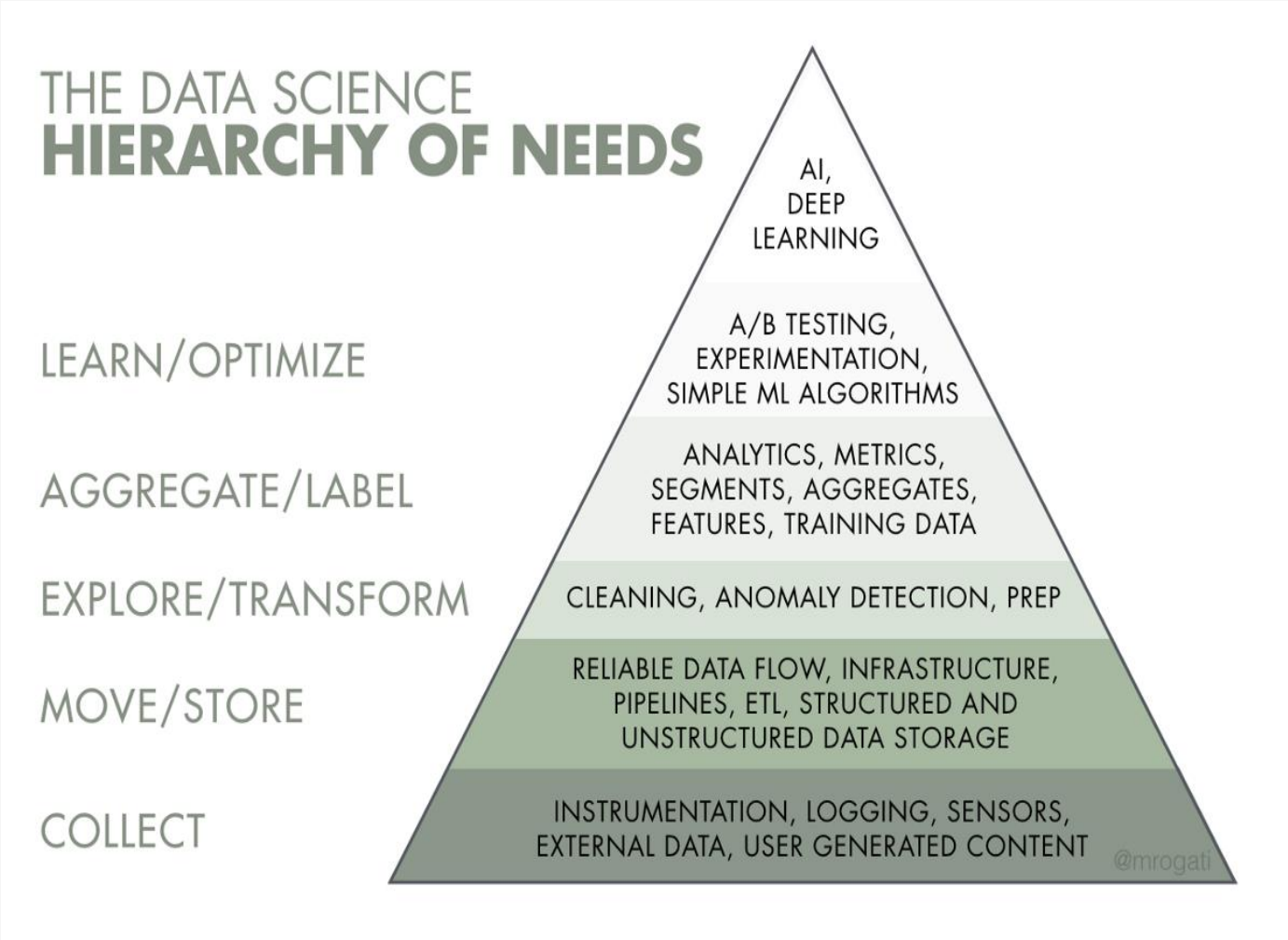
Three type of Data Science jobs

Data Science: Expectation and Reality

https://www.youtube.com/watch?v=8LucP1wiX1g&t=214s

3 common job classifications:

- Data Engineers
- Data Analytics
- Data Scientists



Data as a new natural resource or a new asset class

Data Processing Chain, Data Pipeline, Data Flow

Datafication, Internet of Things

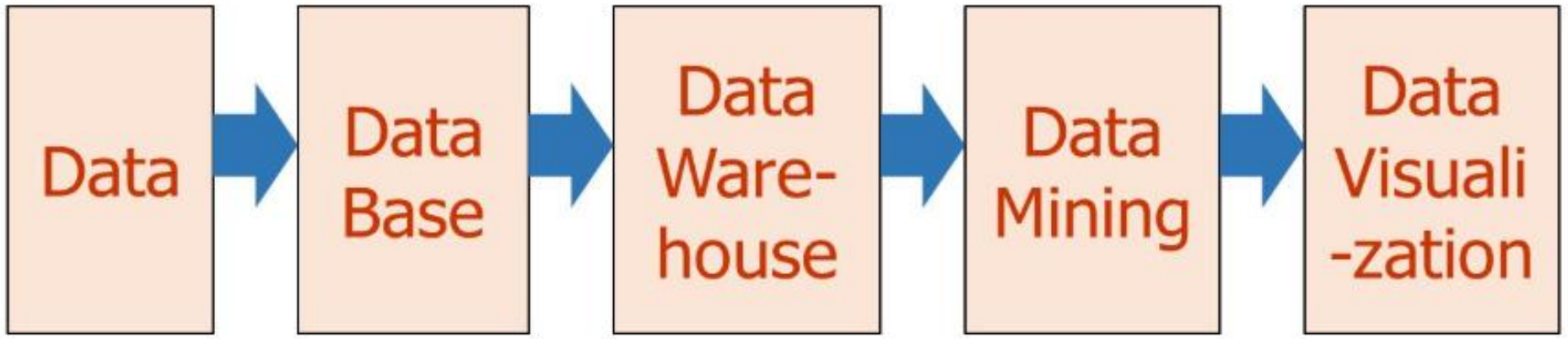
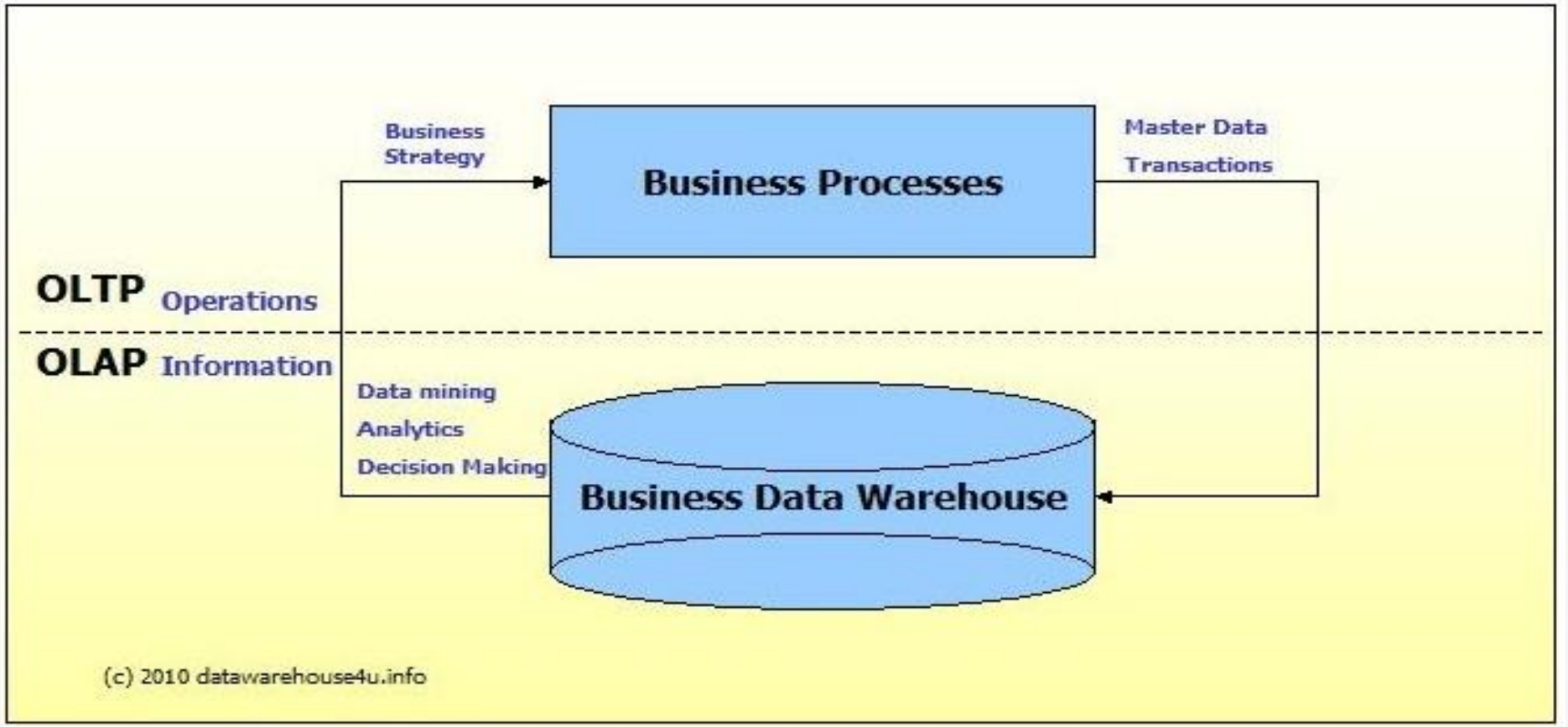


Figure 1.2: Data Processing Chain

Database vs Data Warehouse

OLTP (On-Line Transaction Processing) vs OLAP (On-Line Analytical Processing)



Data Warehouse Definition

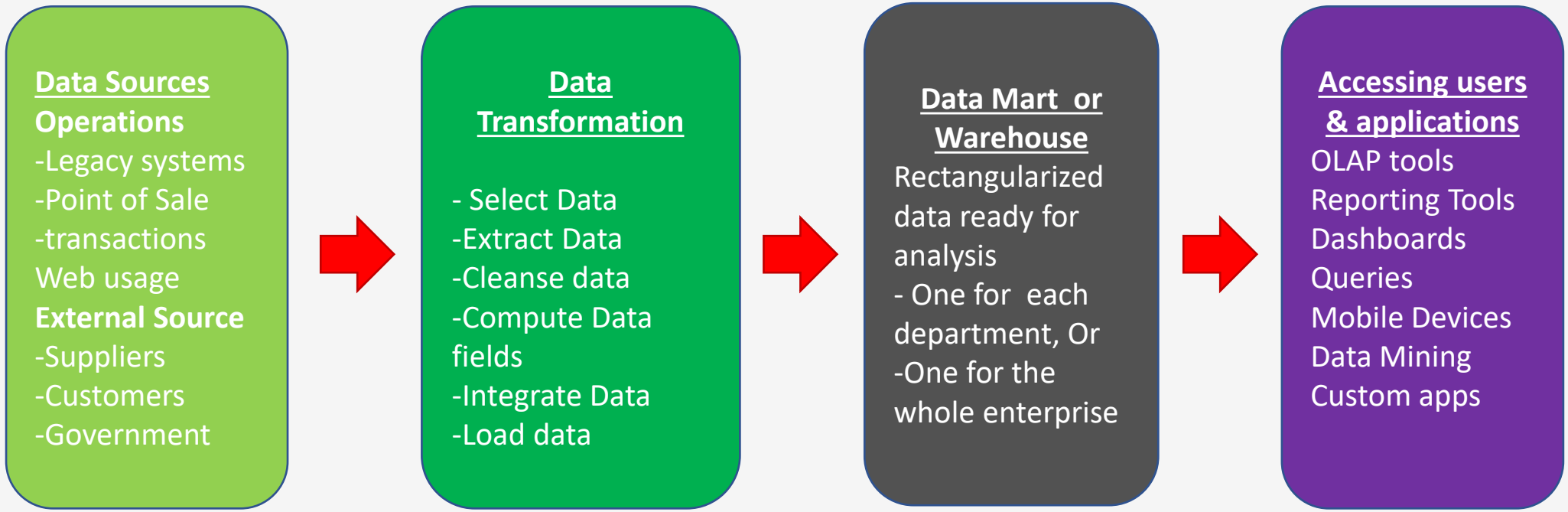
Data warehouse (DW) is an organized collection of integrated, subject-oriented databases designed to support business decision functions.

- organized at the right level of granularity (usually some unit of time)
- provides clean enterprise-wide data in a standardized format for reports, queries and analysis.
- has to be constantly kept up-to-date to be useful
- Has clear Metadata (Data about data)

A DW is physically and functionally separate from an operational transactional database.

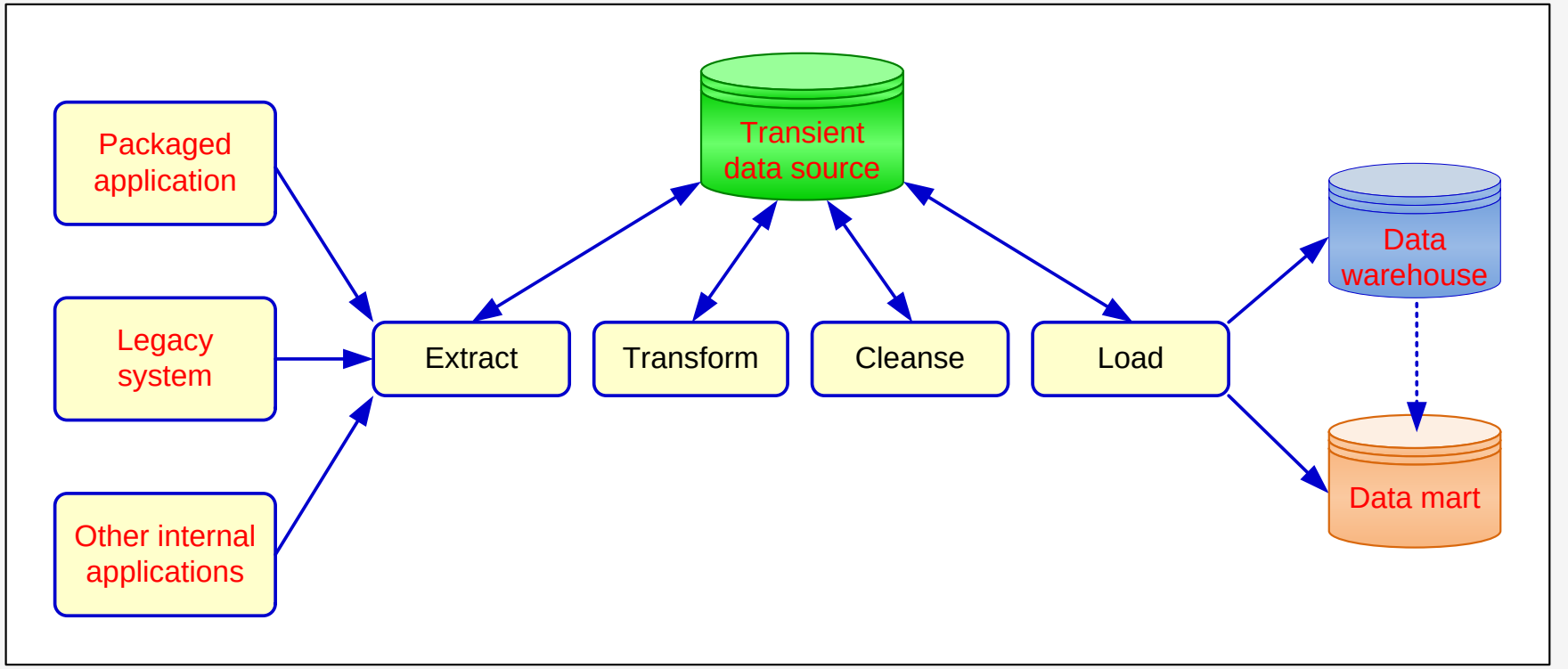
- Creating a DW for analysis and queries represents significant investment in time and effort.

A Conceptual Framework for Data Warehouse



ETL Process for creating a Datawarehouse

Extraction, Transformation, and Load (ETL) process



DataWarehouse Concepts

Watch this 8 min video presentation on Data Warehousing concepts and approaches

https://www.youtube.com/watch?v=zTs5zjSXnvs

Quick reminder of what is 1-2-3 Normal form
https://www.dummies.com/programming/sql/sql-first-second-and-third-normal-forms/

Database vs Data Warehouse

Function	Database	Data Warehouse
Purpose	Data stored in databases can be used for many purposes including day-to-day operations	Data in DW is cleansed data useful for reporting and analysis
Granularity	Highly granular data including all activity and transaction details	Lower granularity data; rolled up to certain key dimensions of interest
Complexity	Highly complex with dozens or hundreds of data files, linked through common data fields	Typically organized around a large fact tables, and many lookup tables
Size	Database grows with growing volumes of activity and transactions. Old completed transactions are deleted to reduce size.	Grows as data from operational databases is rolled-up and appended every day. Data is retained for long-term trend analyses
Architectural choices	Relational, and object-oriented, databases	Star schema, or Snowflake schema
Data Access mechanisms	Primarily through high level languages such as SQL. Traditional programming access DB through Open DataBase Connectivity (ODBC) interfaces	Accessed through SQL; SQL output is forwarded to reporting tools and data visualization tools

What's next

Now we have the data stored in Data Warehouse

What's next?

Data Mining
Data Visualization
…

Recap 1

- Goal and purpose and why Data Science
- Real-life Data Analytic Application
- Different type of data
- Different type of Databases
- Learn Python as fast and as much as possible

- Make sure you had sent me an email at least once
- HW Zero

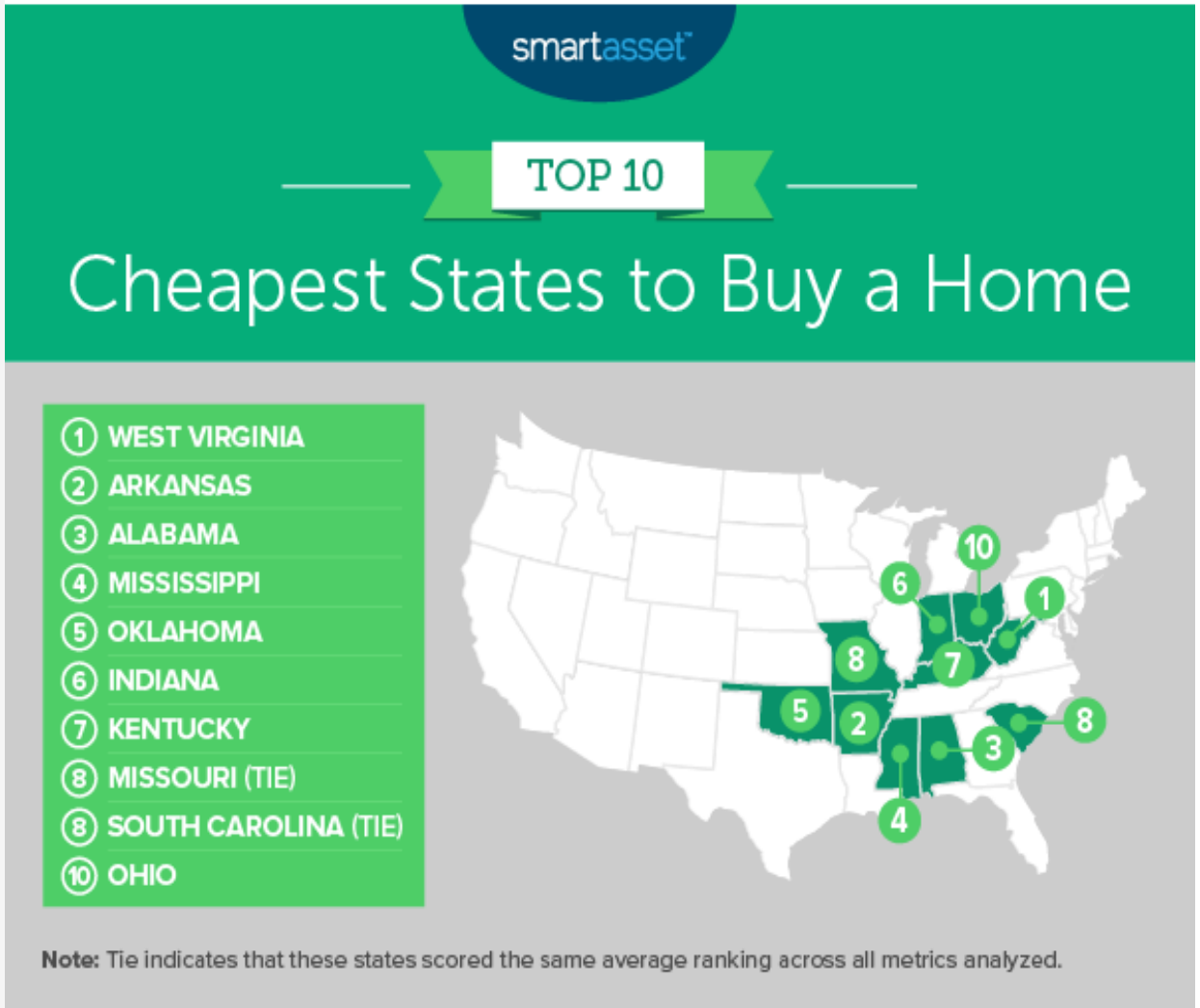
Data Science Real Life Example

Photograph by Breno Assis

Massachusetts ranked as the most expensive state to buy a home in, according to a report from personal-finance website SmartAsset. The analysis assessed 48 states and Washington, D.C., based on the following metrics (Delaware and Louisiana were not included due to insufficient data):

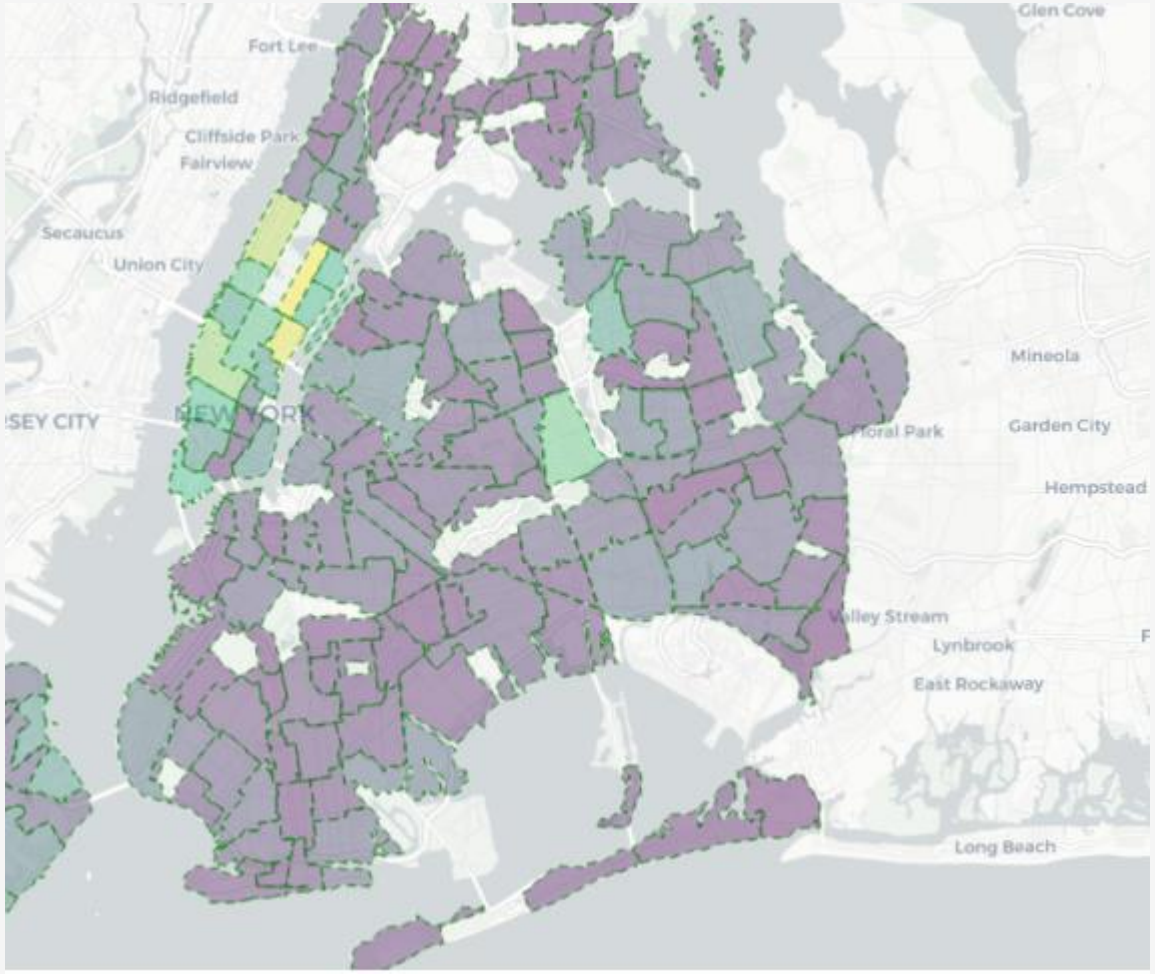
- Effective property tax rate, based on U.S. Census Bureau data
- Median listing price and price per square foot, according to Zillow ZG
- Median value for homes in the bottom third of the market
- Average closing costs, according to SmartAsset's own closing cost calculator

All the states were ranked for each of these metrics, and then researchers calculated their average ranking. This then determined the index value they were assigned on a scale from 0 to 100, with 100 being the cheapest. Massachusetts' index came in at zero, while West Virginia was the most affordable state with a value of 100.

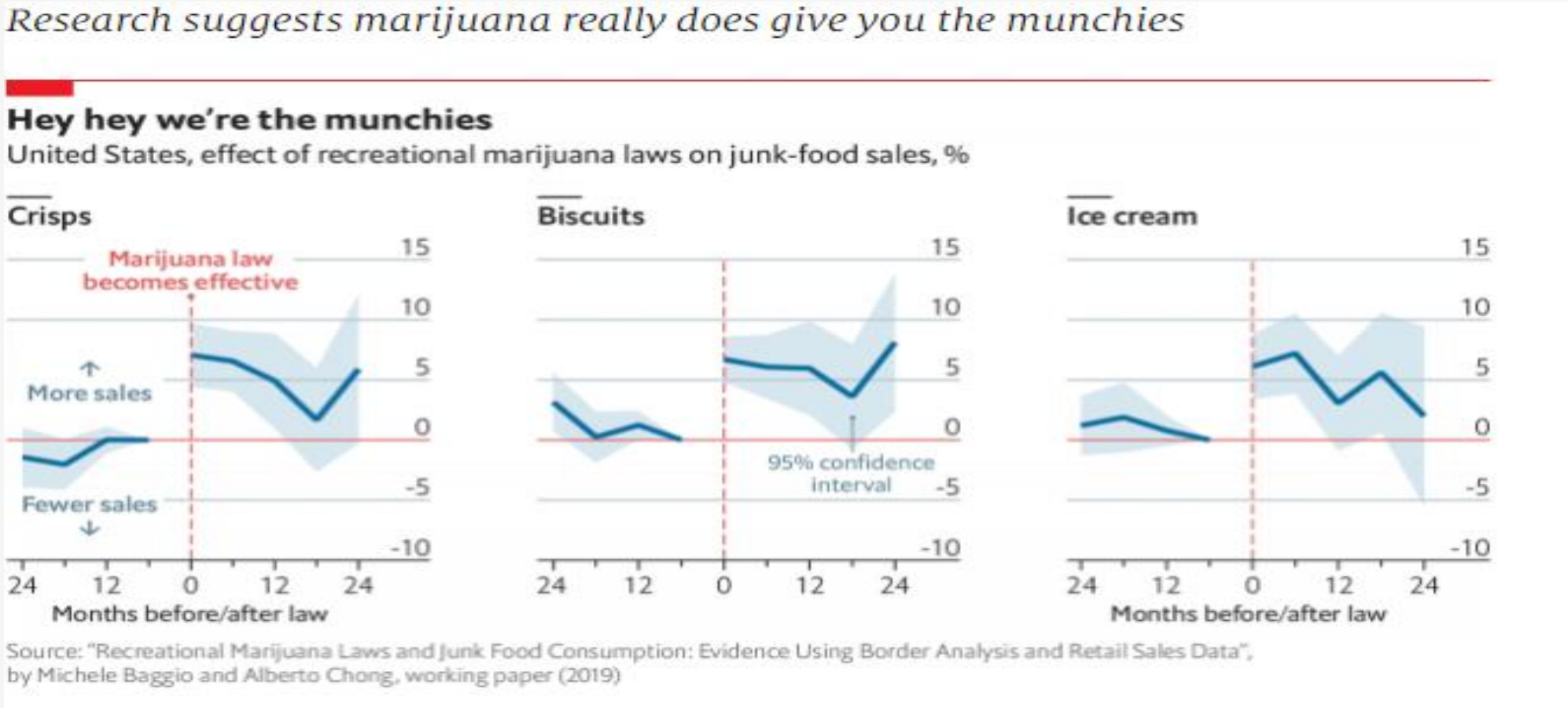


Queens neighborhoods saw drop in home sale prices in August

https://qns.com/story/2019/09/06/the-price-is-wrong-these-queens-neighborhoods-saw-drop-in-home-sale-prices-in-august/



Legal weed is linked to higher junk-food sales



https://www.economist.com/graphic-detail/2019/08/16/legal-weed-is-linked-to-higher-junk-food-sales?al_applink_data=%7B%22target_url%22%3A%22https%3A%5C%2F%5C%2Fecon.trib.al%5C%2FIoQh5j8%22%2C%22extras%22%3A%7B%22fb_app_id%22%3A128869720483178%7D%2C%22referer_app_link%22%3A%7B%22url%22%3A%22fb%3A%5C%2F%5C%2F%5C%2F%3Fapp_id%3D128869720483178%22%2C%22app_name%22%3A%22Facebook%22%7D%7D

Learn Python and Pandas as quickly as you can

- Dive into Python
- Dive into Pandas

Learn Python as fast and as much as possible

- Free Python course on https://www.datacamp.com/
- There is a chapter in Python in the textbook
- DiveIntoPython.pdf & cheatsheets from course resources on Blackboard
  - https://www.datacamp.com/community/data-science-cheatsheets
  - https://ehmatthes.github.io/pcc/cheatsheets/README.html
- Great free book specifically for numerical computation
  - https://www.amazon.com/Programming-Computations-Introduction-Simulations-Computational-ebook-dp-B078YGVNSF/dp/B078YGVNSF/ref=mt_other?_encoding=UTF8&me=&qid=

Specific Python book for numerical computations

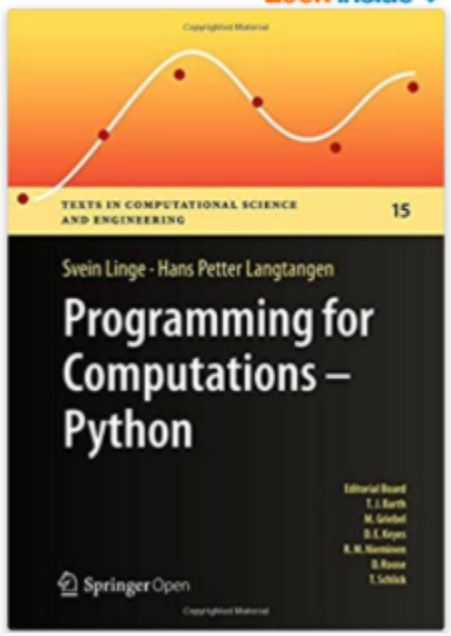
Programming for Computations - Python: A Gentle Introduction to Numerical Simulations with Python (Texts in Computational Science and Engineering Book 15) 1st ed. 2016 Edition, Kindle Edition

by Svein Linge (Author), Hans Petter Langtangen (Author) | Format: Kindle Edition

4.4 out of 5 stars 646 ratings

Book 15 of 24: Texts in Computational Science and Engineering

Look inside



eTextbook	Hardcover	Paperback	Other Sellers
$0.00	from $221.62	$59.99 - $67.26	See all 3 versions

Buy

$0.00

Print List Price: ~~$59.99~~ Save $59.99 (100%)

eTextbook features:

- Highlight, take notes, and search in the book
- In this edition, page numbers are just like the physical edition
- Create digital flashcards instantly

Read with the free Kindle apps (available on iOS, Android, PC & Mac) and on Fire Tablet devices. See all supported devices

Buy now with 1-Click

Deliver to:

Your Kindle Library

This title is not supported on Kindle E-readers or Kindle for Windows 8 app. Learn more

Enter a promotion code or Gift Card

Sold by: Amazon.com Services LLC

More Buying Choices

New (1) from

Python CheatSheet

Beginner's Python Cheat Sheet

Variables and Strings

Variables are used to store values. A string is a series of characters, surrounded by single or double quotes.

Hello world

```
print("Hello world!")
```

Hello world with a variable

```
msg = "Hello world!"
print(msg)
```

Concatenation (combining strings)

```
first_name = 'albert'
last_name = 'einstein'
full_name = first_name + ' ' + last_name
print(full_name)
```

Lists

A list stores a series of items in a particular order. You access items using an index, or within a loop.

Lists (cont.)

List comprehensions

```
squares = [x**2 for x in range(1, 11)]
```

Slicing a list

```
finishers = ['sam', 'bob', 'ada', 'bea']
first_two = finishers[:2]
```

Copying a list

```
copy_of_bikes = bikes[:]
```

Tuples

Tuples are similar to lists, but the items in a tuple can't be modified.

Making a tuple

```
dimensions = (1920, 1080)
```

If statements

If statements are used to test for particular conditions and respond appropriately.

Conditional tests

```
equals            x == 42
not equal         x != 42
greater than      x > 42
```

Dictionaries

Dictionaries store connections between pieces of information. Each item in a dictionary is a key-value pair.

A simple dictionary

```
alien = {'color': 'green', 'points': 5}
```

Accessing a value

```
print("The alien's color is " + alien['color'])
```

Adding a new key-value pair

```
alien['x_position'] = 0
```

Looping through all key-value pairs

```
fav_numbers = {'eric': 17, 'ever': 4}
for name, number in fav_numbers.items():
    print(name + ' loves ' + str(number))
```

Looping through all keys

```
fav_numbers = {'eric': 17, 'ever': 4}
for name in fav_numbers.keys():
    print(name + ' loves a number')
```

Looping through all the values

```
fav_numbers = {'eric': 17, 'ever': 4}
for number in fav_numbers.values():
    print(str(number) + ' is a favorite')
```

Python CheatSheet

```
first_name = 'albert'
last_name = 'einstein'
full_name = first_name + ' ' + last_name
print(full_name)
```

Lists

A list stores a series of items in a particular order. You access items using an index, or within a loop.

Make a list

```
bikes = ['trek', 'redline', 'giant']
```

Get the first item in a list

```
first_bike = bikes[0]
```

Get the last item in a list

```
last_bike = bikes[-1]
```

Looping through a list

```
for bike in bikes:
    print(bike)
```

Adding items to a list

```
bikes = []
bikes.append('trek')
bikes.append('redline')
bikes.append('giant')
```

Making numerical lists

```
squares = []
for x in range(1, 11):
    squares.append(x**2)
```

If statements

If statements are used to test for particular conditions and respond appropriately.

Conditional tests

```
equals            x == 42
not equal         x != 42
greater than      x > 42
  or equal to     x >= 42
less than         x < 42
  or equal to     x <= 42
```

Conditional test with lists

```
'trek' in bikes
'surly' not in bikes
```

Assigning boolean values

```
game_active = True
can_edit = False
```

A simple if test

```
if age >= 18:
    print("You can vote!")
```

If-elif-else statements

```
if age < 4:
    ticket_price = 0
elif age < 18:
    ticket_price = 10
else:
    ticket_price = 15
```

```
fav_numbers = {'eric': 17, 'ever': 4}
for name in fav_numbers.keys():
    print(name + ' loves a number')
```

Looping through all the values

```
fav_numbers = {'eric': 17, 'ever': 4}
for number in fav_numbers.values():
    print(str(number) + ' is a favorite')
```

User input

Your programs can prompt the user for input. All input is stored as a string.

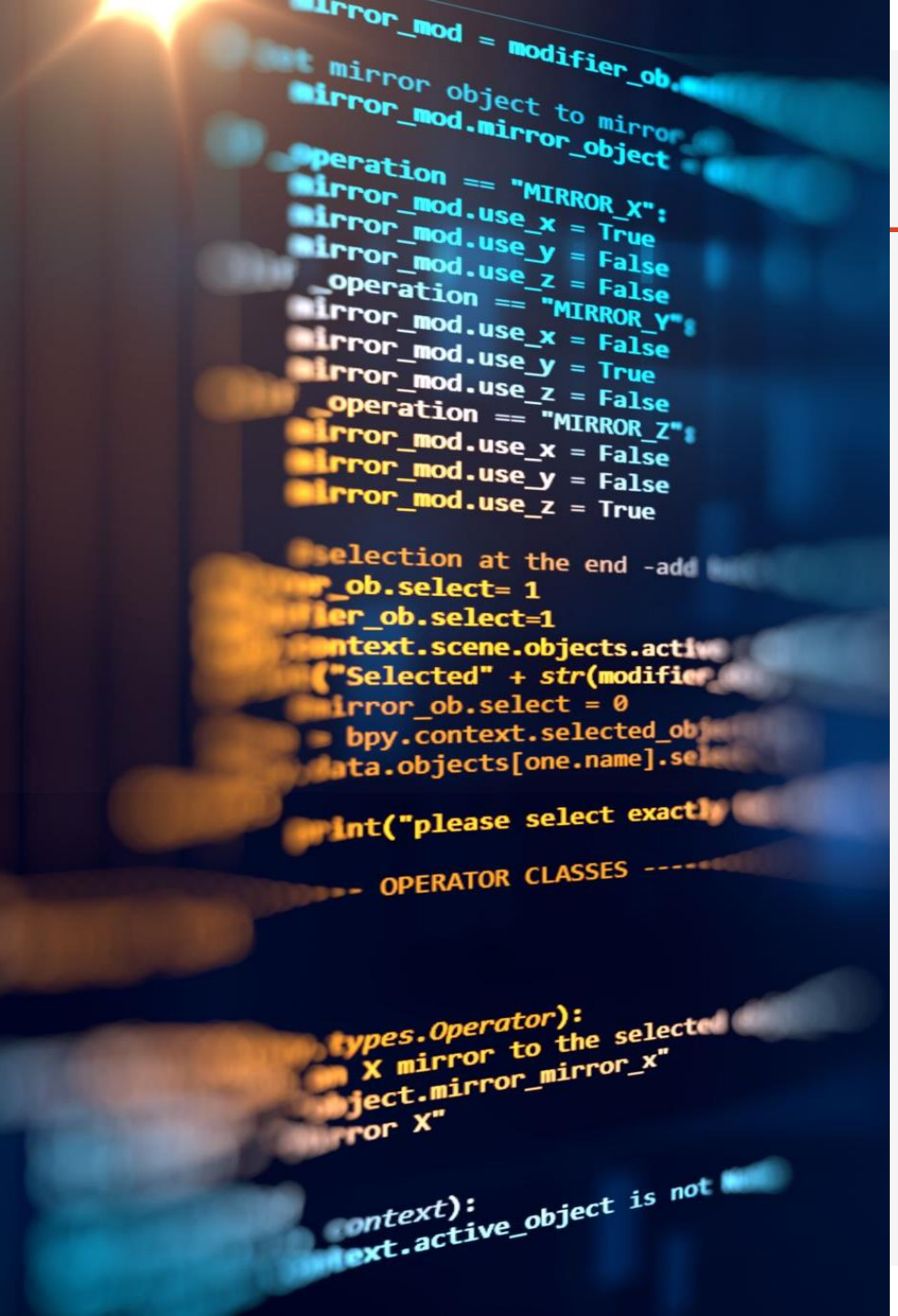
Prompting for a value

```
name = input("What's your name? ")
print("Hello, " + name + "!")
```

Prompting for numerical input

```
age = input("How old are you? ")
age = int(age)

pi = input("What's the value of pi? ")
pi = float(pi)
```



Python core concepts checklist

- How to create a class
- Distinguish between class methods versus instance method
- Familiar with the various built-in data type, know how to parse a date string into a date object
- Know what is a list, and how to loop through each element in the list, know list comprehension
- …

.

Learning by doing

Data Mining 101

Now we have the data in a datastore, how are we going to get useful information out?

SQL Query => Simple Aggregation (mean) => Simple Statistics (standard deviation) => Hypothesis Testing => Data Mining => Artificial Intelligence

- Make sure you master SQL well
- Low hanging fruits;
- Best bang for the bucks!

Learning by doing

https://www.w3resource.com/sql-exercises/

SQL CheatSheet

SQL CHEAT SHEET http://www.sqltutorial.org

QUERYING DATA FROM A TABLE

```
SELECT c1, c2 FROM t;
```
Query data in columns c1, c2 from a table

```
SELECT * FROM t;
```
Query all rows and columns from a table

```
SELECT c1, c2 FROM t
WHERE condition;
```
Query data and filter rows with a condition

```
SELECT DISTINCT c1 FROM t
WHERE condition;
```
Query distinct rows from a table

```
SELECT c1, c2 FROM t
ORDER BY c1 ASC [DESC];
```
Sort the result set in ascending or descending order

```
SELECT c1, c2 FROM t
ORDER BY c1
LIMIT n OFFSET offset;
```
Skip *offset* of rows and return the next n rows

```
SELECT c1, aggregate(c2)
FROM t
GROUP BY c1;
```
Group rows using an aggregate function

```
SELECT c1, aggregate(c2)
FROM t
GROUP BY c1
HAVING condition;
```
Filter groups using HAVING clause

QUERYING FROM MULTIPLE TABLES

```
SELECT c1, c2
FROM t1
INNER JOIN t2 ON condition;
```
Inner join t1 and t2

```
SELECT c1, c2
FROM t1
LEFT JOIN t2 ON condition;
```
Left join t1 and t1

```
SELECT c1, c2
FROM t1
RIGHT JOIN t2 ON condition;
```
Right join t1 and t2

```
SELECT c1, c2
FROM t1
FULL OUTER JOIN t2 ON condition;
```
Perform full outer join

```
SELECT c1, c2
FROM t1
CROSS JOIN t2;
```
Produce a Cartesian product of rows in tables

```
SELECT c1, c2
FROM t1, t2;
```
Another way to perform cross join

```
SELECT c1, c2
FROM t1 A
INNER JOIN t2 B ON condition;
```
Join t1 to itself using INNER JOIN clause

USING SQL OPERATORS

```
SELECT c1, c2 FROM t1
UNION [ALL]
SELECT c1, c2 FROM t2;
```
Combine rows from two queries

```
SELECT c1, c2 FROM t1
INTERSECT
SELECT c1, c2 FROM t2;
```
Return the intersection of two queries

```
SELECT c1, c2 FROM t1
MINUS
SELECT c1, c2 FROM t2;
```
Subtract a result set from another result set

```
SELECT c1, c2 FROM t1
WHERE c1 [NOT] LIKE pattern;
```
Query rows using pattern matching %, _

```
SELECT c1, c2 FROM t
WHERE c1 [NOT] IN value_list;
```
Query rows in a list

```
SELECT c1, c2 FROM t
WHERE  c1 BETWEEN low AND high;
```
Query rows between two values

```
SELECT c1, c2 FROM t
WHERE  c1 IS [NOT] NULL;
```
Check if values in a table is NULL or not

SQL Check-list

- What are primary keys
- How to select subset data
- How to join two tables
- How to group by data

Note: We will NOT spend too much time on SQL, because many of the functions can be done in Python and you should learn it in a formal database class. However, you should make sure you know SQL well !
And I will have at least one (basic) question on SQL in Mid-term

Some useful resources

MS SQL Server 2017 Free Edition

- https://www.microsoft.com/en-us/sql-server/sql-server-editions-express

MS SQL Sample Database

- https://docs.microsoft.com/en-us/sql/samples/adventureworks-install-configure?view=sql-server-2017

MS Azure Data Studio Download

- https://docs.microsoft.com/en-us/sql/azure-data-studio/download?view=sql-server-2017

Installer

1. MS SQL Server for Windows
   https://www.microsoft.com/en-us/sql-server/sql-server-editions-express

1. Postgres for Mac
   https://www.postgresql.org/download/macosx/

3. MySQL for Mac:
   https://dev.mysql.com/downloads/mysql/

Recap from last week

- Data warehouse (OLAP vs OLTP)
- Difference between DB vs DW
- SQL (relational database)
  - https://www.w3resource.com/sql-exercises/
- BIDM cycle
- Structured vs Unstructured Data
- ETL, Data Pipeline
- Data Engineers vs Data Scientists
- Basic Python syntax
- Make sure you can run Jupyter
- Finished chapter 1, 2, 3, and 17

SQL Query => Simple Aggregation (mean) => Simple Statistics (standard deviation) => Hypothesis Testing => Data Mining => Artificial Intelligence

A New Journey

Now we know something about Data, what's next?

Remember the Goal of Data Analytics is to find patterns

But how?

Remember in the video of Expectation vs Reality on
Correlation vs Causation

Let's being our journey

Data Analytics Research Process

Define your question or problem you want to solve

=> Make Observations, Collect Data

⇒ Identify possible important factors (Features, Attributes)

⇒ Test whether the factors are important or not

⇒ Continue to find important factors

⇒ Until you feel you think you got it!

Common Dataset Sources

- UCI Machine Learning Repository (UCI Machine Learning Repository)
  - One of the oldest data source
- Kaggle (https://www.kaggle.com/)
  - One of the site that every data scientist must visit
- New York Open Data (https://opendata.cityofnewyork.us/)
  - Real dataset, good source for geographical related data
- Amazon AWS Open Data (https://aws.amazon.com/opendata/)
- Open Data for Nonprofit Research (https://lecy.github.io/Open-Data-for-Nonprofit-Research/)

Common Dataset

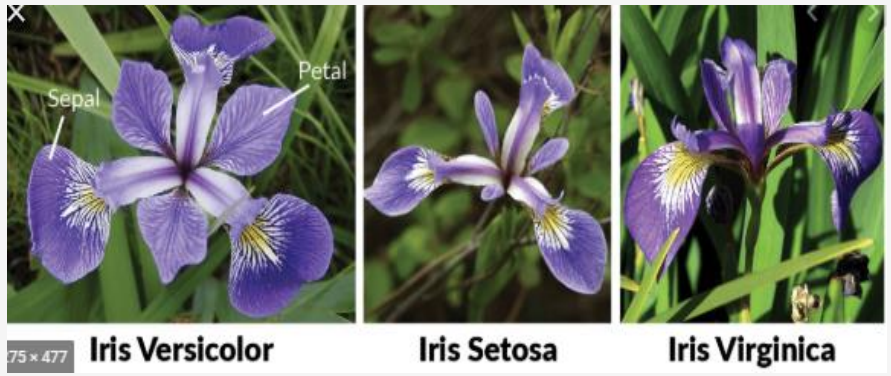
Datasets throughout the course

- Iris (https://archive.ics.uci.edu/ml/datasets/iris)
- US Housing data (https://www.kaggle.com/aariyan101/usa-housingcsv/version/1)
- Boston Housing data (https://archive.ics.uci.edu/ml/machine-learning-databases/housing/) or from scikit-learn
- https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data
- https://www.kaggle.com/c/boston-housing
- Titanic (https://www.kaggle.com/c/titanic/data)
- Adult Income Data from Census https://archive.ics.uci.edu/ml/datasets/adult
- UCI Air Quality (https://archive.ics.uci.edu/ml/datasets/Air+quality)

Common Dataset

Many common dataset can be loaded directly from the Seaborn library package

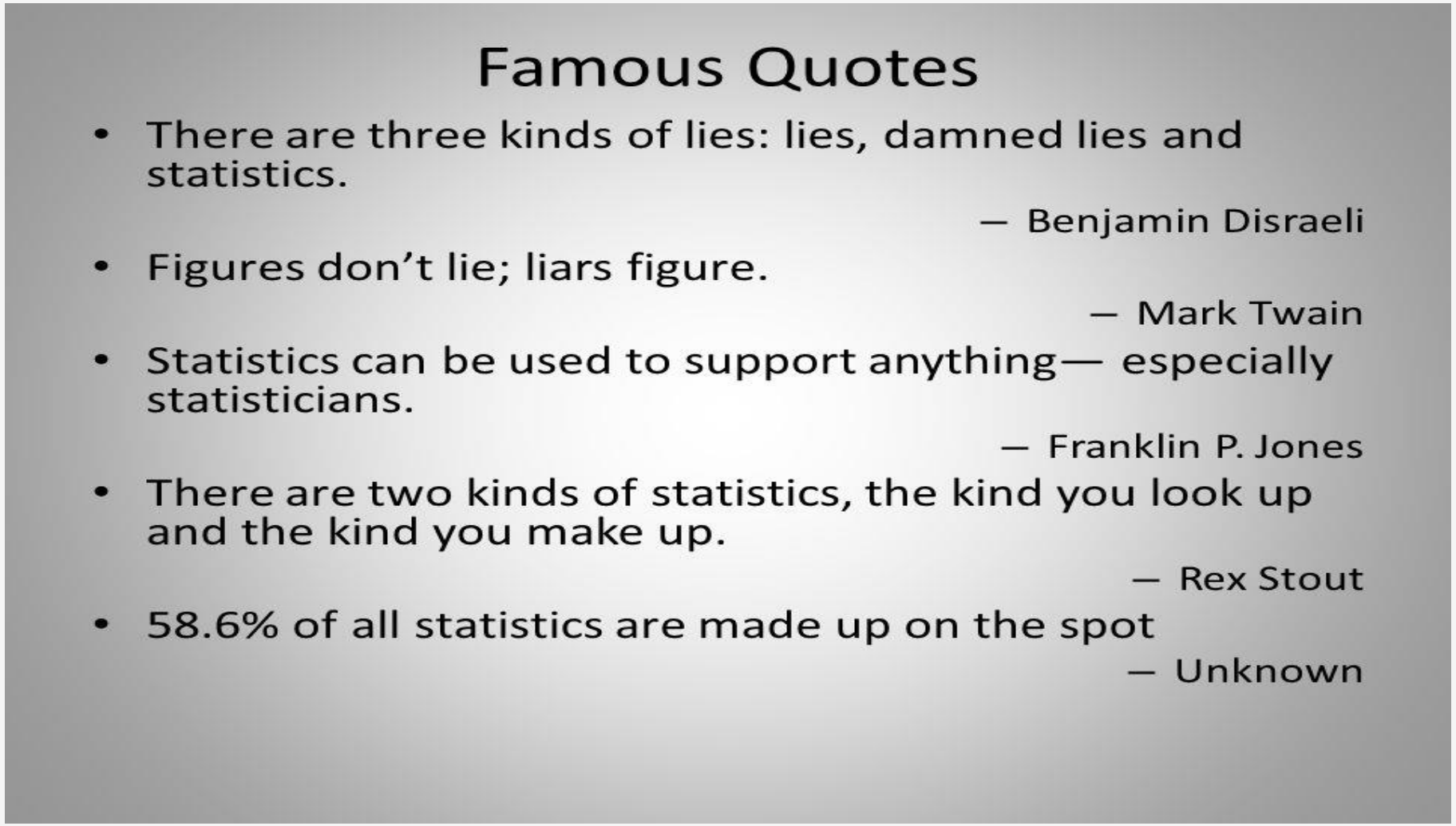
- Import seaborn as sns
- Iris = sns.load_datasets('iris')
- mpg
- Tips
- Titanic



Common Dataset

Learning by doing

Probability and Statistic Review



Then why do we still care?

A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician". – Josh Wills on Quora

http://www.mastersindatascience.org/careers/data-scientist/

Why Statistics Again?

- The world is not deterministic, in other words we need to deal with randomness
- We only live once. What we observe is only ONE realization of many possibility. In another universe, you may be the professor while I may be the student
- However, if there are some underlying truth (such as the sun will always rise from the east or a human being will not grow more than 8 feet), you will see pretty much the same thing again and again if you can make multiple observations from the same underlying mechanism.
- Observations will have a lot of noise. The Signal to Noise ratio will be extremely important.
- Question to ask is whether the conclusion you draw from your observation is statistically significant.

Importance of Statistics

- Cannot just say I am a genius. I know what I am doing. Believe me!
- Statistics is the language of Randomness
- Can only qualify your statements with certain probability
- Interested only in conclusions that are Statistically Significance
- This is what Data Analytics is all about.
- Science behind drawing meaningful conclusions from observations

In fact … Data science first appeared in a statistics paper

In 2001, William S. Cleveland published a research paper that coined the term "Data Science" the first time

Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics

Article *in* International Statistical Review 69(1) · March 2001 *with* 410 Reads

An action plan to enlarge the technical areas of statistics focuses on the data analyst. The plan sets out six technical areas of work for a university department and advocates a specific allocation of resources devoted to research in each area and to courses in each area. The value of technical work is judged by the extent to which it benefits the data analyst, either directly or indirectly. The plan is also applicable to government research labs and corporate research organizations. 1 Summary of the Plan This document describes a plan to enlarge the major areas of technical work of the field of statistics. Because the plan is ambitious and implies substantial change, the altered field will be called "data science." The focus of the plan is the practicing data analyst. A basic premise is that technical areas of data science should be judged by

Computer Science + data mining = Make statistics a lot more technical
= Data Science

Importance of Statistics

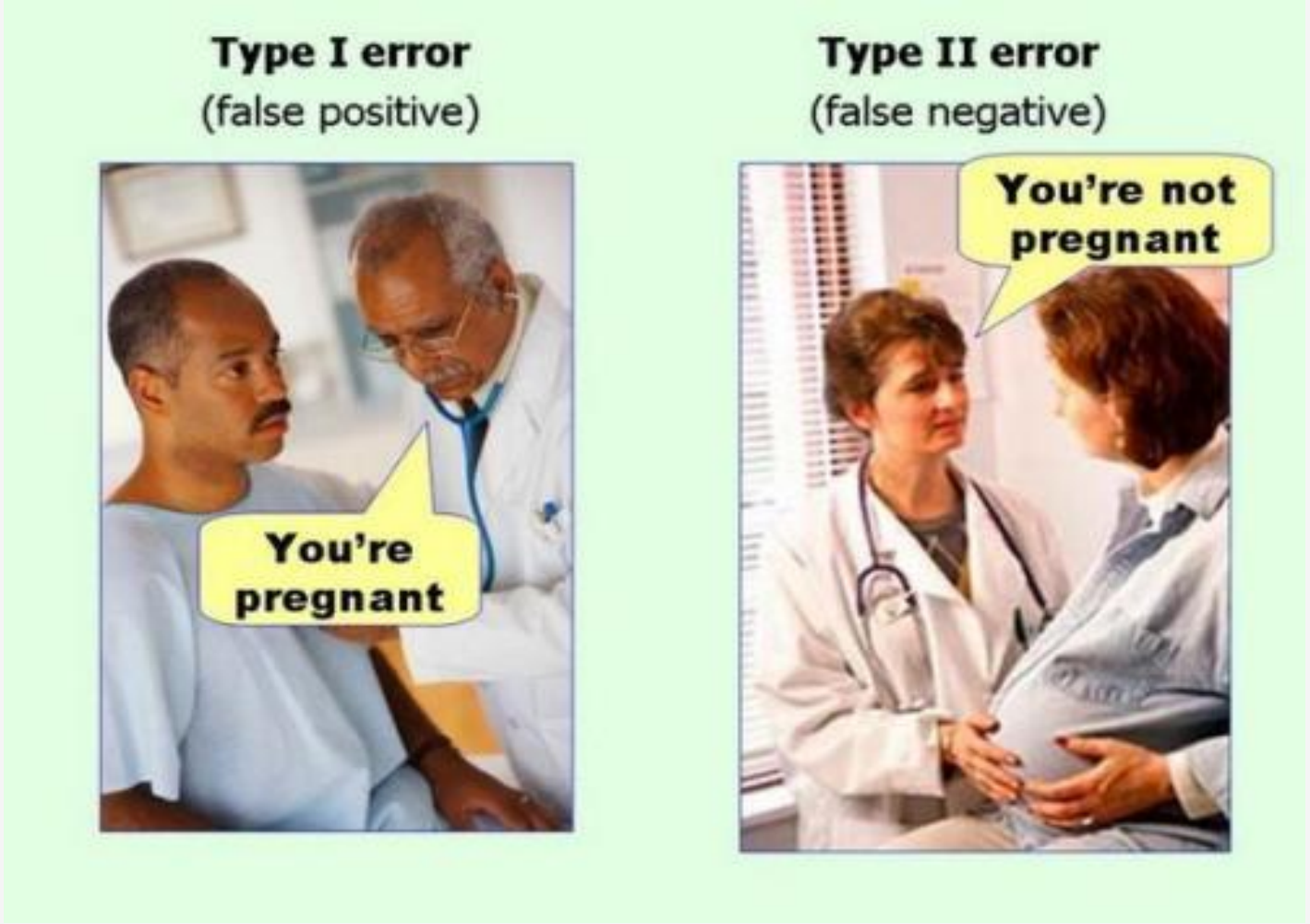
Some even prefers referring the study as Statistical Learning over Machine Learning

From Introduction To Statistical Learning:

> Statistical learning refers to a set of tools for modeling and understanding complex datasets. It is a recently developed area in statistics and blends with parallel developments in computer science and, in particular, machine learning.
>
> Since that time, inspired by the advent of *machine learning* and other disciplines, statistical learning has emerged as a new subfield in statistics, focused on supervised and unsupervised modeling and prediction.

Know your Statistics (terms you that need to understand well)

- Mean, Standard Deviation
- Distribution
- Law of large numbers
- Statistical Significance
- Survival Bias
- Bias vs Variance



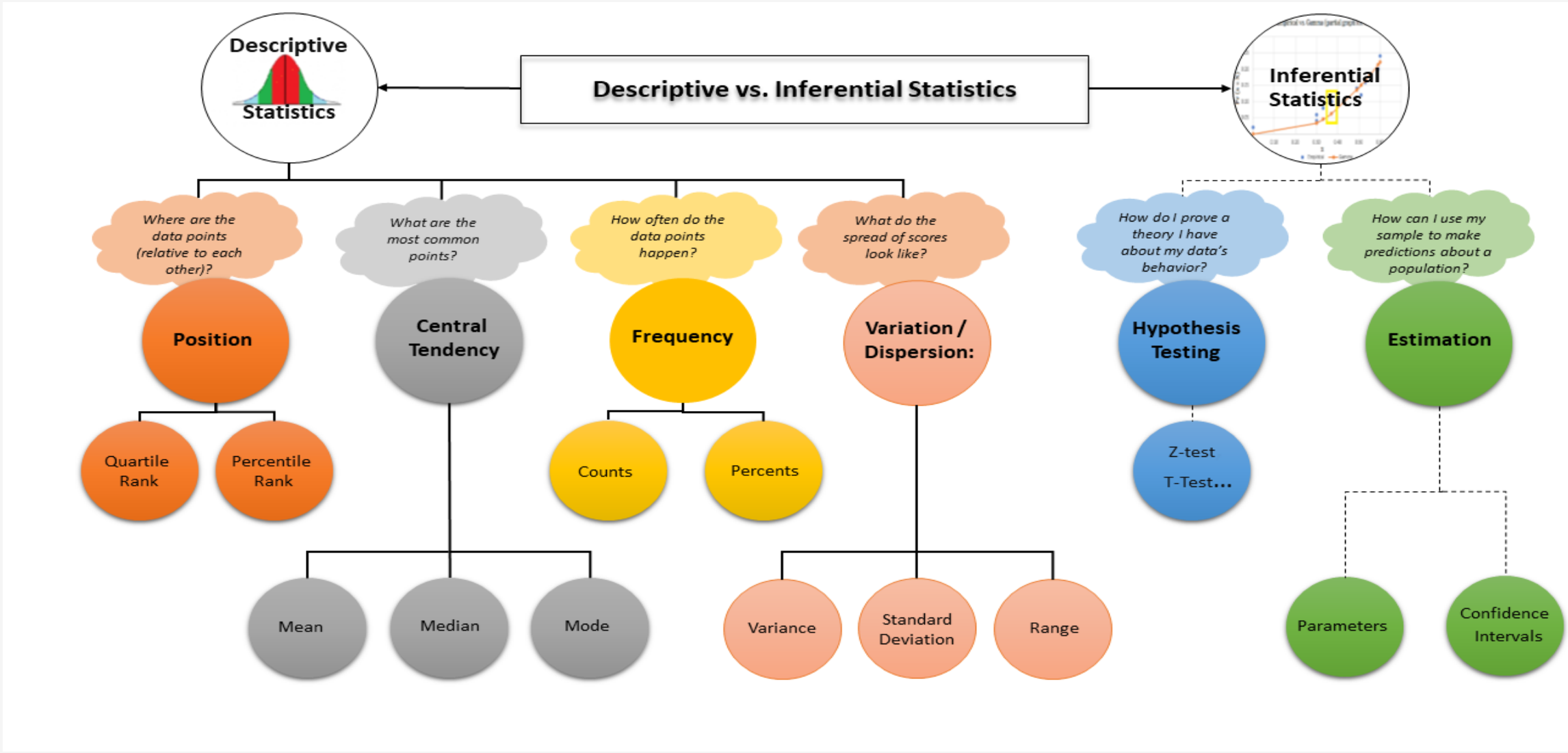
Two kind of Statistics

Descriptive Statistics vs Inferential Statistics

- Descriptive Statistics consists of organizing and summarizing data.
  - Mean, Standard Deviation, Skew, Quantile, Ranks

- Inferential Statistics (Predictive Statistics) consists of using data you have collected to form conclusions
  - Hypothesis testing
  - Estimation (Use sample mean to predict population mean)

Two kind of Statistics



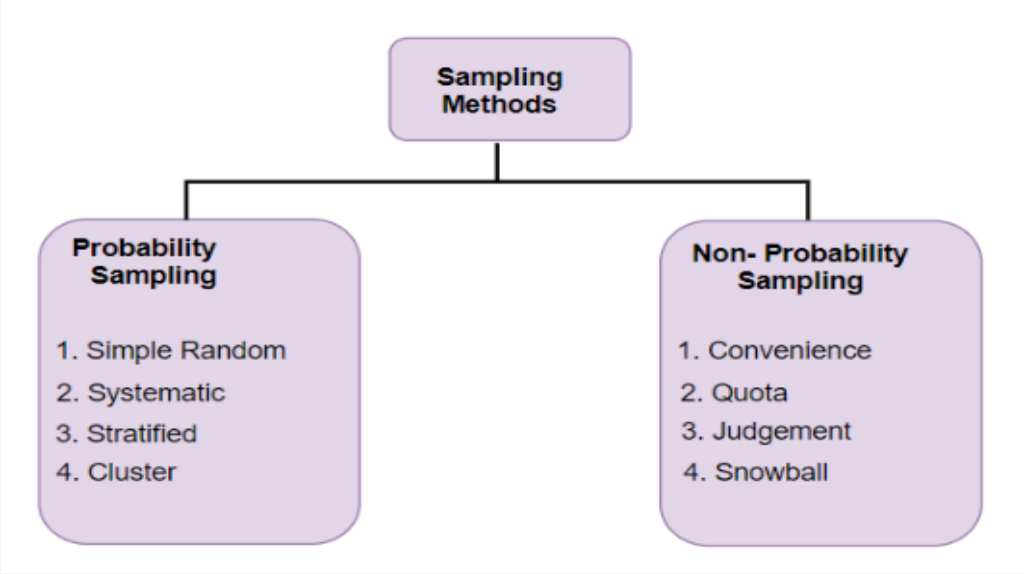
Sampling

Population vs Sample

- The population is the entire group you are interested in studying.
- A sample is a subset of the population. That is to say, it is a select group of information taken from a population.

Sampling Methods

- Random Sampling
- Stratified Sampling
- Systematic Sampling
- Convenience Sampling



https://www.analyticsvidhya.com/blog/2019/09/data-scientists-guide-8-types-of-sampling-techniques/?utm_source=feedburner&utm_medium=email&utm_campaign=Feed%3A+AnalyticsVidhya+%28Analytics+Vidhya%29

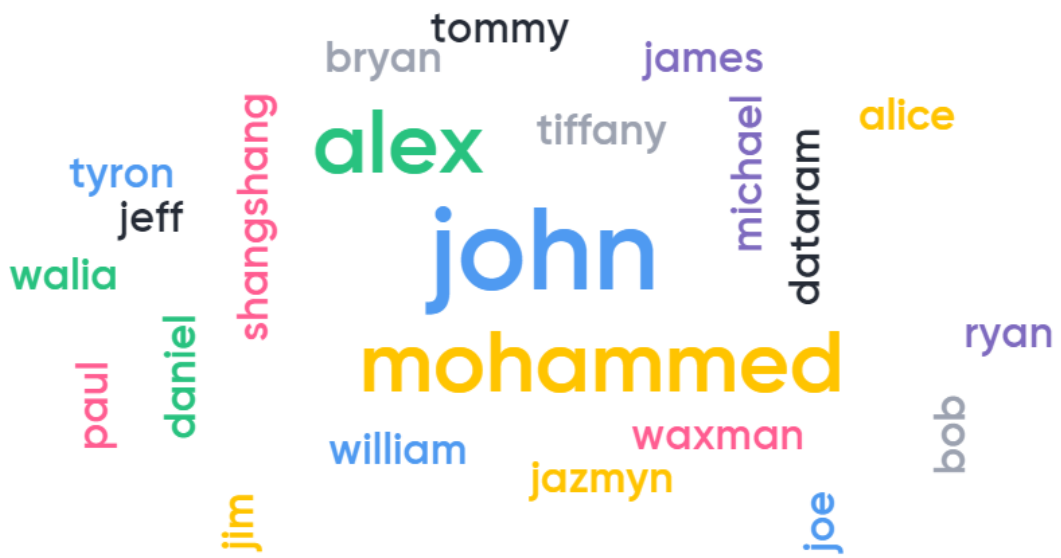
Sampling

Let’s answer the question:
What is the most common name in US?

Sampling

Go to **www.menti.com** and use the code **12 12 64**

To help answer the question of what is the most common name in US, please submit your first name.



Importance of Correct Sampling

What is the problem of what we just did?

Important consideration in Sampling

- Systematic Bias
- Survival Bias
- Size of the samples (cost)

Excellent Online Resource for Statistics Review

Assume you have the Math 241 (Probability and Statistics) pre-requisites

http://www.statisticslectures.com/topics/statistics/ is an excellent online resource for quick review

- Basics of Probability
- Discrete and Continuous Random Variables
- Probability Distribution
- Mean and Expected Value
- Law of Large Numbers
- Central Limit Theorem
- Normal Distributions
- Sampling
- Hypothesis Testing
- Type I and Type II Errors
- P-value
- One-tail tests
- Conditional Probability
- Bayes Rules

Descriptive Statistics

Central tendency refers to the measure used to determine the center of a distribution of data. It is used to find a single score that is most representative of an entire data set

Mean, Median and Mode

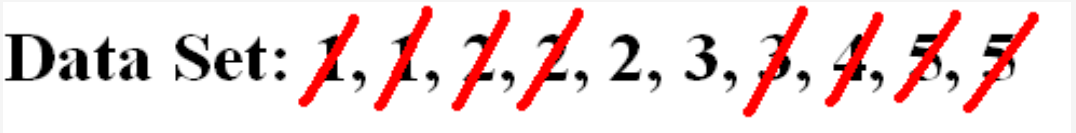
- Mean is the most common single statistics (number) to describe an entire data set.
- However, it is very sensitive to outliers.
  - Example: mean of the age of students in a college class: { 18, 19, 19, 17, 60}
- Median is the number that lies in the middle after the data set is sorted.
- Mode is simply the most frequently occurring value

Example Dataset:

1, 1, 2, 2, 2, 3, 3, 4, 5, 5

$$\bar{X} = \frac{\sum x}{n} \qquad \frac{1+1+2+2+2+3+3+4+5+5}{10} = \frac{28}{10} = 2.8$$

Mean is 2.8
Median is (2+3)/2 = 2.5
Mode is 2 because it occurs the most frequently

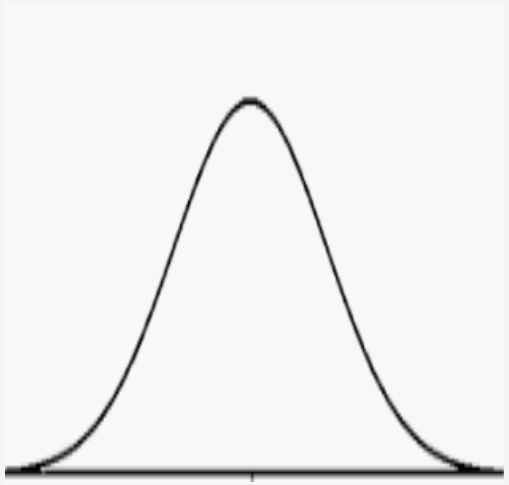


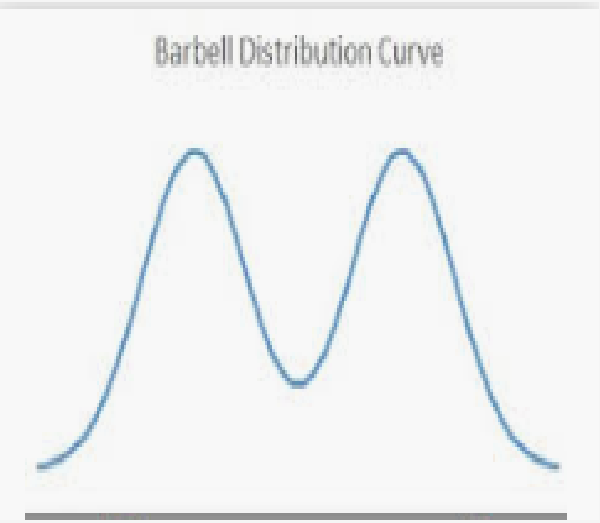
Descriptive Statistics

Two very different distribution could have the same mean and median

Example: dataset1 { 16, 18, 18, 18, 18, 18, 18, 20 }
dataset2 { 8, 8, 8, 18, 18, 28, 28, 28 }

So, we need more descriptive statistics to describe a distribution





Standard Deviation, Skew, Kurtosis

Standard Deviation or Variance

Dispersion refers to how spread out a data set is about the mean.

Variance and Standard Deviation are two measures of dispersion within a data set.

Example Dataset: {1, 2, 2, 3, 4, 5}

μ denotes the mean

$$3.35 + 0.69 + 0.69 + 0.03 + 1.37 + 4.71 = 10.84$$

$$\sigma^2 = \frac{10.84}{6} = 1.81$$

Population Variance
$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$

Population Standard Deviation
$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$

x	μ	x - μ	$(x - \mu)^2$
1	2.83	1 – 2.83 = (-1.83)	(-1.83)^2 = 3.35
2	2.83	2 – 2.83 = (-0.83)	(-0.83)^2 = 0.69
2	2.83	2 – 2.83 = (-0.83)	(-0.83)^2 = 0.69
3	2.83	3 – 2.83 = (0.17)	(0.17)^2 = 0.03
4	2.83	4 – 2.83 = (1.17)	(1.17)^2 = 1.37
5	2.83	5 – 2.83 = (2.17)	(2.17)^2 = 4.71

Second Equivalent Formula For Variance

Population Variance

$$\sigma^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{N}}{N}$$

Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{N}}{N}}$$

Figure 3.

ı this problem, N is the size of our data set(6). The other values are calculated like this:

$$\sum x^2 = 1^2 + 2^2 + 2^2 + 3^2 + 4^2 + 5^2 = 59$$

$$(\sum x)^2 = (1 + 2 + 2 + 3 + 4 + 5)^2 = (17)^2 = 289$$

After plugging in all the values, we again find a variance of 1.81, and a standard deviation of 1.35.

Sample Variance vs Population Variance

N denotes the Population size, n is the sample size

Sample Variance

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

Sample Standard Deviation

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

Population Variance	Population Standard Deviation
$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$	$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$

Sample Variance

$$s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n - 1}$$

Sample Standard Deviation

$$s = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n - 1}}$$

Population Variance

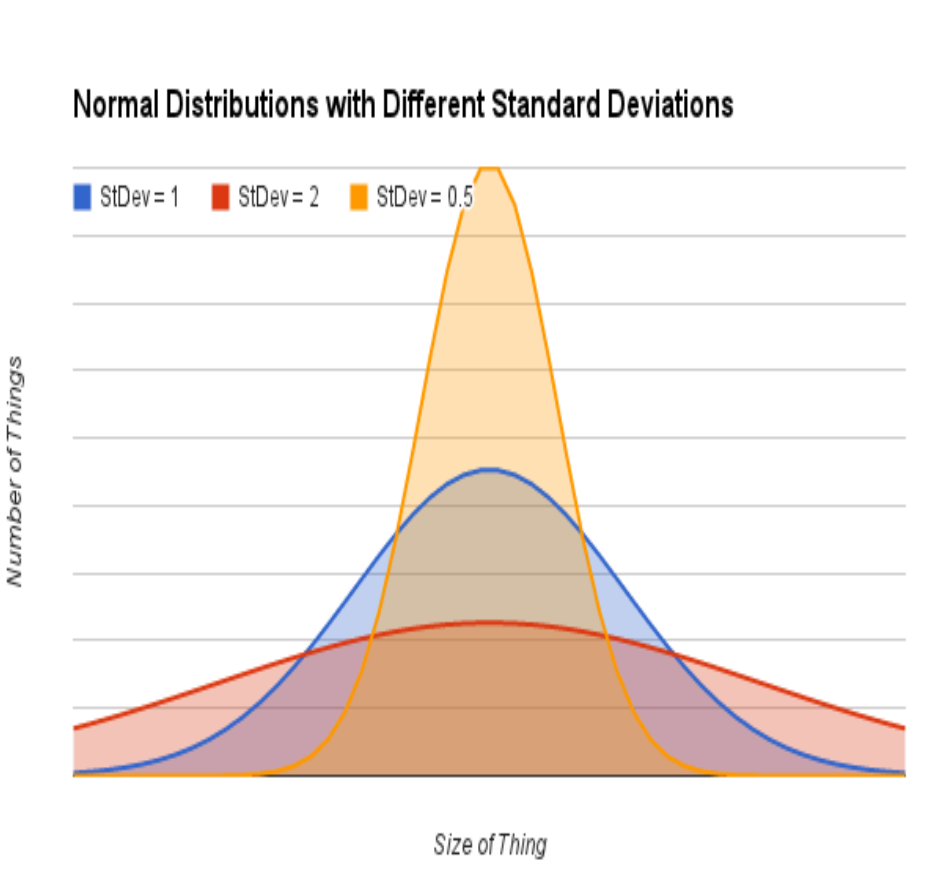
$$\sigma^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{N}}{N}$$

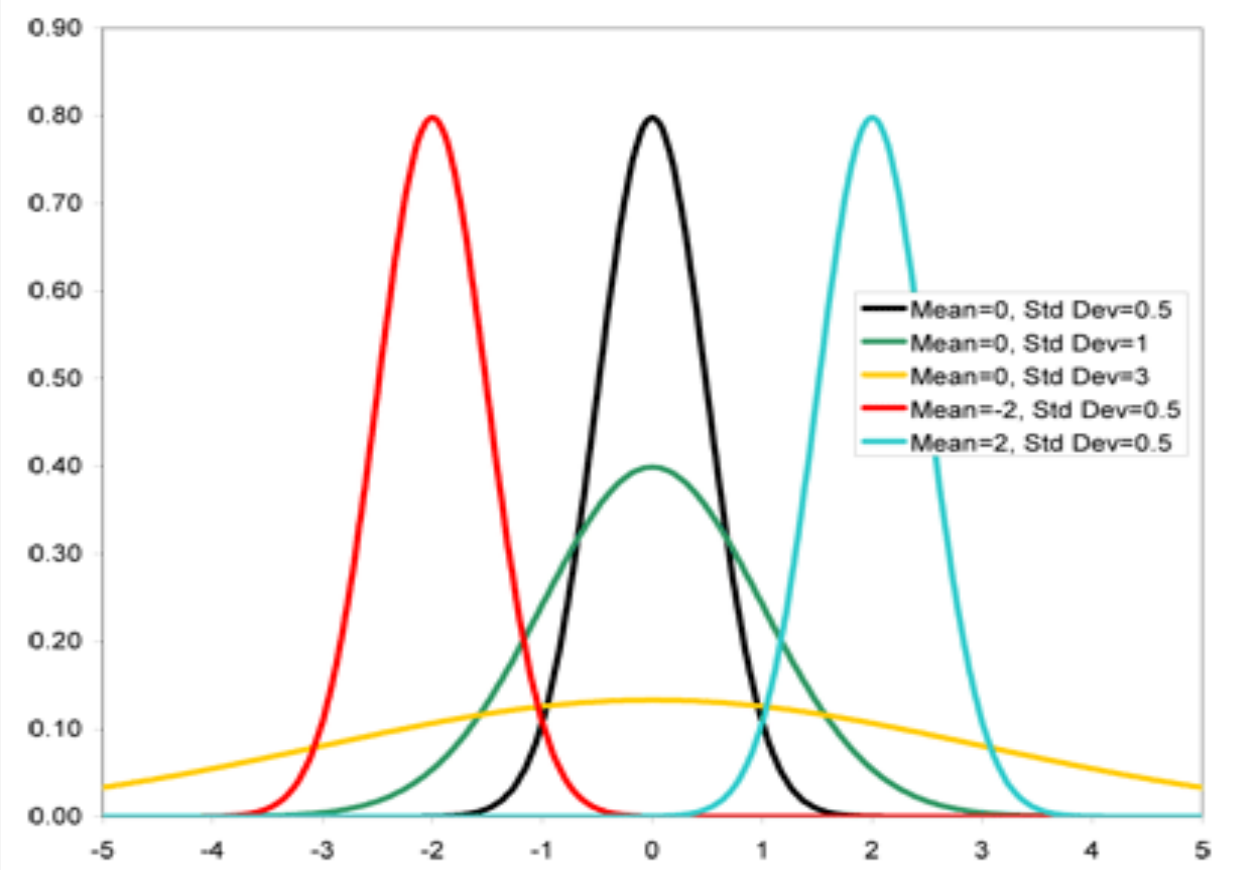
Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{N}}{N}}$$

Standard Deviation

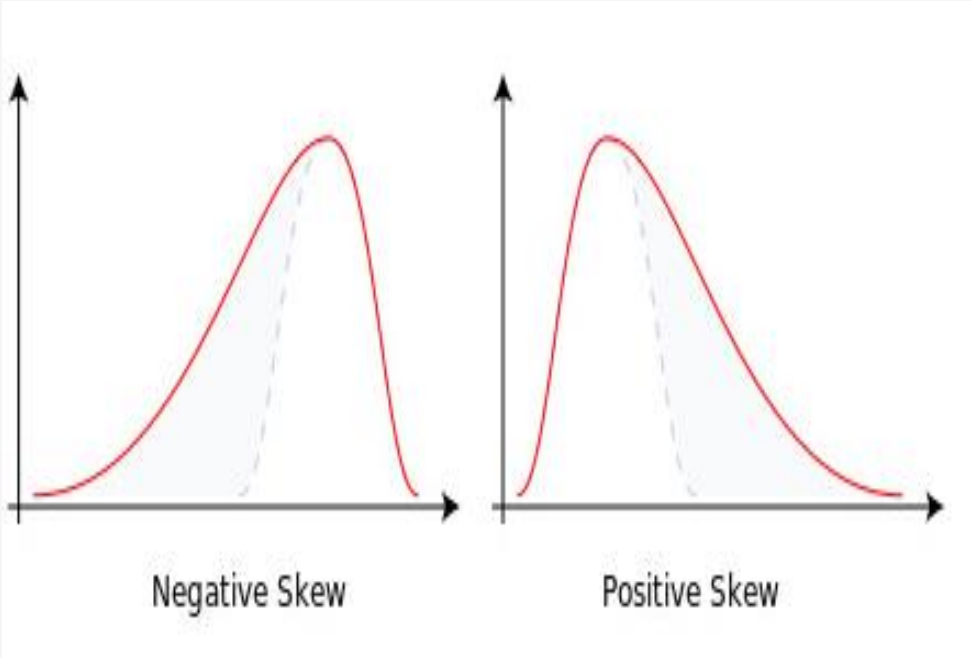
Standard Deviation as a metric to describe how “wide” or “spread-out” the distribution is.

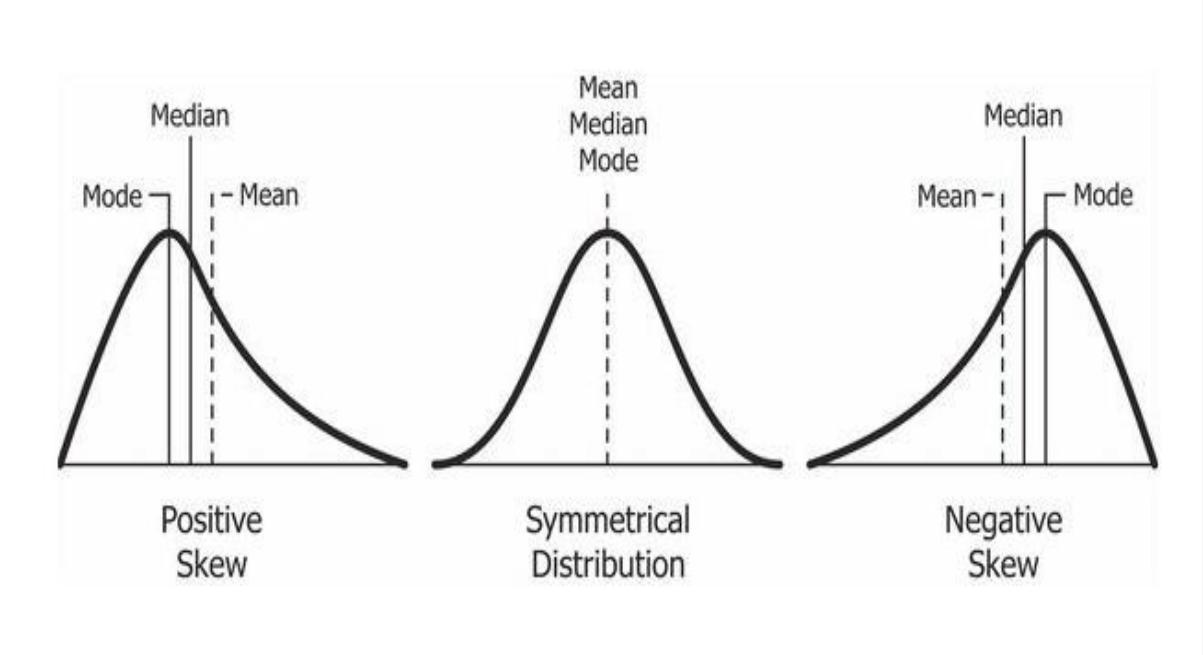




Skew

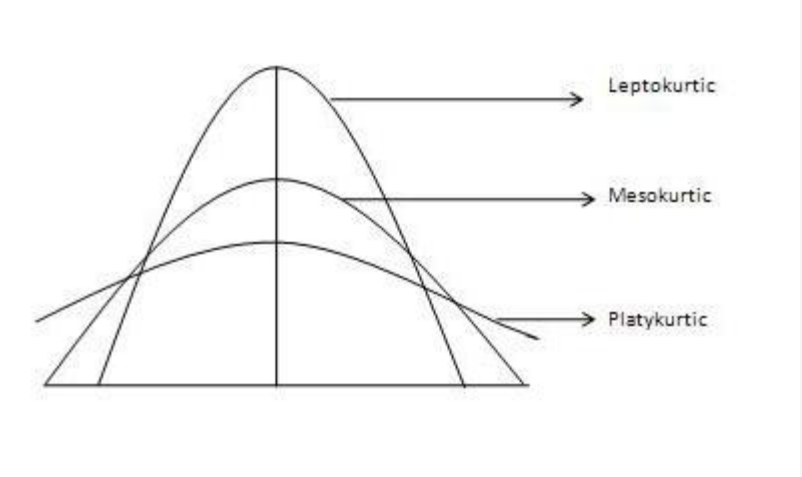
Skewness is the degree of distortion from the symmetrical bell curve or the normal curve. It measures the lack of symmetry in data distribution. It differentiates extreme values in one versus the other tail. A symmetrical distribution will have a skewness of 0.





Kurtosis

Kurtosis, on the other hand, refers to the pointedness of a peak or the tails in the distribution curve. The main difference between skewness and kurtosis is that the former talks of the degree of symmetry, whereas the latter talks of the degree of peakedness (or tailedness) in the frequency distribution.



Mesokurtic: This distribution has kurtosis statistic similar to that of the normal distribution. The standard normal distribution has a *kurtosis of three.*
Leptokurtic (*Kurtosis > 3*): tails are fatter, has more outliers. Peak is higher and sharper than Mesokurtic
Platykurtic: (*Kurtosis < 3*): tails are thinner, has less outliers than the normal distribution. The peak is lower

In Pandas the kurtosis definition is slightly different. Normal distribution has a zero Kurtosis. Leptokurtic kurtosis is > 0 and Platykurtic kurtosis is < 0

Appendix: Statistics Review

Descriptive and Inferential Statistics http://www.statisticslectures.com/topics/descriptiveinferential/

Population vs Sample http://www.statisticslectures.com/topics/samplingmethods/

Parameters, Statistics and Sampling Errors http://www.statisticslectures.com/topics/parametersstatistics/

Distribution of Sample Mean http://www.statisticslectures.com/topics/distributionsamplemean/

Mean and Expected value of a probability http://www.statisticslectures.com/topics/meanexpectedvaluediscrete/

Variance and Standard deviations http://www.statisticslectures.com/topics/variancestandarddeviationdiscrete/

Law of Large Numbers, Central Limit Theorem http://www.statisticslectures.com/topics/centrallimittheorem/

Skew and Kurtosis explained:
https://keydifferences.com/differences-between-skewness-and-kurtosis.html
https://codeburst.io/2-important-statistics-terms-you-need-to-know-in-data-science-skewness-and-kurtosis-388fef94eeaa
https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-whats-with-the-different-formulas-for-kurtosis/

Statistics Review

Watch the Recommended Statistics Lectures as much as you could

http://www.statisticslectures.com/topics/statistics

CSCI381 Data Analytics -- logistics

Instructor: Dr. Alex Pang

Email: chiuyan.pang@qc.cuny.edu

Lectures: Mon, Wed (8:00pm – 9:15pm)

Pre-requisites:

- CSCI 313 (Data Structures)
- Math 241 (Prob & Stat)

Teaching Assistant: None

Office hours: 9:15 to 9:45 pm after class

Course Objective:

At the end of this course students should

1. have a good overview of the data science professions and modern data analytics platforms.

2. have acquired expertise in using Python as his/her data analysis and model development platform

3. have developed a good analytical mindset in drawing insights on data and making recommendations

4. have understood some of the most common machine learning techniques and feel comfortable in pursuing more advanced skill set in machine learning areas.

CSCI381 Data Analytics -- logistics

Instructor: Dr. Alex Pang

Email: chiuyan.pang@qc.cuny.edu

Lectures: Mon, Wed (8:00pm – 9:15pm)

Pre-requisites:

- CSCI 313 (Data Structures)
- Math 241 (Prob & Stat)

Teaching Assistant: None

Office hours: 9:15 to 9:45 pm after class

Course Objective:

At the end of this course students should

1. have a good overview of the data science professions and modern data analytics platforms.

2. have acquired expertise in using Python as his/her data analysis and model development platform

3. have developed a good analytical mindset in drawing insights on data and making recommendations

4. have understood some of the most common machine learning techniques and feel comfortable in pursuing more advanced skill set in machine learning areas.

5. be able to present themselves as smarter than they actually are

Course objective – Make yourselves looks smart

Imagine your boss ask you to join a meeting where either your colleague or some sale representative or research to present their “super-wonderful-one-of-a-kind” great Data Mining model that your company cannot resist not to use.

And you have absolutely no ideas what are the theory behind the models

What can you still say or ask that can impress others and convince them that you are an expert that they need your advice on project?

First most useful word your will learn from me -- Bias

How did you collect the data? What are your sampling methods? ….

So, your data seems to have XXX bias.

Example: does your data include enough Latinos, Female, etc and etc

Does your data have Bias?

Remember Data Mining starts with Data

not Mining

https://cmotions.nl/en/5-typen-bias-data-analytics/

Review of Standard Deviation, Skew and Kurtosis

Standard Deviation
large SD => wide distribution => heterogeneity
Small SD => narrow distribution => homogeneity

Skew
Positive => lots of bigger values
Negative => lots of smaller values

Kurtosis
Positive => More outliers than normal distribution
Negative => Less outliers than normal distribution

The height distribution taken from Computer Science class in Queen College will have a mean __similar__ (higher/lower/similar) than the whole college and a _____ (positive/zero/negative) skews

The height distribution taken from the basketball Team in Queen College will have a mean _higher___ (higher or lower) than the whole college and a ___positive or zero_____ (positive/zero/negative) skews

The height distribution taken from Computer Science class in Queen College will have a mean _higher___ (higher or lower) than the whole college and ____positive_____ (positive/zero/negative) skews <u>if we know many are also in the basketball Team</u>

Questions

What are the factors that drive house prices?

Questions

What are the factors that drive house prices in a city?

Mortgage Rates
Unemployment Rates
Local School performance
…

Questions

How would you determine which factors are really important in 5 minutes (ie without developing any models)?

Covariance and Correlation

Covariance measures the linear relationship between two variables.

- **Positive covariance**: Indicates that two variables tend to move in the same direction.
- **Negative covariance**: Reveals that two variables tend to move in inverse directions

Covariance can range from negative infinity to positive infinity.

Correlation is the scaled measure of covariance. It is dimensionless. In other words, the correlation coefficient is always a pure value and not measured in any units.

Correlation is between -1 and +1

ρ(X,Y) – the correlation between X and Y
Cov(X,Y) – the covariance between X and Y
σ_X – the standard deviation of X
σ_Y – the standard deviation of Y

$$\text{Cov}(X, Y) = \frac{\sum(X_i - \bar{X})(Y_j - \bar{Y})}{n}$$

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

https://corporatefinanceinstitute.com/resources/knowledge/finance/covariance/

Covariance and Correlation

Pearson product moment correlation

The Pearson correlation evaluates the linear relationship between two continuous variables. A relationship is linear when a change in one variable is associated with a proportional change in the other variable.

For example, you might use a Pearson correlation to evaluate whether home price increase in a city is related to the unemployment rate in that area.

Spearman rank-order correlation

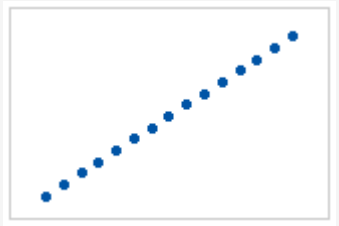
The Spearman correlation evaluates the monotonic relationship between two continuous or ordinal variables.
In a monotonic relationship, the variables tend to change together, but not necessarily at a constant rate.
The Spearman correlation coefficient is based on the ranked values for each variable rather than the raw data.

Spearman correlation is often used for ordinal variables. For example, you might use a Spearman correlation to study how the order in which employees complete a test exercise is related to the months they have been employed.

In a scatterplot, Pearson Correlation coefficients measure linear relationship while Spearman is more concerned on whether the relationships is monotonic or not.

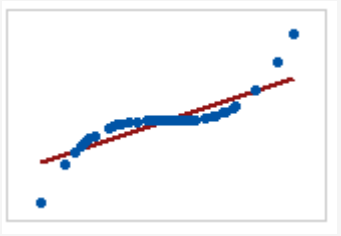
Pearson vs Spearman Correlation

Fig 1



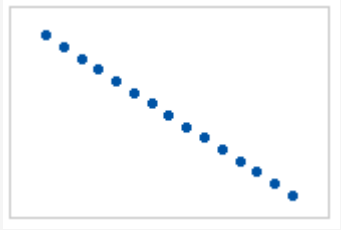
Pearson: +1
Spearman: +1

Fig 2



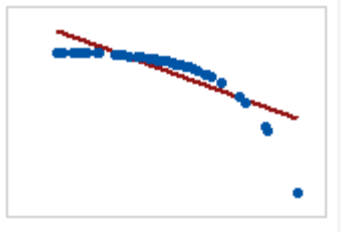
Pearson: ?
Spearman: ?

Fig 3



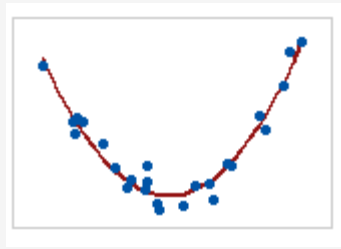
Pearson: -1
Spearman: -1

Fig 4



Pearson: ?
Spearman: ?

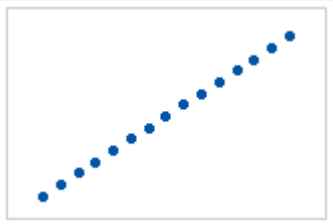
Fig 5



Pearson: ?
Spearman: ?

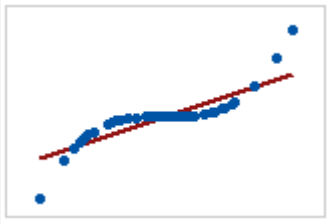
Pearson vs Spearman Correlation

Fig 1



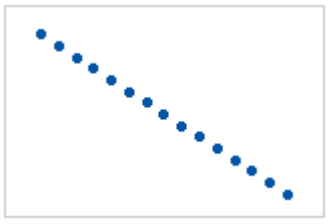
Pearson: +1
Spearman: +1

Fig 2



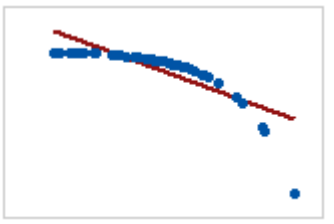
Pearson: +0.85
Spearman: +1

Fig 3



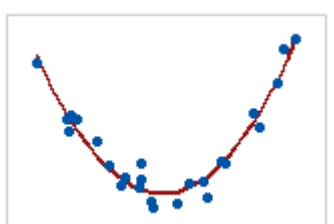
Pearson: -1
Spearman: -1

Fig 4



Pearson: -0.85
Spearman: -1

Fig 5



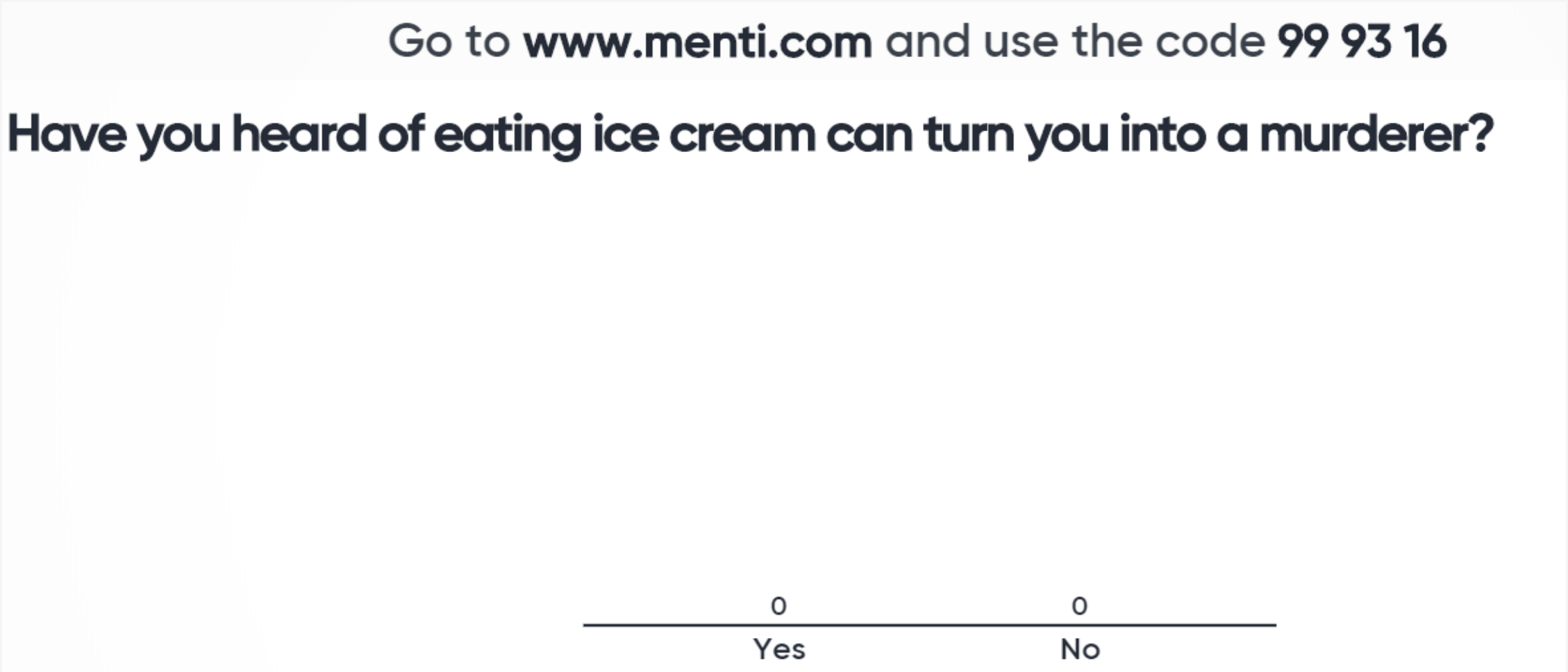
Pearson: 0
Spearman: 0

Zero correlation does not mean the variables are independent

Low correlation does not mean there is no dependence between two variables

https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/supporting-topics/basics/a-comparison-of-the-pearson-and-spearman-correlation-methods/

Questions


Go to www.menti.com and use the code 99 93 16
Have you heard of eating ice cream can turn you into a murderer?
0
Yes
0
No

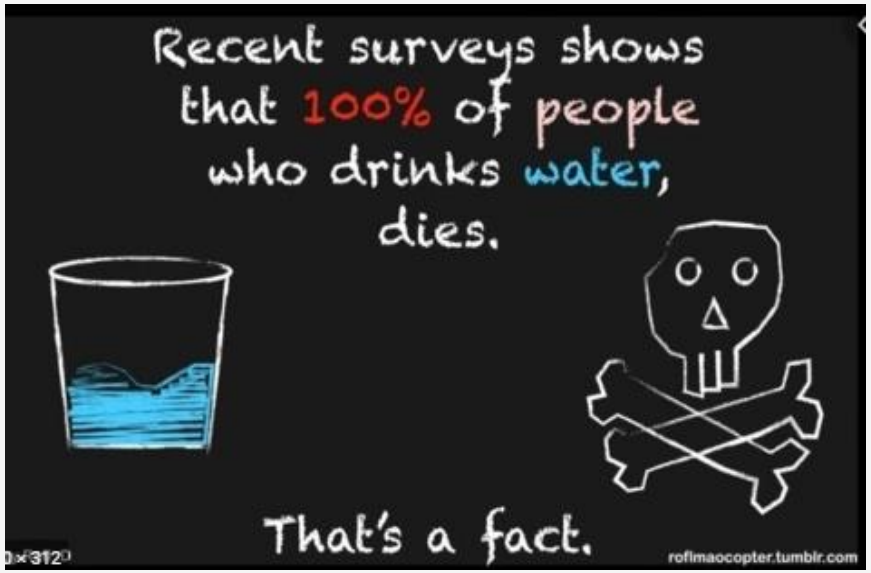
Correlation and Causation

Causation will lead to high correlation, but high correlation may not necessarily imply causation relationship

Classic Example: Murder rates goes up when ice cream sales go up

The rates of violent crime and murder have been known to jump when ice cream sales do. But, presumably, buying ice cream doesn't turn you into a killer (unless they're out of your favorite kind?)

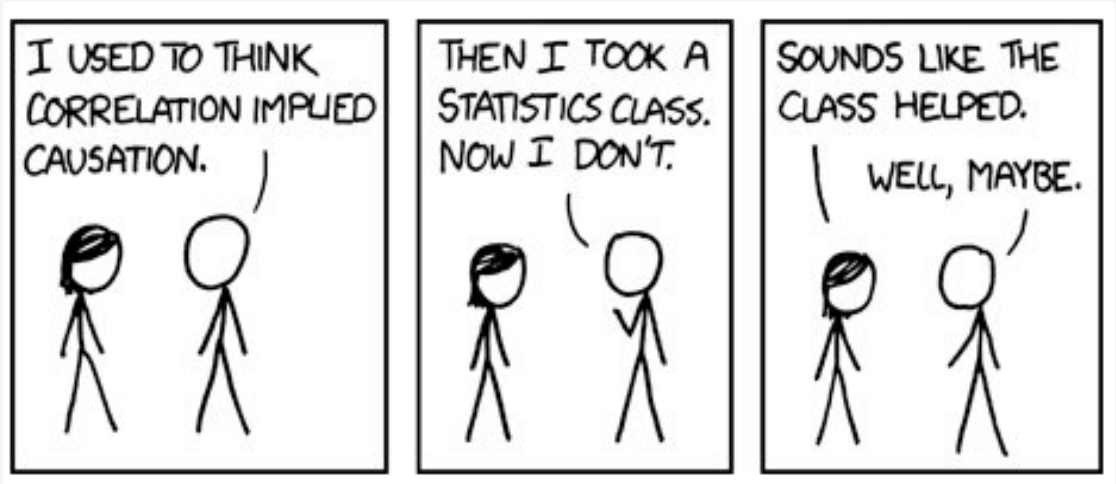
But, correlation is still one good tool to identify driving factors.



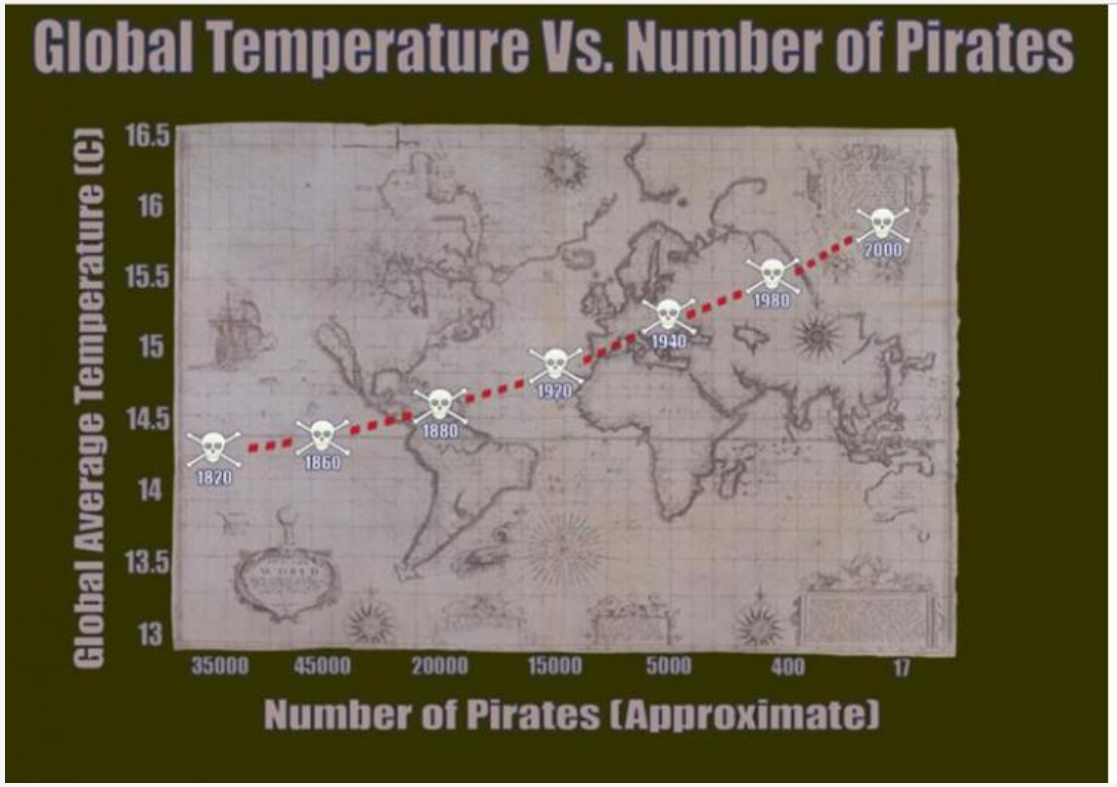
https://science.howstuffworks.com/innovation/science-questions/10-correlations-that-are-not-causations.htm

https://www.georanker.com/correlation-vs-causality-differences-and-examples

Correlation and Causation

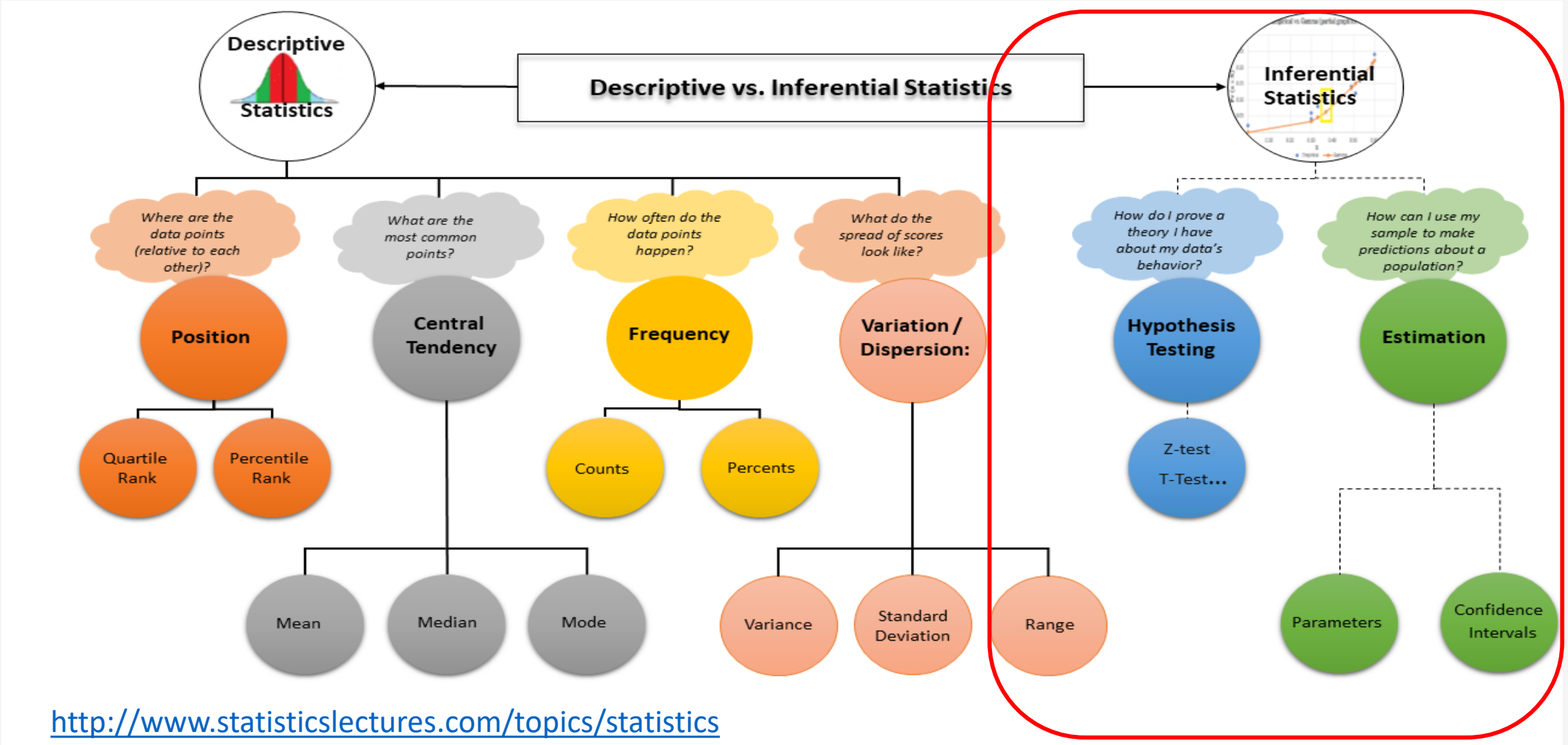


Global Warming caused by Lack of Pirate



https://www.sisense.com/blog/global-warming-caused-lack-pirates-bad-graph-lessons/

Inferential Statistics / Predictive Statistics



http://www.statisticslectures.com/topics/statistics

Inferential Statistics – making estimations of the population from samples

Parameters: A characteristic that describes a population is called a parameter. Because it is often difficult (or impossible) to measure an entire population, parameters are most often estimated

http://www.statisticslectures.com/topics/parametersstatistics/

Statistic: A characteristic that describes a sample is called a statistic. Statistics are most often used to estimate the value of unknown parameters

http://www.statisticslectures.com/topics/distributionsamplemean/

- Distribution of Sample Mean: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

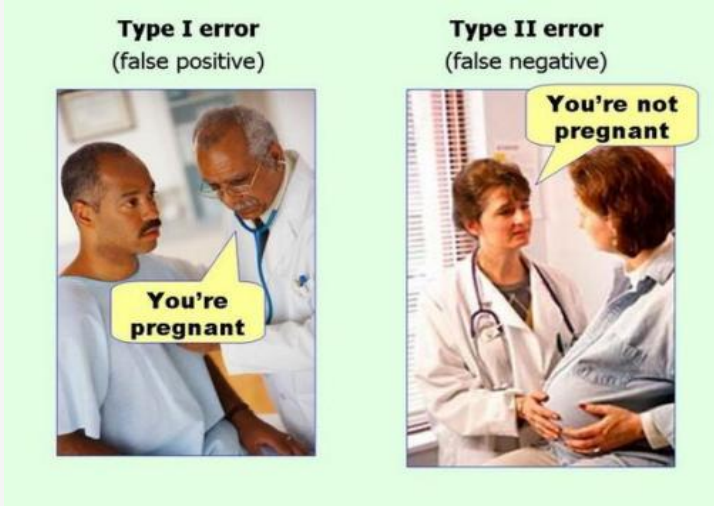
http://www.statisticslectures.com/topics/centrallimittheorem/

- The Central Limit Theorem: Independent of the actual distribution of the population, if we take a big enough sample size, when we repeat taking sample again and again, the distribution of the sample mean follows a normal distribution.

- That is why we can often use the normal distribution behind hypothesis testing

Hypothesis Testing

- Type I error (false positive, too excited to claim something non-existence)
- Type II error (false negative, failed to realize something real is going no)



- Null Hypothesis (nothing to see, life is as usual)
- Alternate Hypothesis (something is going on)

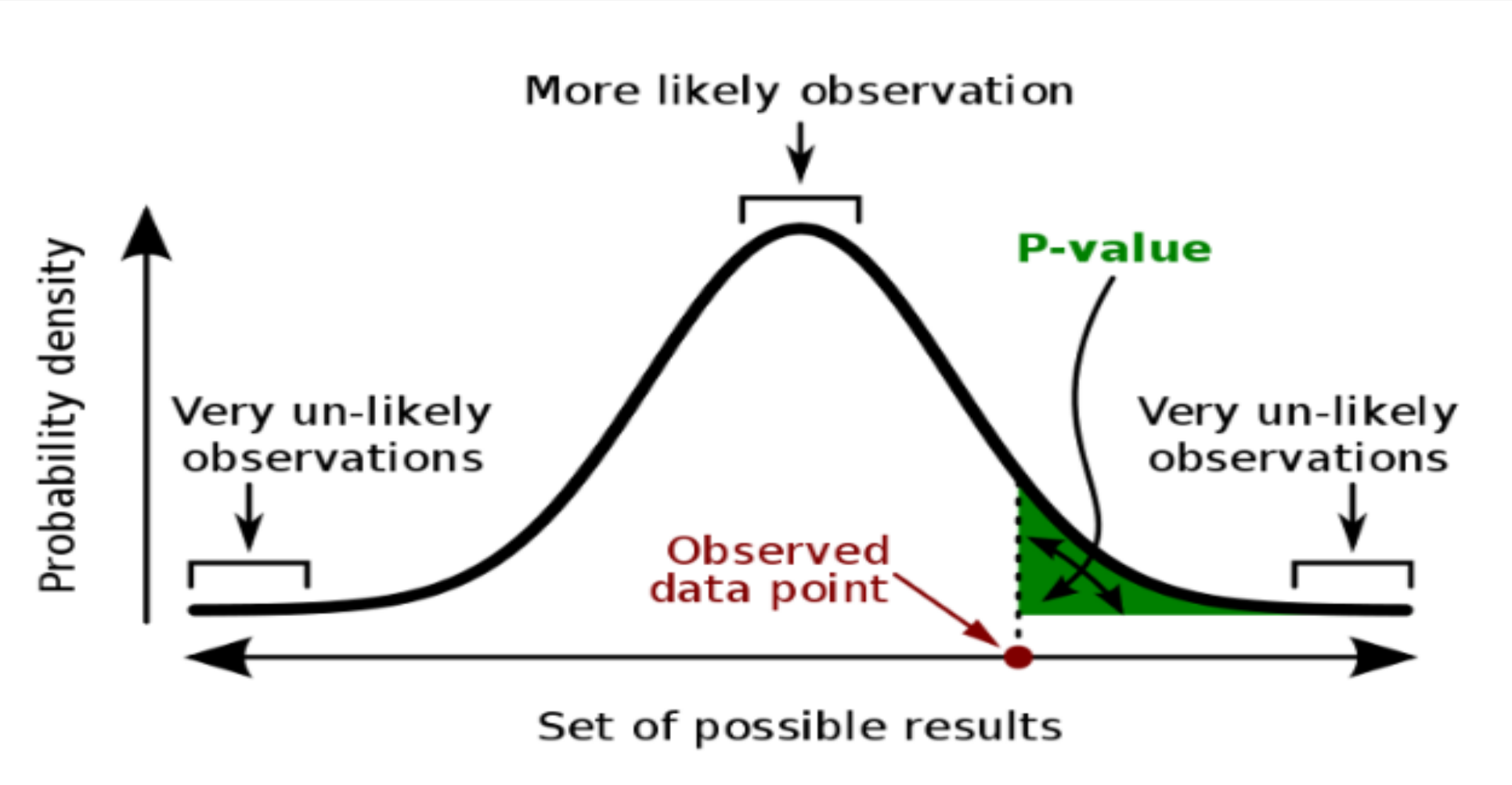
1. Define Null and Alternative Hypotheses
2. State Alpha
3. State Decision Rule
4. Calculate Test Statistic
5. State Results
6. State Conclusion

http://www.statisticslectures.com/topics/typeonetypetwoerrors/

http://www.statisticslectures.com/topics/onetailtwotail/

http://www.statisticslectures.com/topics/onesamplez/

P-value and Confidence interval



https://www.youtube.com/watch?v=vemZtEM63GY

Online Statistics Review

Watch this online Statistics Lectures as much as you can

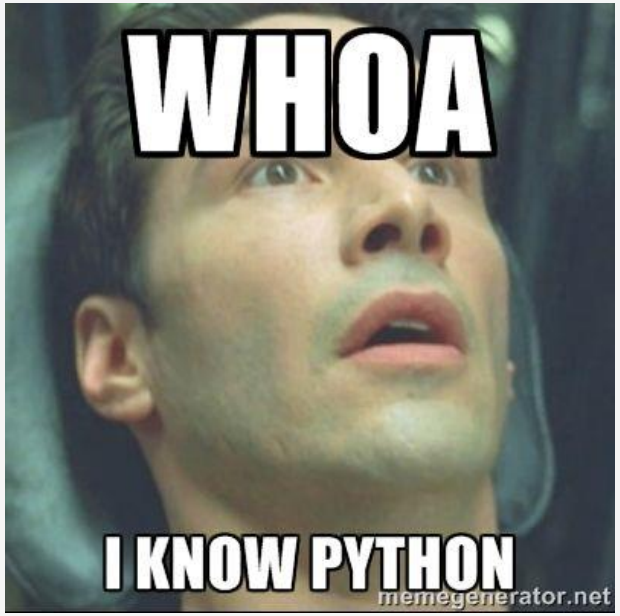
- http://www.statisticslectures.com/topics/statistics/

TO-DO Task

Read Chapter 4 Data Mining of the Textbook

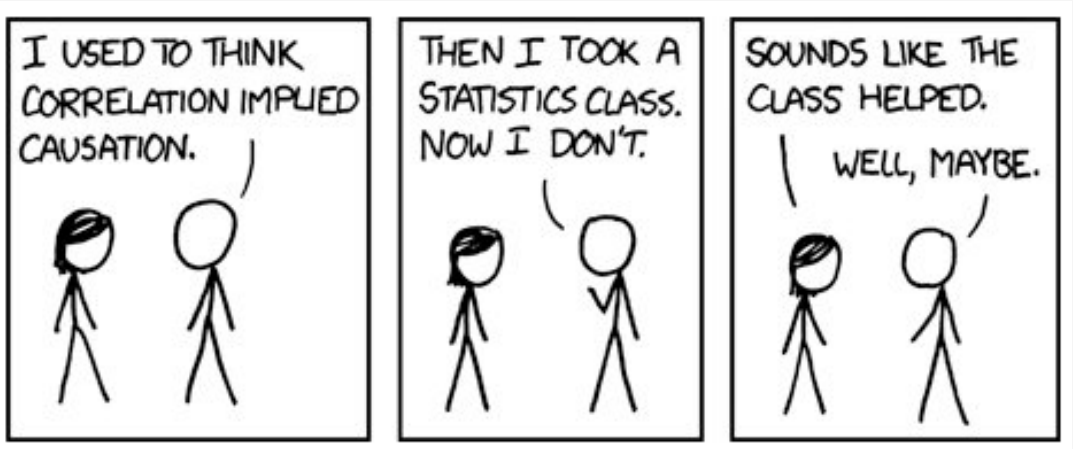
(first part of the chapter, especially on data cleansing and preparation)

Recap


WHOA
I KNOW PYTHON
memegenerator.net


DataFrame?


ONE DOES NOT SIMPLY
SAY NO TO PANDA


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Use SQL to retrieve data

Use SQL to calculate average and group by

Use Pandas to load data

Use Pandas to calculate average and group by

Know to ask the right questions

And filter out outliers and missing values

Know Mean and Standard Deviation

Know positive Skew vs negative Skew

Know Kurtosis is related to outliers

Know how to pronounce
Leptokurtic
Mesokurtic
Platykurtic

Know Correlation is between -1 and 1

Know ice cream does not make you a murderer

And zero correlation does not mean independence

Know the difference between Pearson correlation and Spearman correlation

Again One More Time

Know SQL

https://www.w3resource.com/sql-exercises/

Know Pandas

http://www.datacamp.com

Know your Statistics

http://www.statisticslectures.com/topics/statistics

Exploratory Data Analysis (EDA)

Before building any sophisticated model, we need to do EDA first.

EDA is the first step in your data analysis. You take a broad look at patterns, trends, outliers, unexpected results and so on in your existing data, using visual and quantitative methods to get a sense of the story this tells. You're looking for clues that suggest your logical next steps, questions or areas of research.

- Dataset summary
- Missing data
- Basic Statistics
- Basic graphs
- Basic relationship

https://www.sisense.com/blog/exploratory-data-analysis/

Some of the tasks in EDA

- Spotting mistakes and missing data
- Mapping out the underlying structure of the data
- Identifying the most important variables
- Listing anomalies and outliers
- Test a hypotheses / check assumptions related to a specific model
- Establish a parsimonious model (one that can be used to explain the data with minimal predictor variables)
- Estimate parameters and figuring out the associated confidence intervals or margins of error.

Data Cleansing (Garbage in Garbage out, 80/20 rules)

Most data scientists spend only 20 percent of their time on actual data analysis and 80 percent of their time finding, cleaning, and reorganizing huge amounts of data

- Duplicate data removed
- Missing values need to be filled (or handled)
- Data elements should be comparable (similar units)
- Continuous values may need to be binned
- Outlier data need to be removed
- Ensure dataset has no systematic biases for the phenomena under analysis
- Be sure dataset has enough information density

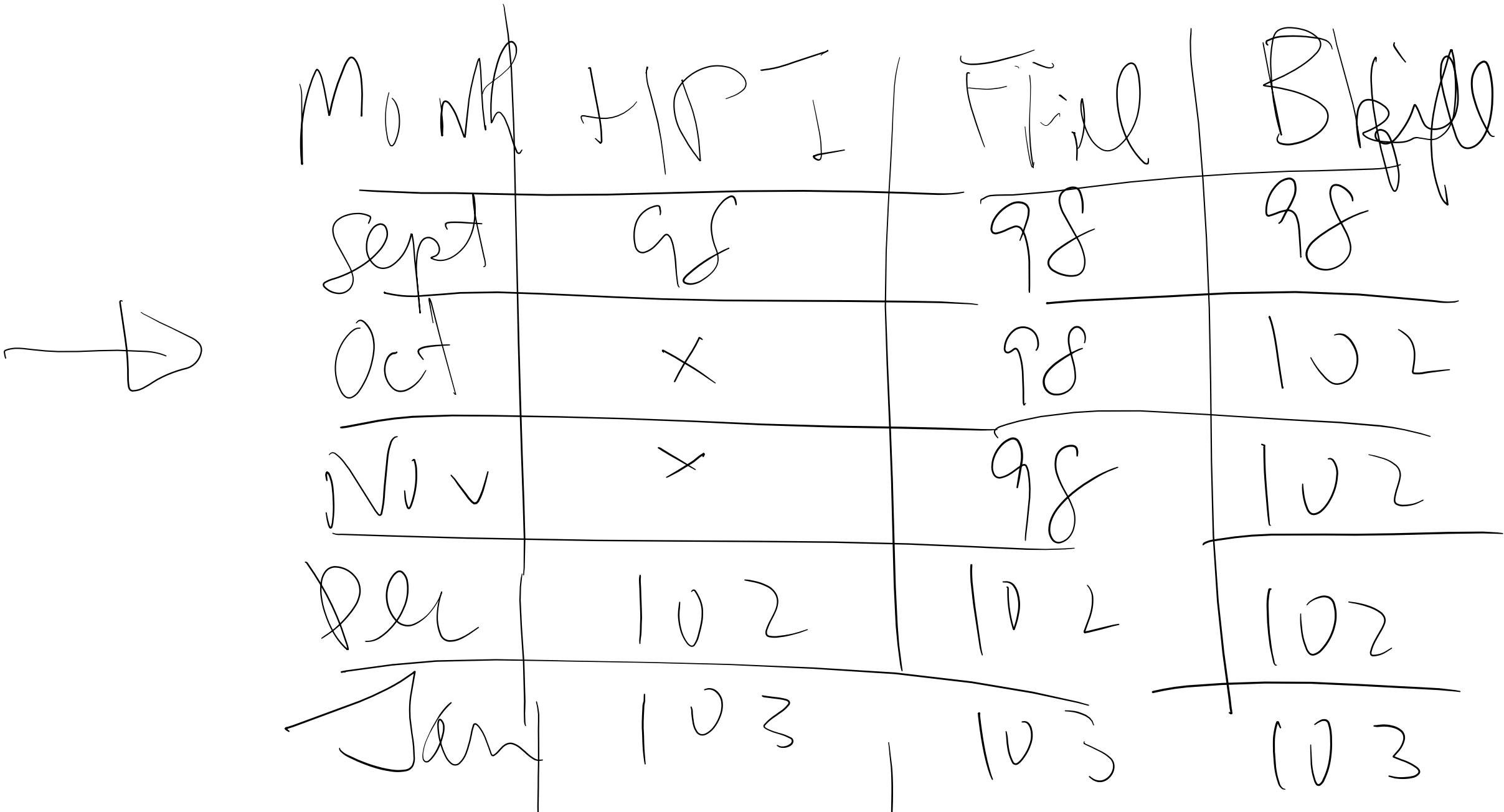
How to handle missing values

- Deletion
  - Pro: most easy way and no ambiguity
  - Con: can apply only if we have enough data, may introduce systematic bias

- Imputation
  - Use Mean, Median or Mode
    - Pro: Easy to understand, ok most of the time
    - Con: may introduce systematic bias

  - For Time Series data,
    - Use last observed data (forward fill) (df.fillna(method='ffill'))
    - Use latest available data (backward fill) (df.fillna(method='bfill'))

  - More advanced method such as use nearest neighbor


Month
TIP
Fin
Bill
Sept
95
98
98
Oct
×
98
102
Nov
×
98
102
Dec
102
102
102
Jan
103
103
103

How to handle missing values

There is no silver bullet

That’s why a critical mind is important

Other aspects of Exploratory Data Analysis

Ask the right questions

The goal of EDA is to explore and develop a high-level intuition and understanding of the data before we dive into any more sophisticated models

Exploratory Data Analysis

Learning by doing

Read Chapter 5 (Data Visualization) of the Textbook

Data Visualizations

Pictures worth a thousand word

Human brain is good at understanding information conveyed through visualizations than just text and numbers

Principle of good Data Visualizations

- Serve a reasonably one single clear purpose
- don’t over-complicated your graphs
- Encourage viewer to compare different pieces of data
- Induce viewer to think of the substance of data
- Avoid distorting what the data have to say
- Show the data with easy to understand scale
- Reveal the data at several levels of detail

Data Visualizations Toolbox

Libaries

- Matplot lib
- Pandas plot
- **Seaborn**
- Plotly
- Geographic plots

Type of Plots

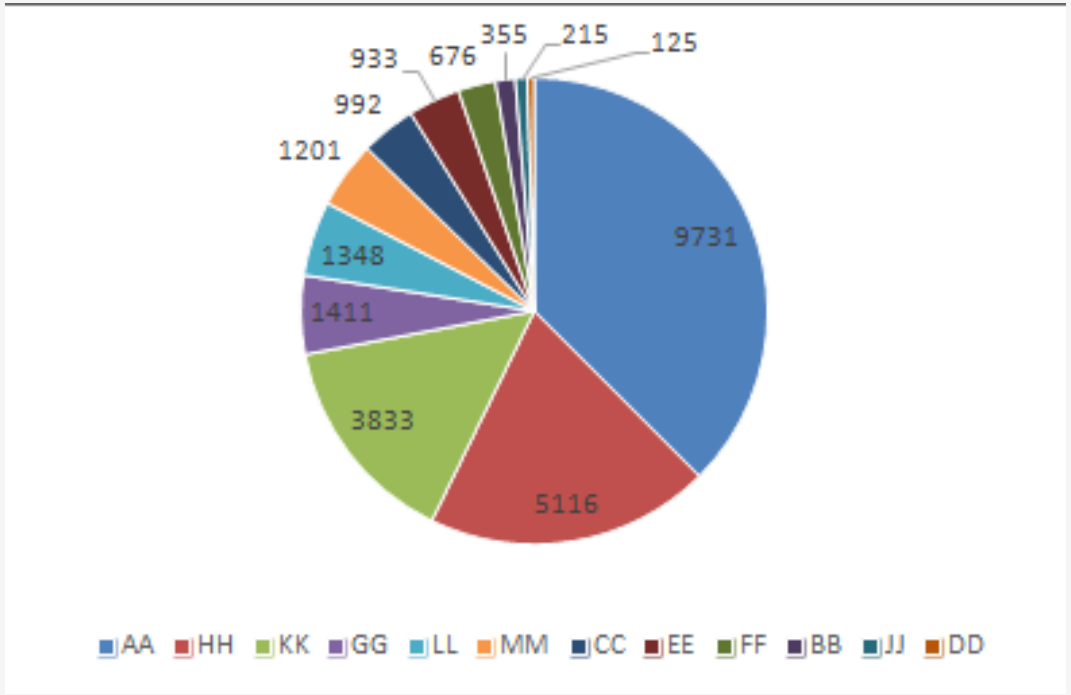
- Pie Chart / Bar Chart
- Scatter Plots
- Distribution Plots
- Box Plots
- **FacetGrid**
- Heatmap

More Advanced visualizations

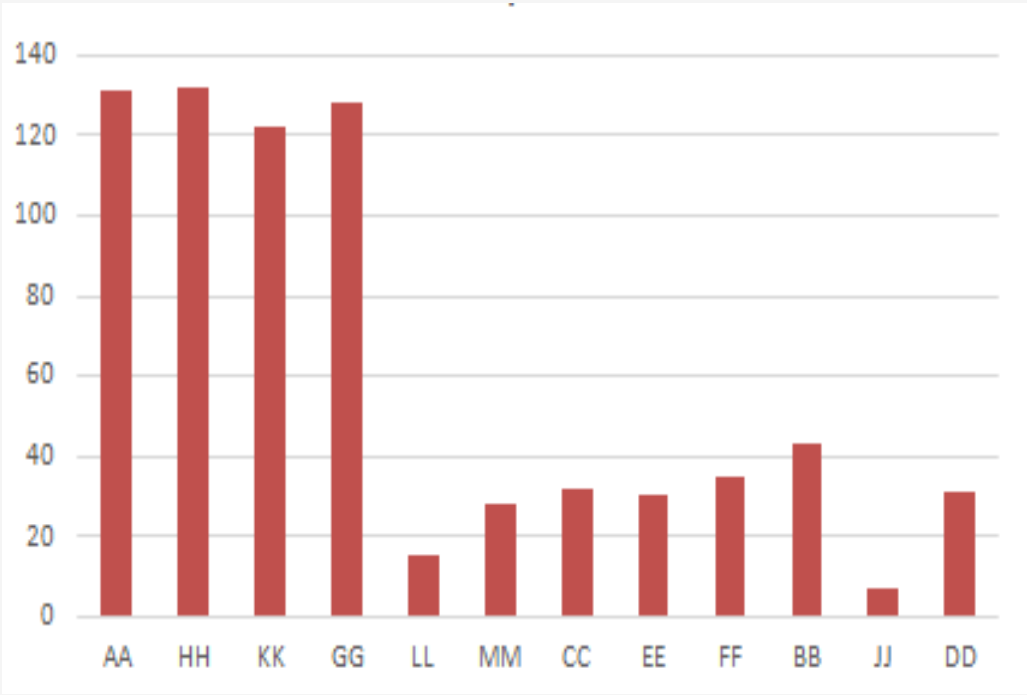
- Geographical Maps (KML)
- Interactive Plots
- Dash
- D3
- 3D plots
- **Folium**

Basic Business Data Visualization

Revenue by Products

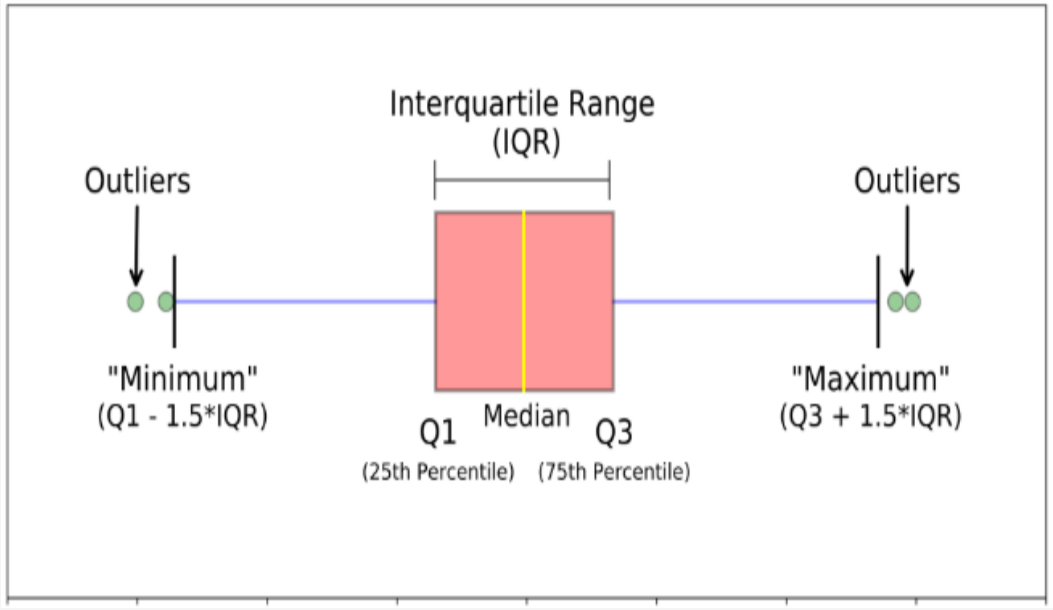


Orders by Products

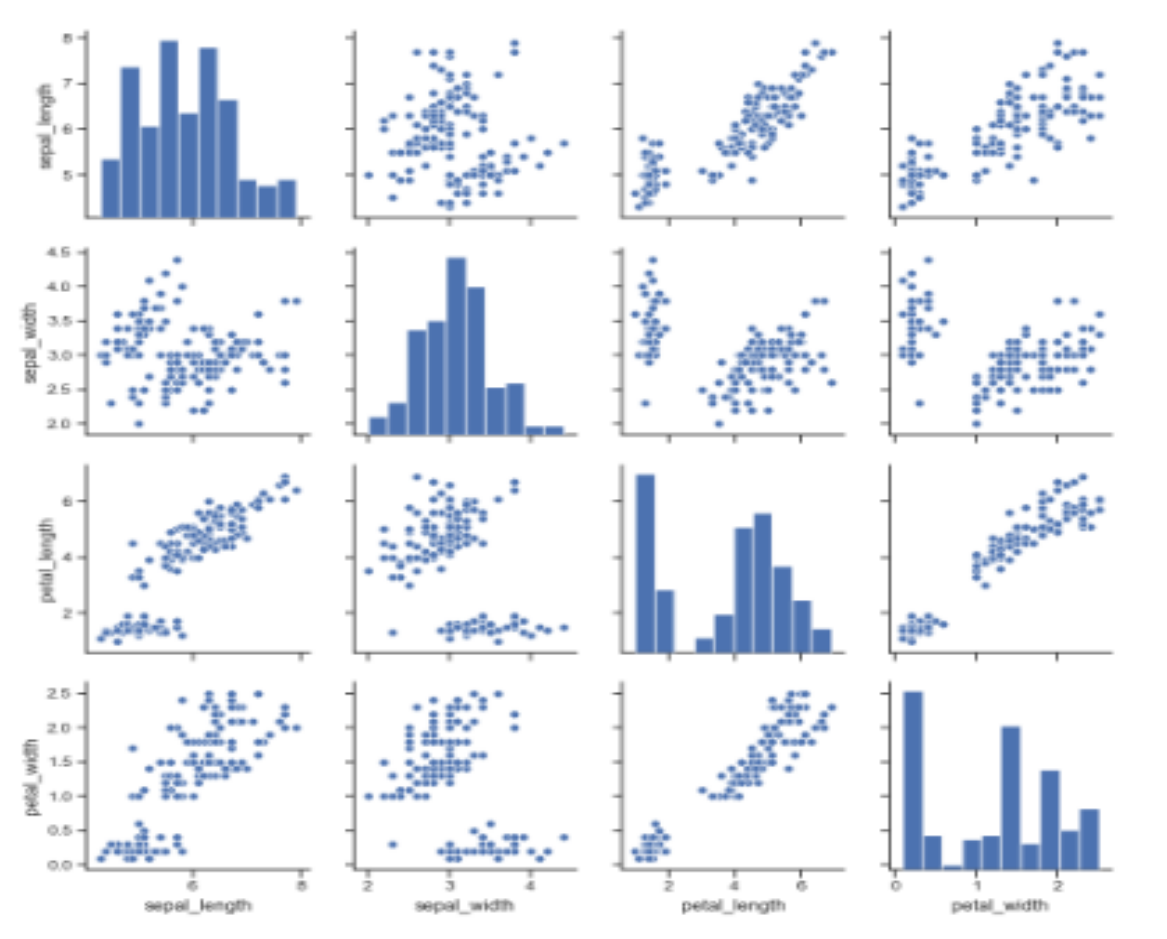


Common Plot type

Box Plot



Pairs Plot



Common Plot type

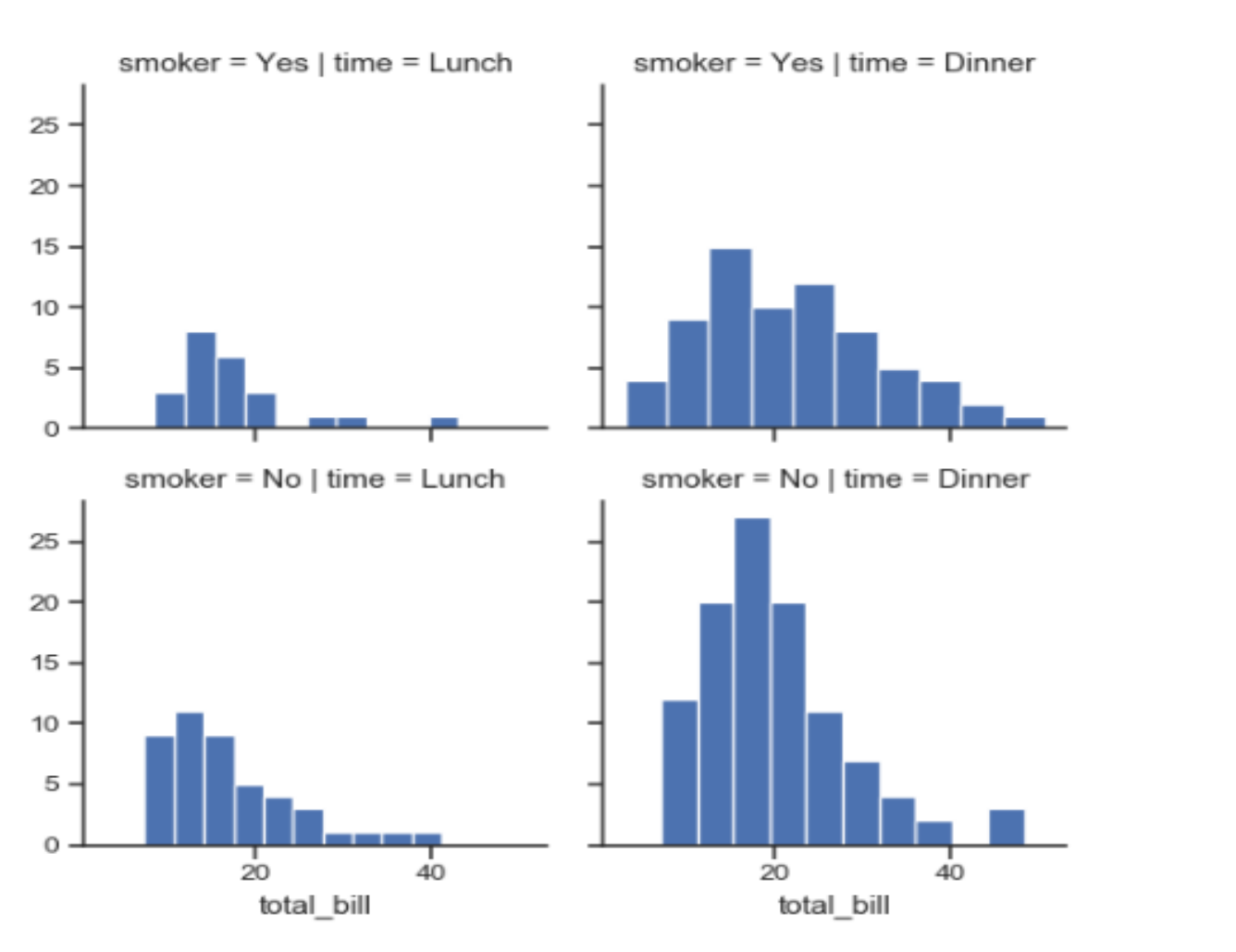
FacetGrid

Tips data set

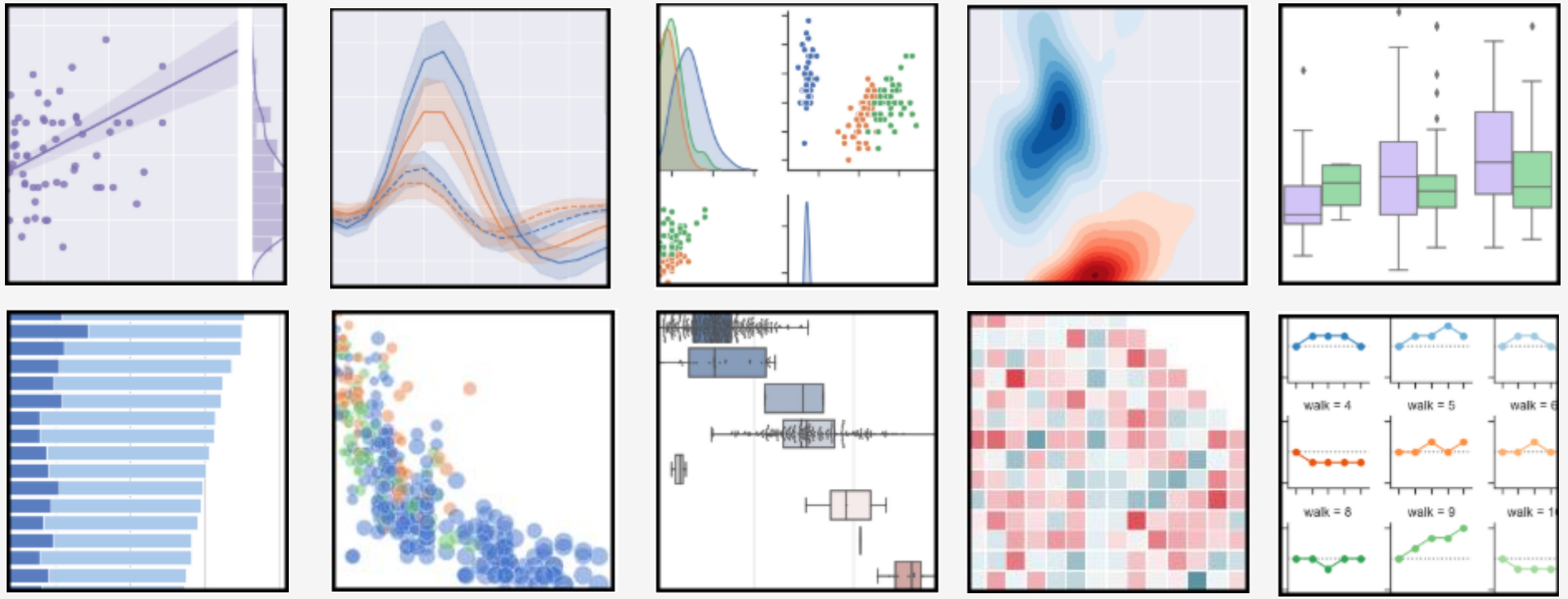
Total_bill

Smoker = Yes or No

Time = Dinner or Lunch



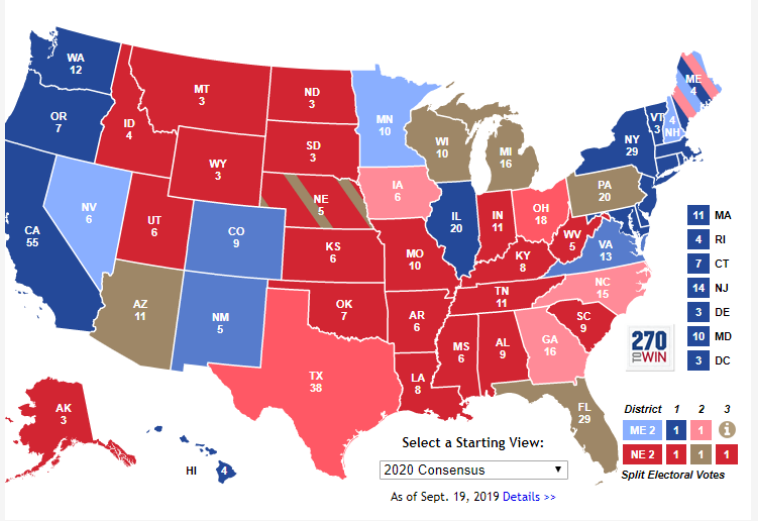
More advanced Types of Visualizations



More advanced Types of Visualizations

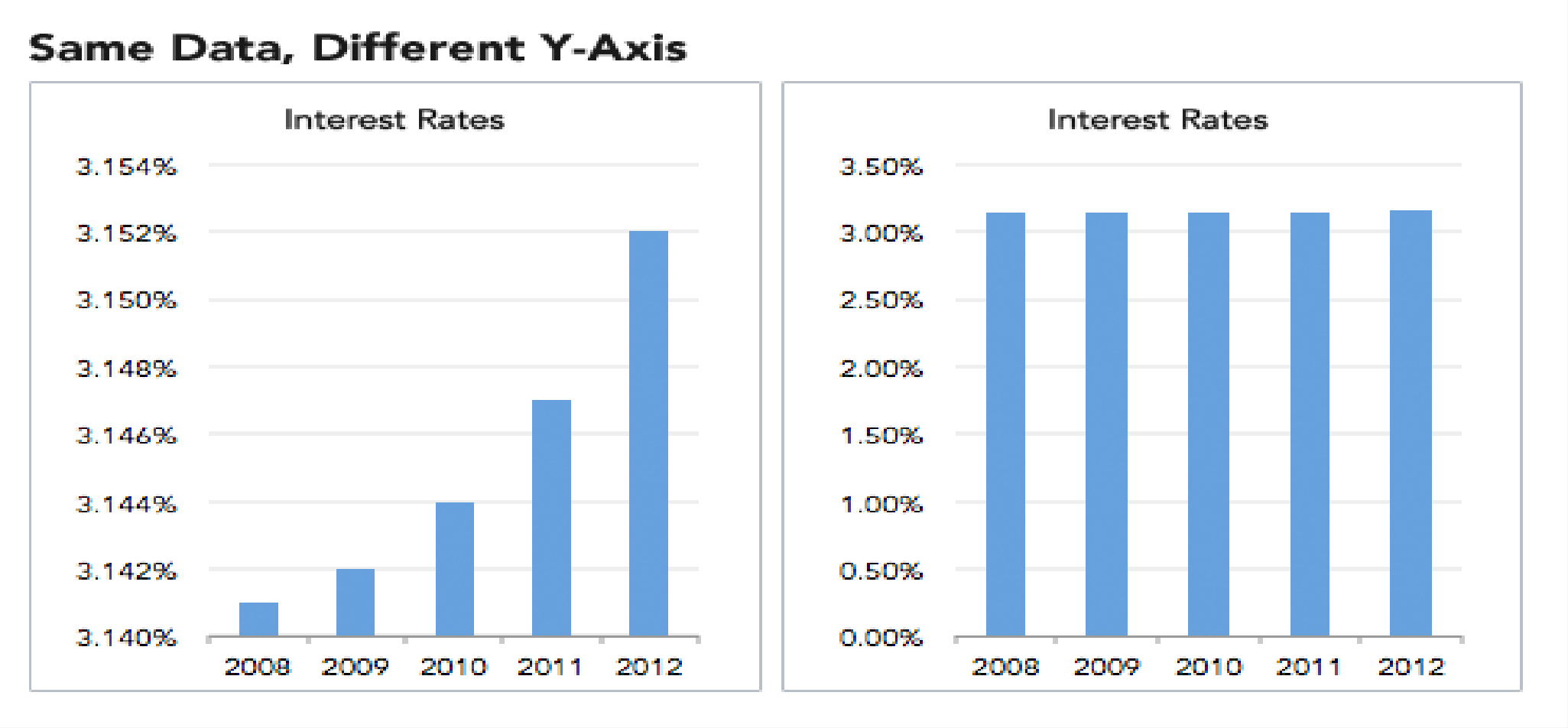
Geographical Data Visualizations

- Google Map
- Folium Python library (pip install folium)
  https://python-visualization.github.io/folium/quickstart.html
- KML (pip install simplekml)
  https://pypi.org/project/simplekml/
- Google Earth
  https://www.google.com/earth/versions/#earth-pro



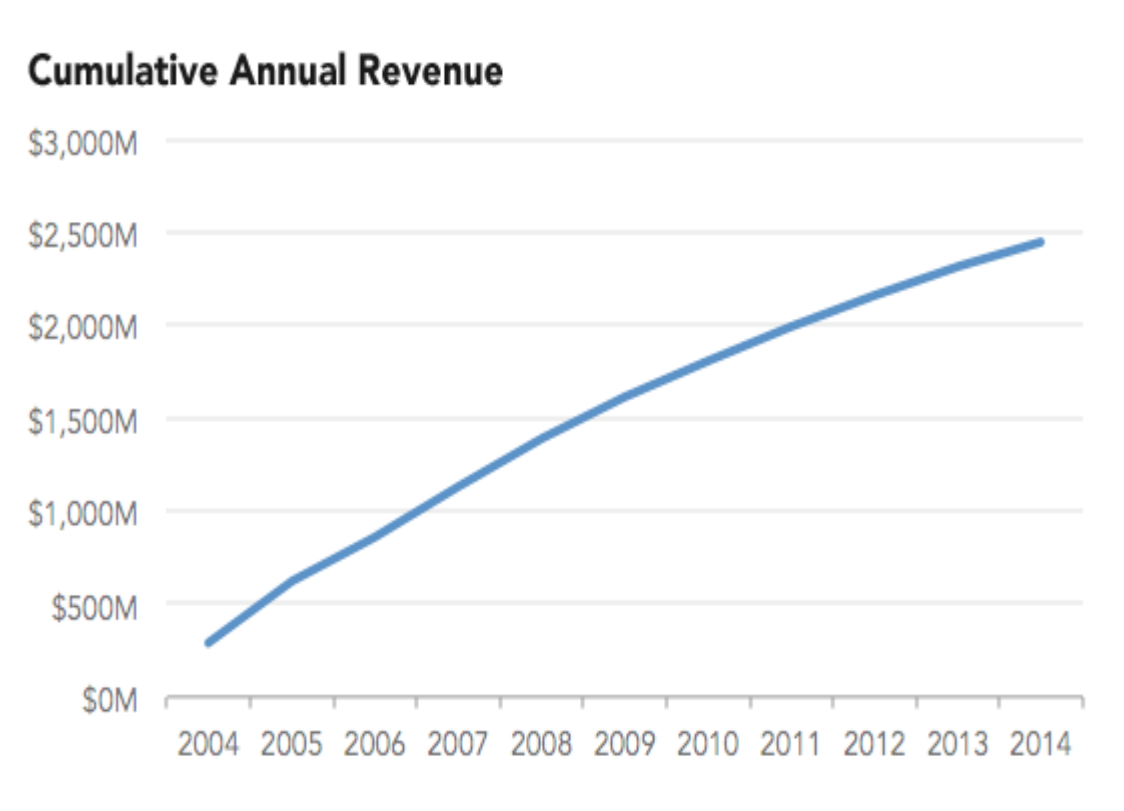
How to misuse or even lie Data Visualization

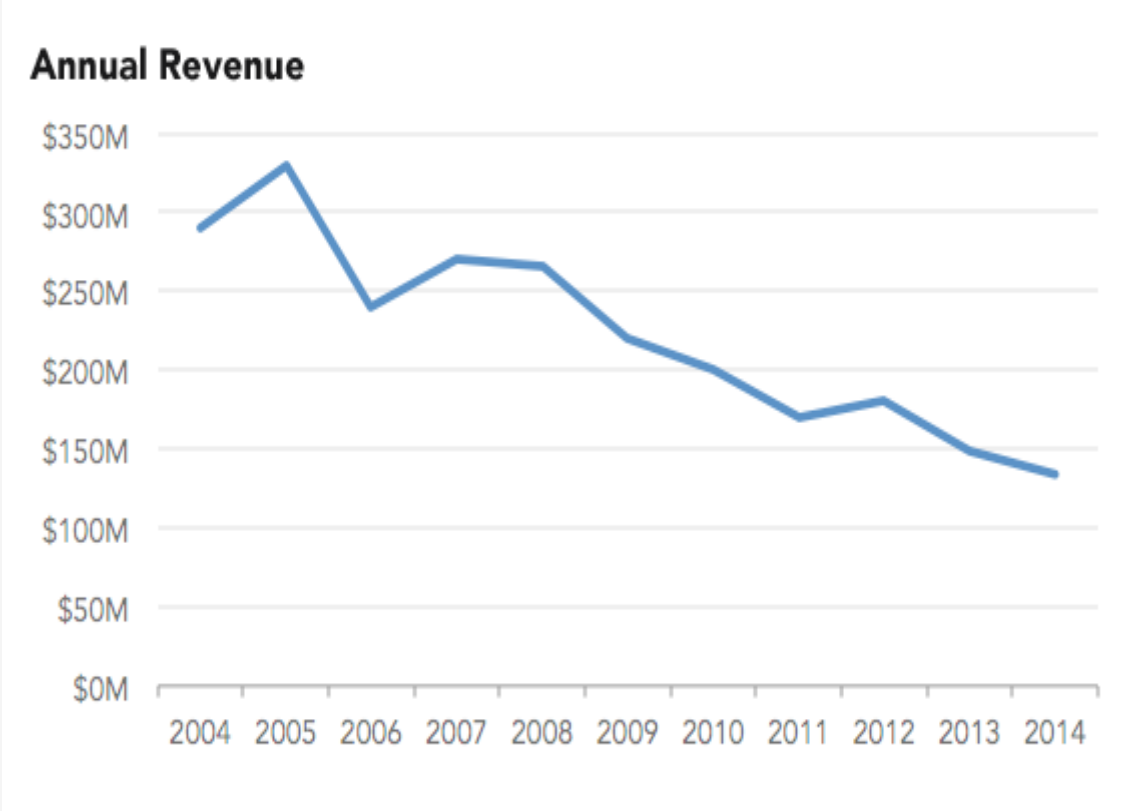
Mess around with your scale in the axis



How to misuse or even lie with Data Visualization

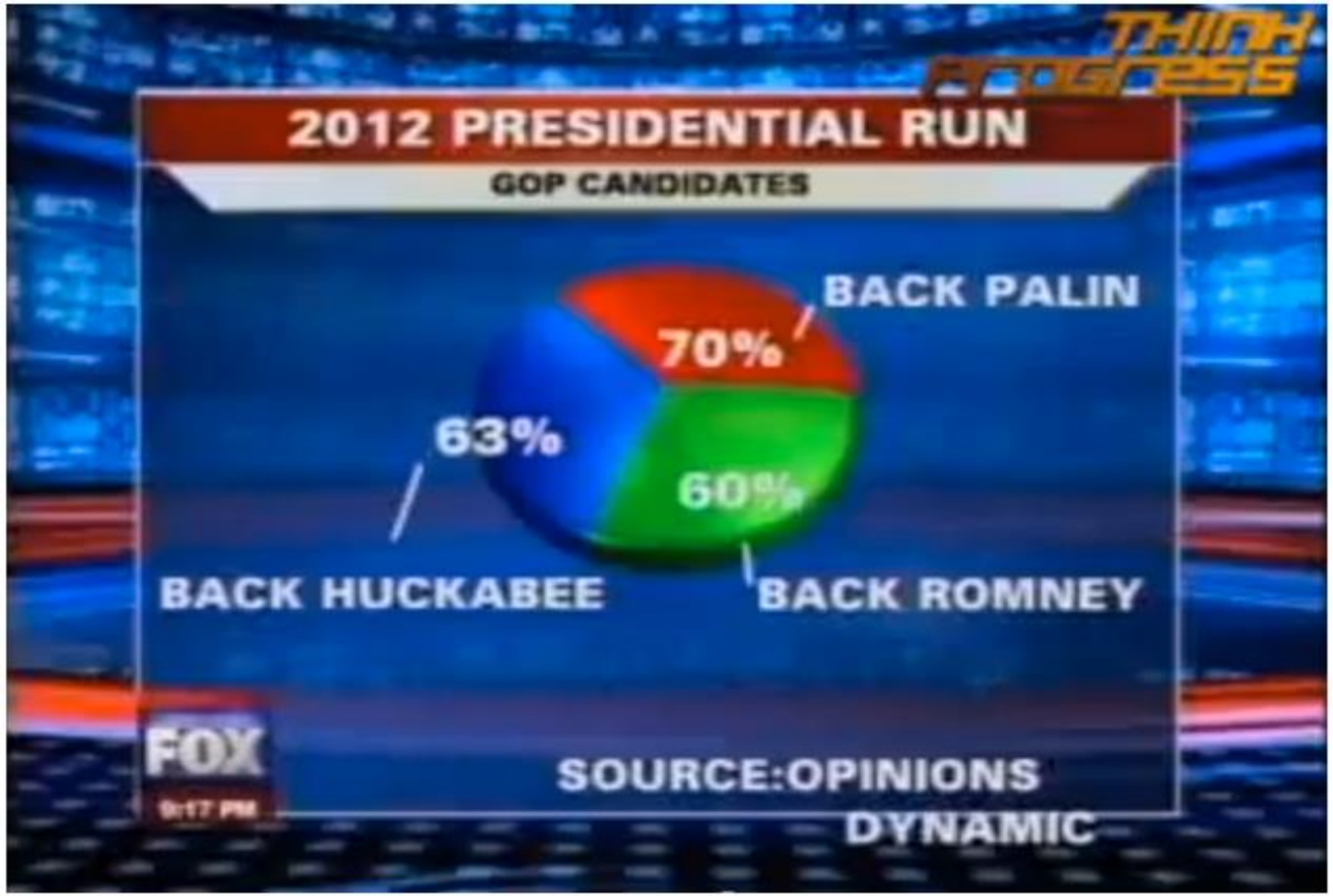
Use Cumulative graph





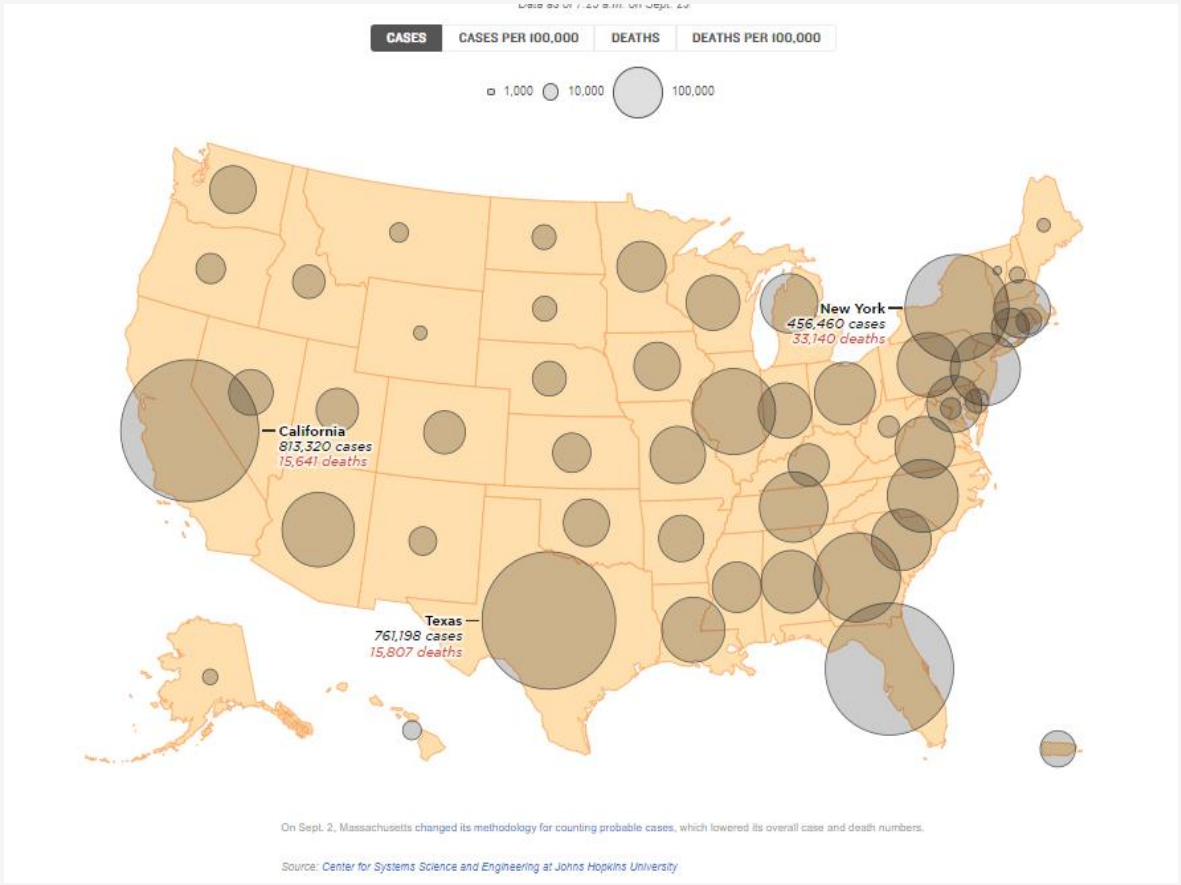
How to misuse or even lie with Data Visualization

Ignore convention

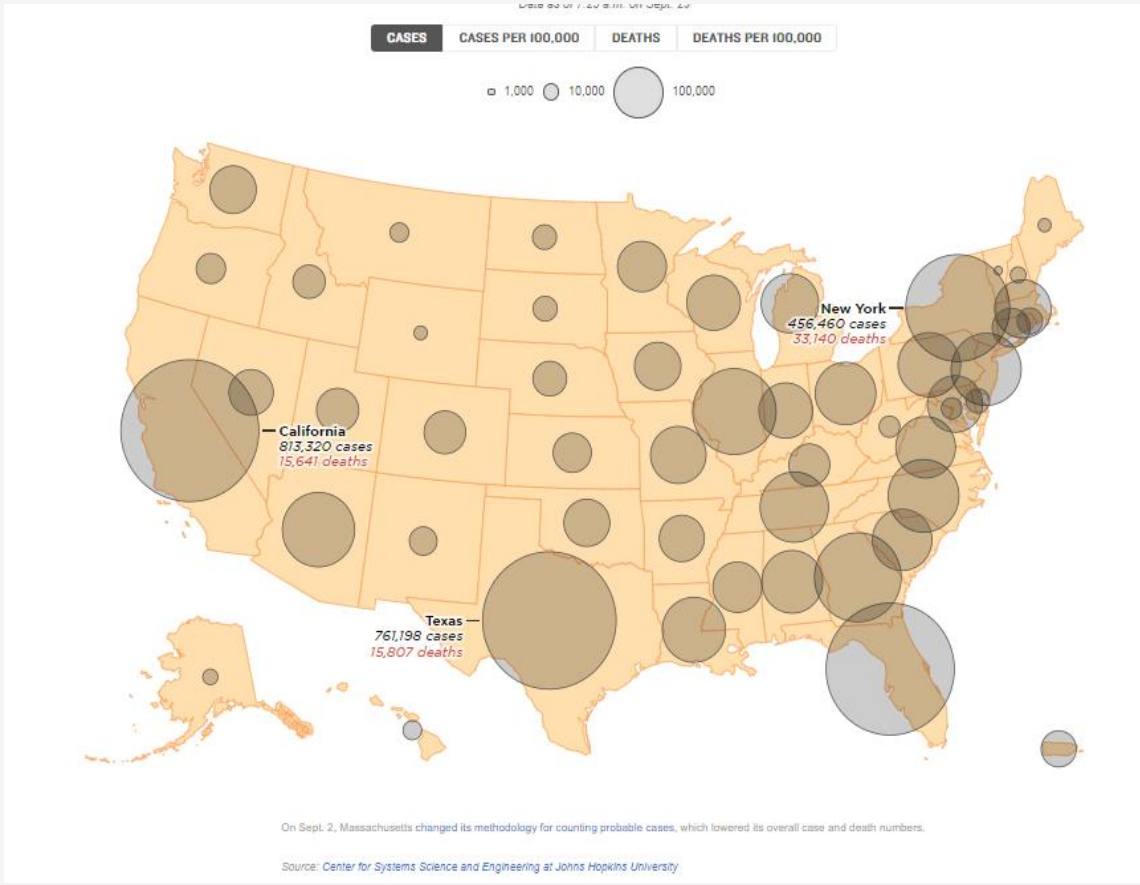


How to misuse or even lie with Data Visualization

Date 1



Date 2 looks like nothing had changed, but in fact the numbers had changed because the scale has changed.
e.g. same circle represents 120,000 instead of 100,000



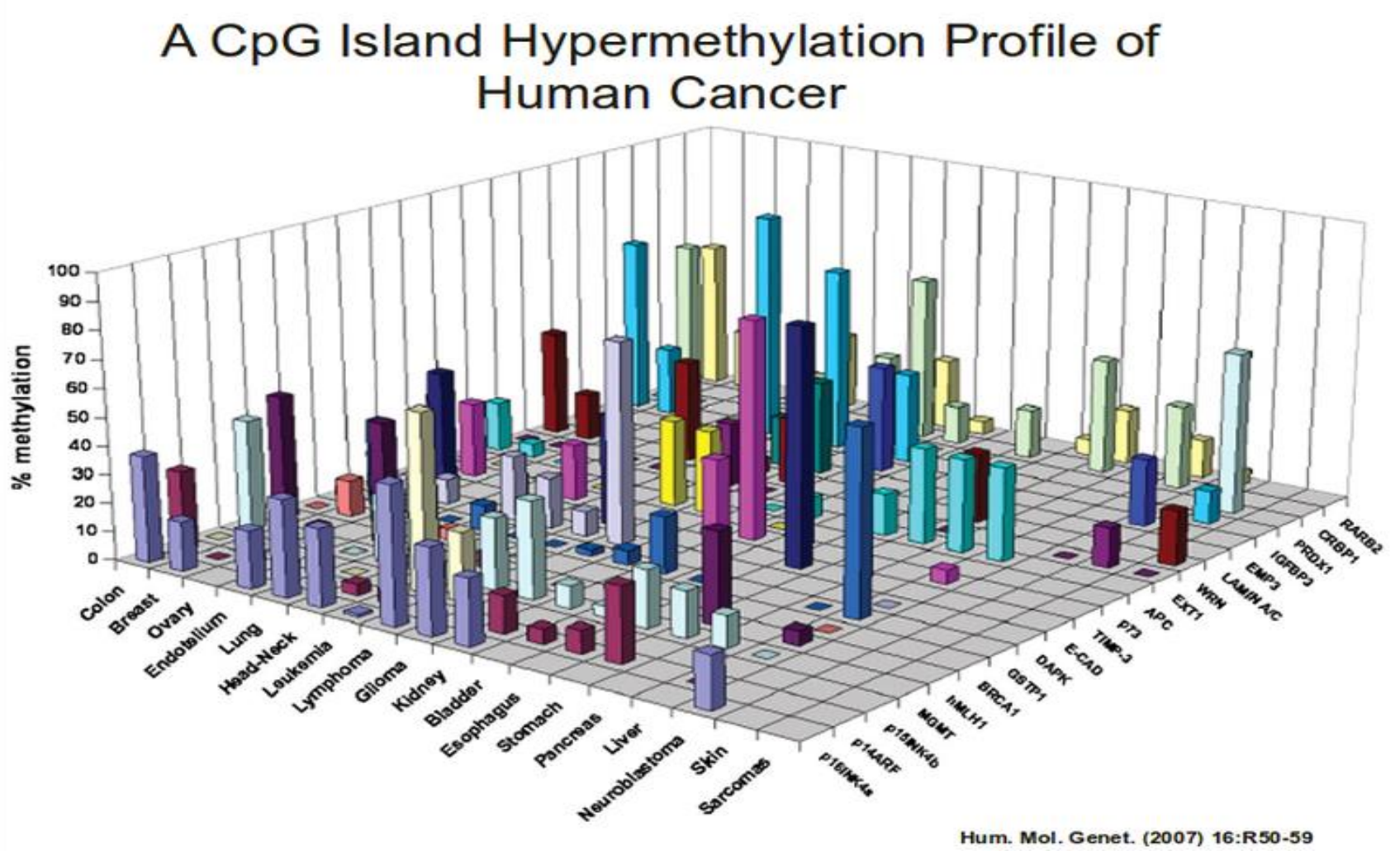
Bad way to tell a story from Data Visualization

Criteria for a bad plot:

Take more than 10 min to understand and get the takeaway from the plot

Storytelling with Data

https://www.amazon.com/Storytelling-Data-Visualization-Business-Professionals-ebook/dp/B016DHQSM2



The other extreme

Tim Cook and other Apple leaders use this clever presentation hack to make their slides memorable — and they borrowed it from Steve Jobs

When Apple CEO Tim Cook began talking about a new release of Apple's mobile operating system (iOS 13), he said: "iOS has the highest customer satisfaction in the industry, with an incredible 97%." The slide had one number in large font — 97%. In smaller font beneath the number, a sentence read: "Customer satisfaction for iOS 12.“ That's it. One number and one sentence.

https://www.businessinsider.com/apples-leaders-use-this-presentation-hack-to-make-slides-memorable-2019-6?utm_content=buffer5934a&utm_medium=social&utm_source=facebook.com&utm_campaign=buffer-bi

Best showcase of how to use Data Visualization

Hans Rosling: Master of Data Visualizations

TED talk on Public health and longevity
https://www.youtube.com/watch?v=hVimVzgtD6w

Income disparities
https://www.youtube.com/watch?v=DoSTNRhoceY
https://www.youtube.com/watch?v=AdSZJzb-aX8Ed/PBS

Accenture talk at CWRU
http://www.youtube.com/watch?v=qprHlIzhgUk .

https://www.gapminder.org/answers/how-does-income-relate-to-life-expectancy/

Data Visualizations Toolbox

Learning by doing

Geographical Data Visualizations

What is common between a Data Scientist and a Hacker?

Answer the question in one single word

Geographical Data Visualizations

Scraper

Let’s do something “fun”

Project 1: Scrap Zillow to list house for sales and shows them in Google Map

- Download Google Chrome https://www.google.com/chrome/
- And install the Instant Data Scrapper
  - https://chrome.google.com/webstore/detail/instant-data-scraper/ofaokhiedipichpaobibbnahnkdoiiah/related?hl=en

- Go to Zillow, enter Queens College and choose “For Rent”. Use the Instant Data Scrapper to download the data.

- Write a Python script to parse the download file to create a HouseForSale CSV file
  Use http://jsonviewer.stack.hu/ to view the JSON data

- Show them in folium

Project 2: Scrap Starbucks website and show them in Google Earth

- Download Parsehub (https://parsehub.com/quickstart)

- Follow the instruction from Parsehub blog to scrap Starbucks website to download their store location near you
  - https://www.parsehub.com/blog/how-to-get-the-locations-of-retail-stores-with-web-scraping/)
  - https://www.starbucks.com/store-locator?map=46.805111,-114.009559,12z

- Download Google Earth (https://www.google.com/earth/versions/#earth-pro)

- Install simplekml (pip install simplekml) and create a Google Kml file

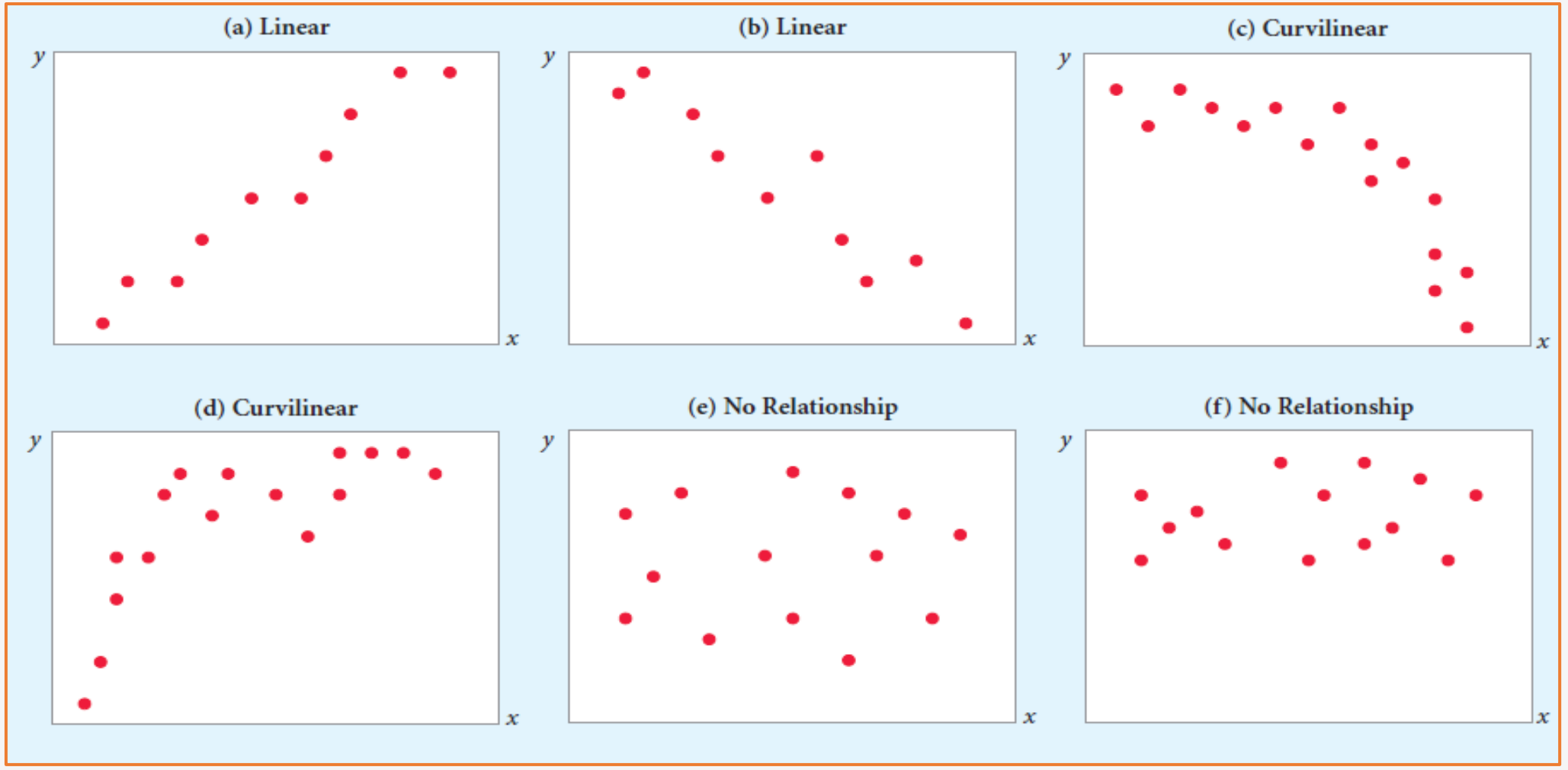
Linear Regression

Read Chapter 7 (Regression) of the Textbook

Regression Models

- To understand the application of regression analysis in data mining
    - Linear/nonlinear
    - Logistic (Logit)
- To understand the key statistical measures of fit

Relationships between variables



When the data shows linear relationship

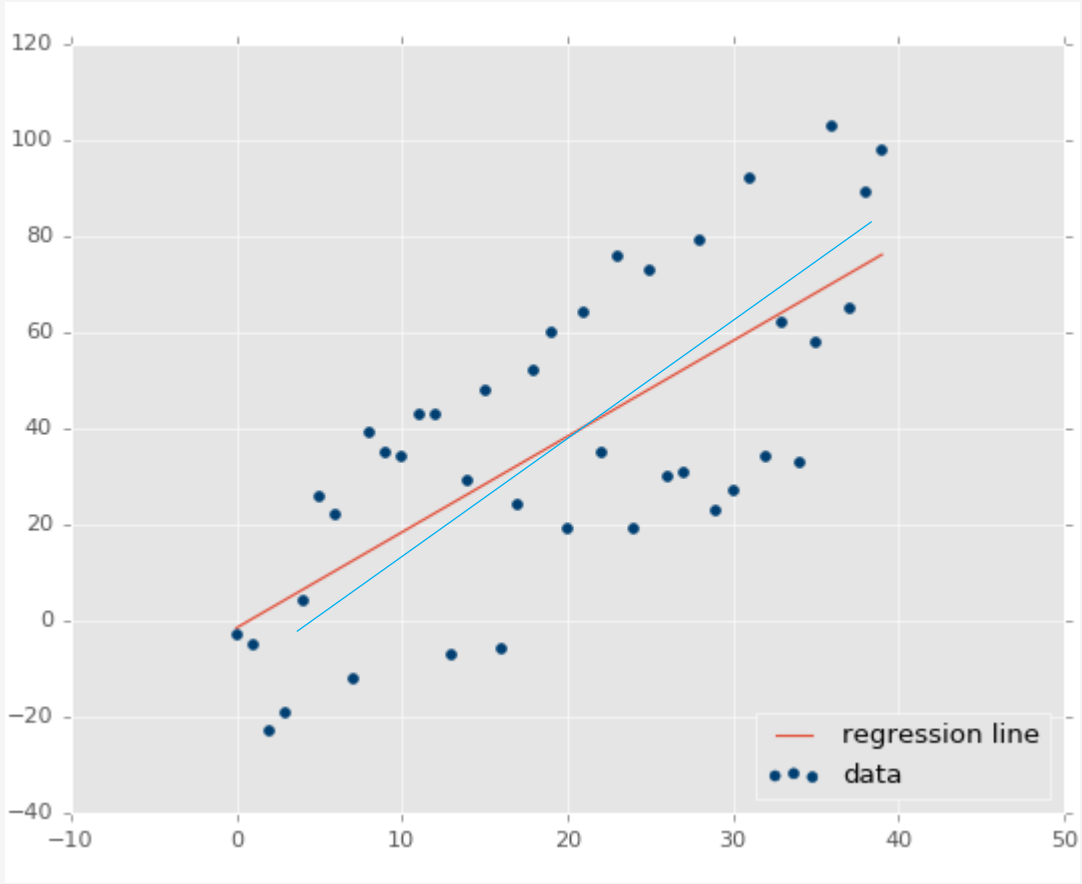
Correlation is high (positive or negative) and Scatter plots display a linear relationship

First model come to mind is

$$Y = m X + b$$

But still, there can be many lines that can “kind of” fit the data as well

Question: How to pick the “best-fit” line?

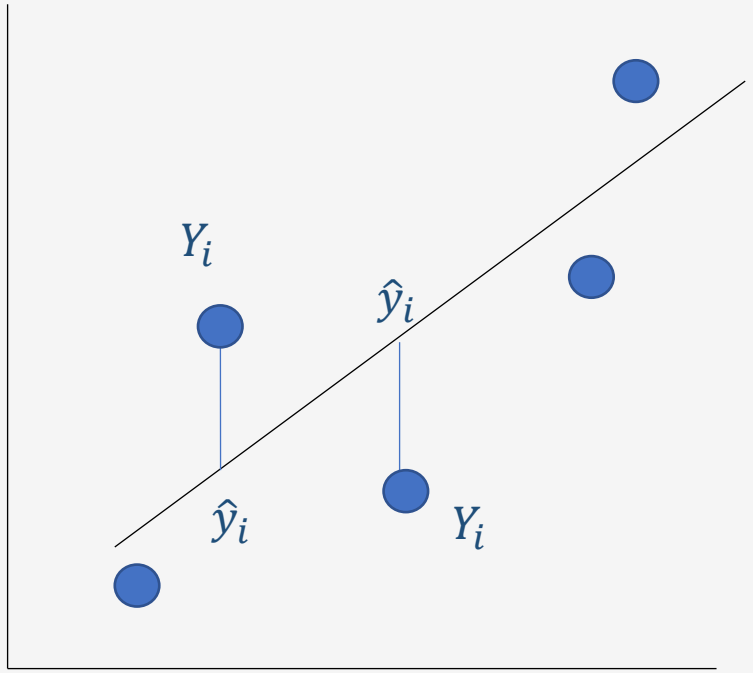


How to find the best fitting line?

Define Mean Squared Error (MSE)
To be the square of the distance between actual and predict Y values

$\hat{y}_i$ = prediction, $Y_i = actual\ value$

$$\text{MSE} = \frac{1}{N}\sum_{i}^{n}(y_i - \hat{y}_i)^2$$



Best fitted line is the line that minimize the MSE =>

Least Square Methods

R-square as metrics for determining “goodness” of the fit

- Determining the relationship between predictor & outcome
- Relationship Among SST, SSR, SSE

$$r^2 = \mathrm{SSR}/\mathrm{SST}$$

$$\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}$$

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

where:

SST = total sum of squares

SSR = sum of squares due to regression

SSE = sum of squares due to error

Higher R-square => Lower SSE => Better Model

R-square is 0% to 100%, anything > 70% is great

Common Theme, Toolbox and Research workflow in Data Science

Apply different algorithms to solve different problems based on the same <Theme> and <Research Workflow>

Algorithms

- SVM
- KNN
- Naïve Bayes
- Neural Network
- Logistics Regression
- NLP

Problems

- Regression
- Classification
- Recommendation System
- Clustering
- Association

Common Theme, Toolbox and Research workflow in Data Science

Will use Linear Regression for many of the general practices in building models, some of them are

- Split the dataset into training set and a testing set
- Use standard metrics to judge model performance
- K-fold cross validation

Learning by doing

Linear Regression Continued

Challenges Number 1 multi-linear regressions

$$Y = \text{beta}_0 + \text{beta}_1 X_1 + \text{beta}_2 X_2 + \ldots$$

- Collinearity
  - Pick the factors with highest correlation first, but what about the second factors?
    - Second highest correlation coefficients or lowest correlation with the first factor, but with high enough correlation with the dependent variable
  - Solution is: find an Orthogonal independent vectors
    - PCA (Principal Components Analysis)

=> Features Engineering

Common Theme, Toolbox and Research workflow in Data Science

Apply different algorithms to solve different problems based on the same <Theme> and <Research Workflow>

Algorithms

- SVM
- KNN
- Naïve Bayes
- Neural Network
- Logistics Regression
- NLP

Problems

- Regression
- Classification
- Recommendation System
- Clustering
- Association

Common Theme in Machine Learning

Linear Regression

Challenge Number 2:

- Relationship is NOT linear
- Solution: may become linear after transformation

$Y = a X^2 + b X + c \quad \Rightarrow \quad Y = b1\ Z1 + b2\ Z2 + b3$
where $Z1 = X^2$ and $Z2 = X$

$N = N_0 \exp(-\lambda * t) \Rightarrow \ln(N/N_0) = -\lambda * t + c$

$\Rightarrow \quad Y = m X + b$
where $Y = \ln(N)$
$X = t$

Polynomial Regression

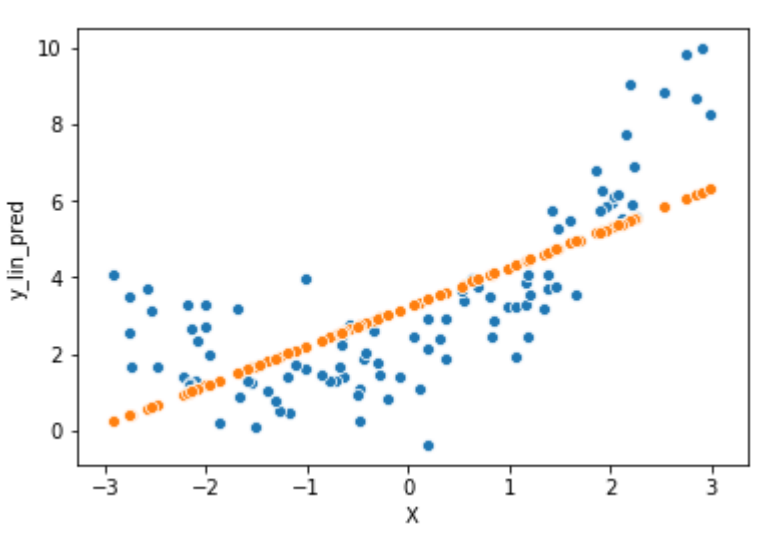
Learning by doing

Simple vs More complicated model

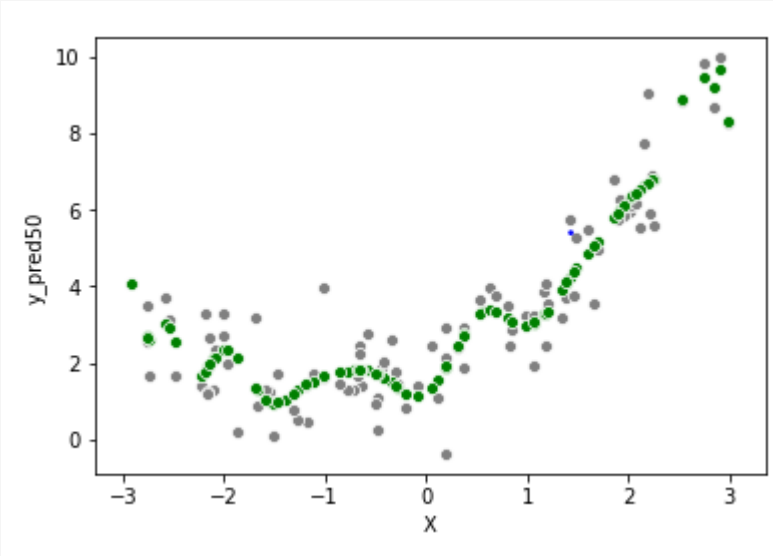
- Using a model with more parameters (more features, more predictors), you are guaranteed to fit your in-sample data (training data) better
    - More parameters => R-squares always increases
- BUT it doesn’t mean you have a better model
    - Adjusted R-squares (R-squares adjusted by penalizing models with more parameters)

Lesson Learned from Polynomial Regression

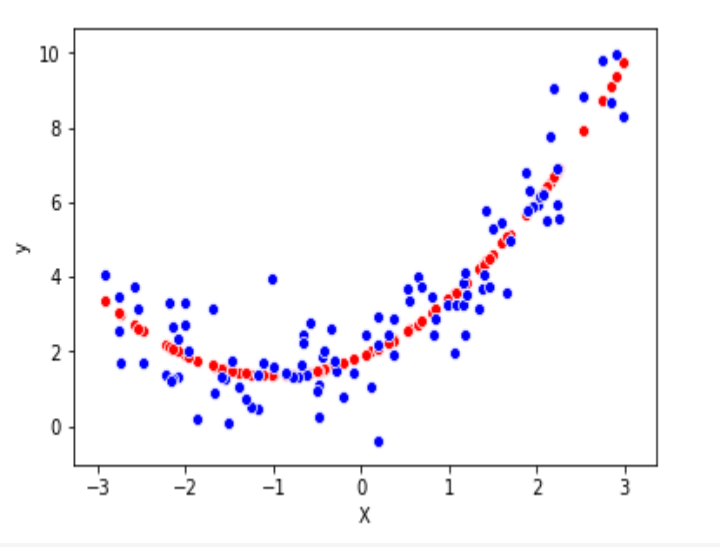
Underfit



Overfit



Good fit



A more sophisticated model tends to have smaller errors in the training set, but can perform worse in testing dataset because it overfit

A too simplistic model will never be able to fit well on both the training set as well as the testing dataset

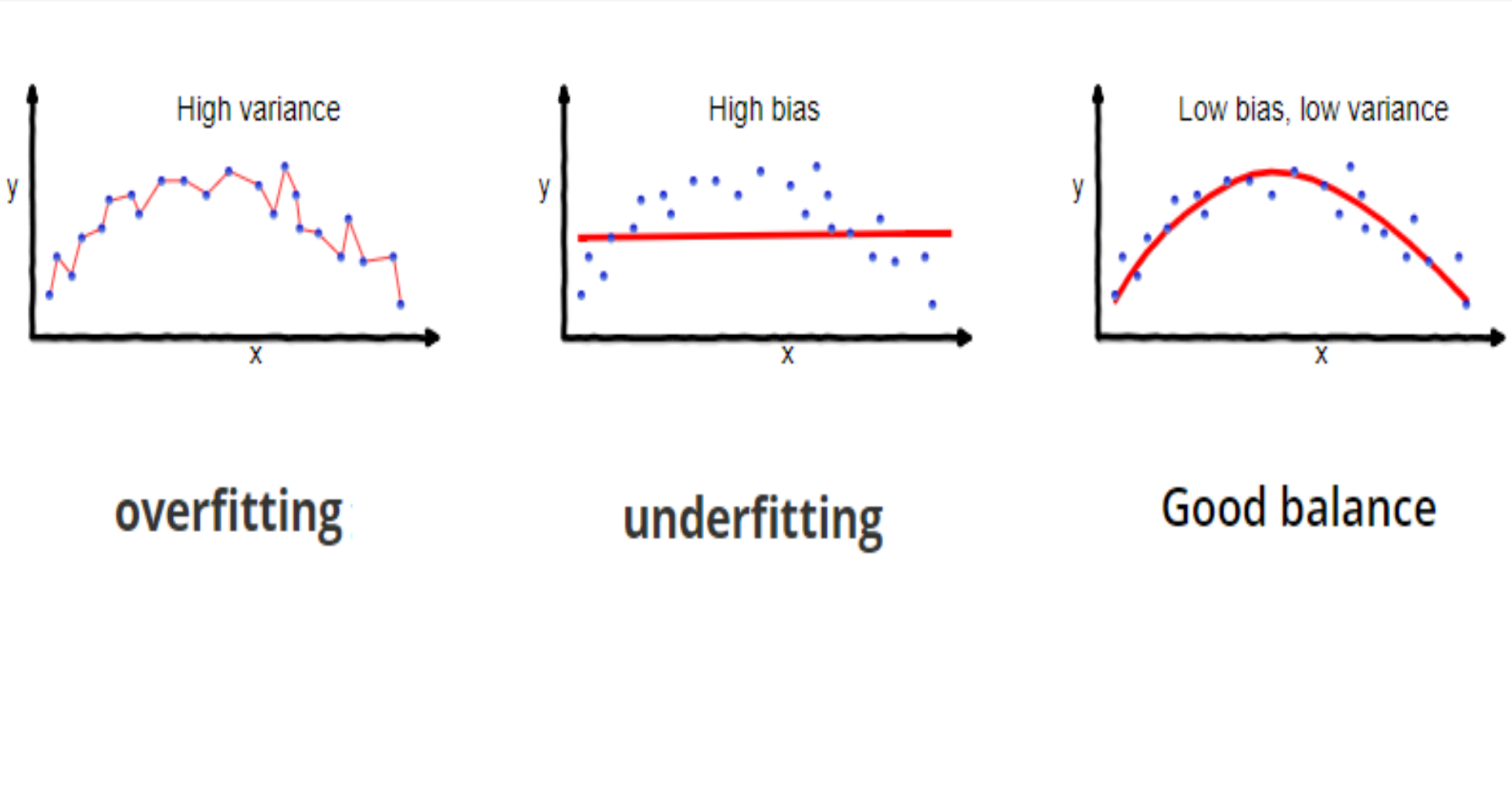
Bias vs Variance

- Bias means your model is intrinsically wrong (off, biased) that you will not fit the data well. If you use a too simplistic model, you will have high bias.

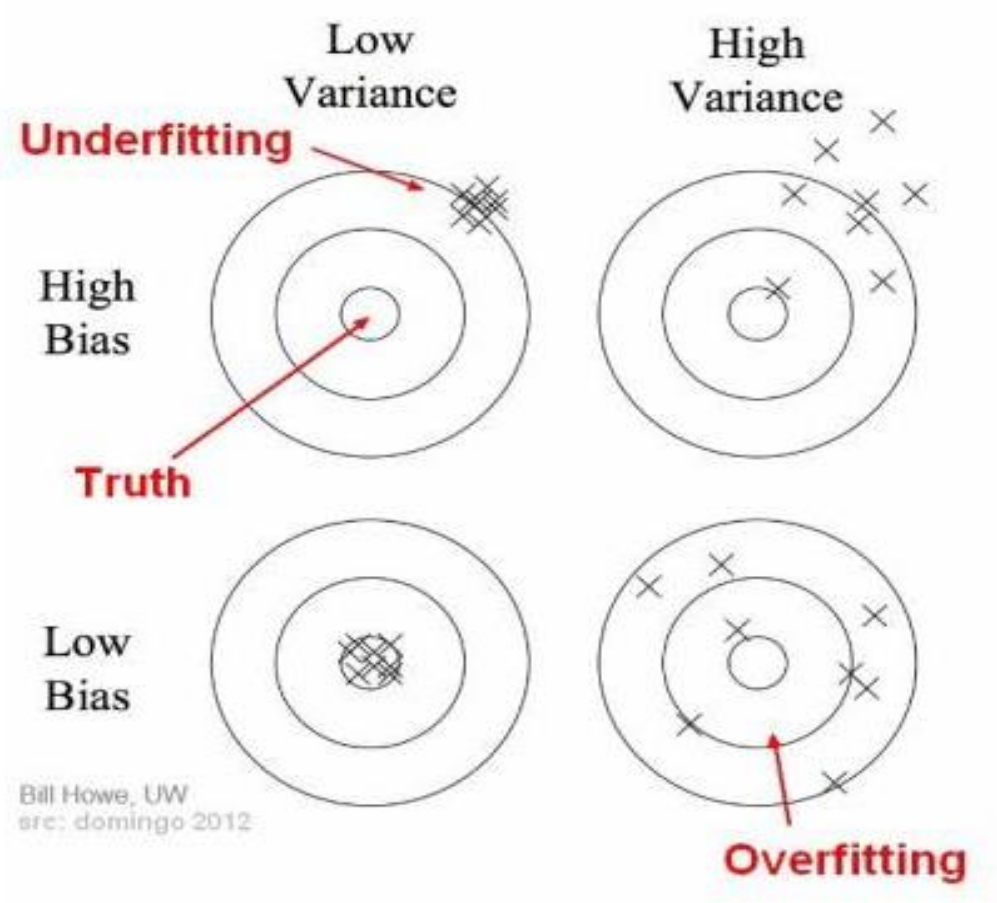
- On the other hand, using a more complicated model, you will have low bias. However, your model will not generalize well to testing dataset (out-of-sample data). The “variance” of your prediction will be high

- We call this the Bias vs Variance trade-off

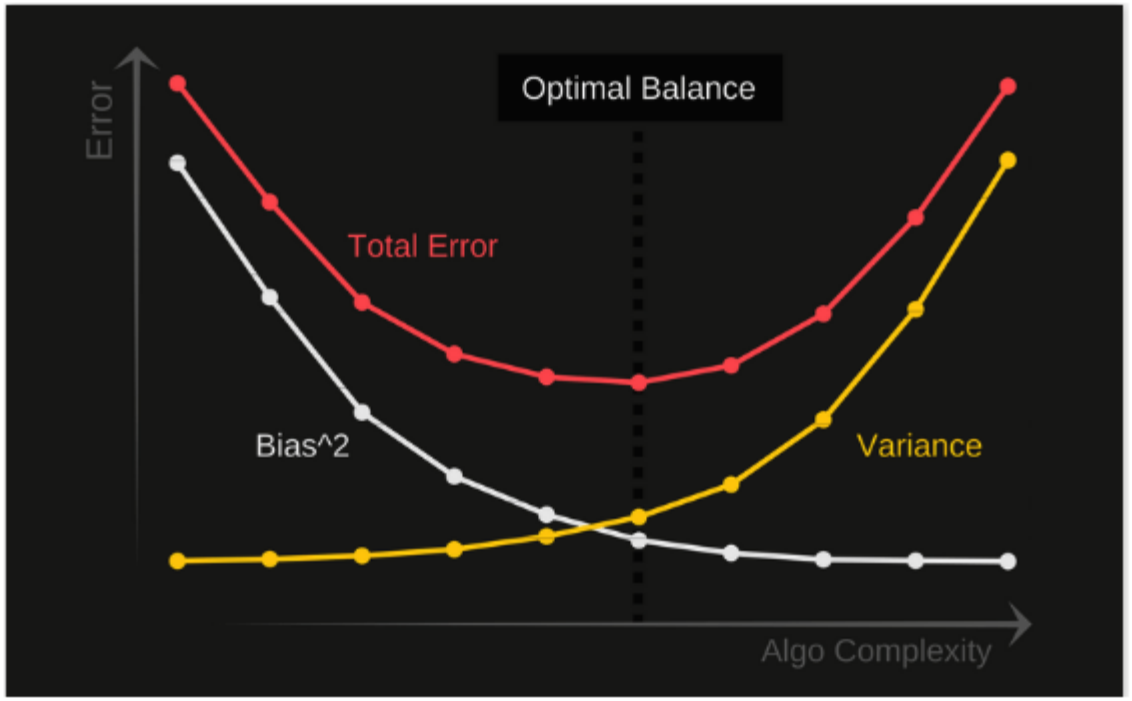
Bias vs Variance Tradeoff



Bias vs Variance Tradeoff



Total Error = Bias^2 + Variance + Irreducible Error



Recall Linear Regression can still apply to non-linear relationship

$Y = a X^2 + b X + c \Rightarrow Y = b1 Z1 + b2 Z2 + b3$
where $Z1 = X^2$ and $Z2 = X$

$N = N_0 \exp(-\text{lambda} * t)$
$\Rightarrow \ln (N/N_0) = - \text{lambda} * t + c$
$\Rightarrow Y = m X + b$ where $Y = \ln (N)$ and $X = t$

If $Y = \log (P / (1-P)) = \text{beta}_0 + \text{beta}_1 * X_1 + \text{beta}_2 * X_2 + \ldots. \text{Beta}_N + X_N$

where P is the probability of something happens

It is called Logistic Regression, which we will cover next

Classification Problem

Linear Regression: Target variable can take any numeric value

Binary Classification Problem: Target variable is either 1 or 0, Yes or No

Multi-class Classification Problem: Target variable is a list of possible values
(such as classify a picture of animal as a cat, dog, bird, fish picture)

$\Rightarrow$ NEXT TOPICS

=>Classification Problem and Logistics Regression

What do you think is my Teaching Style?

pangacademy@gmail.com

A Journey Together

Where are we in our evolution?

AP's learning style

Learning by doing

If there is only one thing you can get out of the class, it will be your familiarity with using Pandas in Jupyter Notebook environment

-Alex Pang 2019

Lots of CheatSheets and online materials suggestions

The best teacher is the one that does not teach by himself or herself

- Alex Pang 2019

AP's learning style

A critical mind is more important than just knowing mechanics

Ask the right question: Who pay more tips? Is there any bias? Why the mean or median is not good enough? Does the result change if we just focus on a subsets of the data?
Is there another underlying hidden factor?

AP's learning style

Intuition and Understanding is more important than detail formula and API calls

If you can't explain in simple English terms, you don't understand it

- Alex Pang 2019

Examples of Intuitions

The height distribution taken from Computer Science class in Queen College will have a mean ____ (higher or lower) than the whole college and a _____ (positive/zero/negative) skews

The skewness of a random variable X is the third standardized moment γ_1, defined as:[4][5]

$$\gamma_1 = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} = \frac{\mathrm{E}\left[(X-\mu)^3\right]}{\left(\mathrm{E}\left[(X-\mu)^2\right]\right)^{3/2}} = \frac{\kappa_3}{\kappa_2^{3/2}}$$

The height distribution taken from the basketball Team in Queen College will have a mean ____ (higher or lower) than the whole college and a _____ (positive/zero/negative) skews

If σ is finite, μ is finite too and skewness can be expressed

$$\begin{aligned}
\gamma_1 &= \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] \\
&= \frac{\mathrm{E}[X^3] - 3\mu\,\mathrm{E}[X^2] + 3\mu^2\,\mathrm{E}[X] - \mu^3}{\sigma^3} \\
&= \frac{\mathrm{E}[X^3] - 3\mu(\mathrm{E}[X^2] - \mu\,\mathrm{E}[X]) - \mu^3}{\sigma^3} \\
&= \frac{\mathrm{E}[X^3] - 3\mu\sigma^2 - \mu^3}{\sigma^3}.
\end{aligned}$$

The height distribution taken from Computer Science class in Queen College will have a mean ____ (higher or lower) than the whole college and _____ (positive/zero/negative) skews <u>if we know many are also in the basketball Team</u>

Examples of Intuitions

Comparing the Graduation Rate distribution with the height distribution of the Queens College students, the Graduation Rate should have a _____ (higher/same/lower) Kurtosis

The household income distribution of a gated community should have a _____ (higher/same/lower) standard deviation than a random sample of the whole population

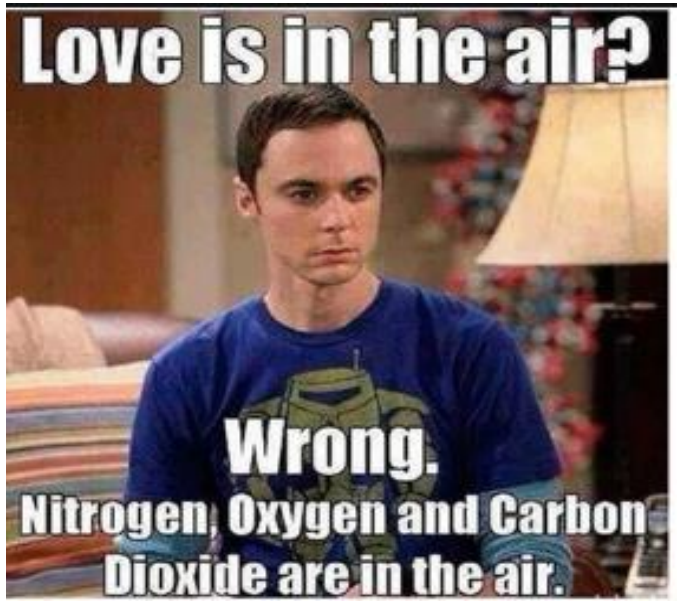
The kurtosis is the fourth standardized moment, defined as

$$\operatorname{Kurt}[X] = \operatorname{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mu_4}{\sigma^4} = \frac{\operatorname{E}[(X-\mu)^4]}{(\operatorname{E}[(X-\mu)^2])^2},$$

The kurtosis is bounded b

$$\frac{\mu_4}{\sigma^4} \geq \left(\frac{\mu_3}{\sigma^3}\right)^2 + 1,$$

Before Sheldon and Penny moves to California, they were actually in NY. And in fact they met in Queens College one day.



Penny: Hey, Sheldon, What are you up to these days?

Sheldon: Not much. But I am taking a Data Analytics class. The professor looks normal, you know, his height falls kind of between the 45% to 50% percentile, slightly below the mean. But his lecture is funny. He is absolutely an outlier. His training was in Physics which I can say with 95% confidence that it should have no correlation with computer science, right? So, I wonder how he extrapolate his Physics knowledge to Computer Science.

Sheldon vs Penny

Penny: Did you enjoy the birthday party yesterday?

Sheldon: There were too many weirdo and all sort of people around. The entropy is a bit too high for my liking. My alternative hypothesis is that the IQ distribution of the people in the party is highly negatively skewed. The kurtosis is obviously negative.

Hum… maybe I can should do a cross-validation to validate my model, but if there are only 30 data points, the central limit theorem will not apply… I will need additional sampling ideas …

CS 381/780 Data Analytics Mid-Term Review Topics

- Data Science overview
  - Datawarehouse vs Transaction based database (OLTP vs OLAP), Entity relationship, Relational data modeling
  - Dataflow (ETL), Pipeline, different data types, unstructured vs structured data, different job functions of data engineers, data analysts and data scientists
- SQL
  - join tables, group by, subqueries and various aggregate functions
- Know your Statistics
  - Sampling, Sample means and standard deviation vs population means and standard deviation, Parameters vs Statistics
  - Central Limit Theorem, Hypothesis Testing, p-values
  - Type I vs Type II errors
  - Mean, Median, Standard Deviation, Skews and Kurtosis
  - Correlation, Pearson correlation vs Spearman correlation
  - Correlation and Causation

CS 381/780 Data Analytics Mid-Term Review Topics

- Exploratory Data Analysis
  - Goals and typical steps, different ways to deal with missing values and outliers (bad data) removal
  - Various form of bias, Systematic Bias, Survival Bias
  - Different forms of mis-use of data visualization
- General Machine Learning principle
  - In-sample data vs out-of-sample data, training vs testing dataset
  - K-fold cross validation, Bias vs Variance, overfit vs underfit
- Linear Regression
  - R-squared, MSE,
  - Adjusted-R squared,
  - Cost function, Gradient descent (won't cover in mid-term)
- Classification and Logistics Regression
  - Odd vs Probability

Important concepts

- large Standard Deviation means a lot of heterogeneous data (lots of difference among the data)
- Low standard deviation means data are similar, homogenous data
- High kurtosis (a lot of outliers), Low kurtosis (not much outliers)

- Positive skew means there are more data points on the high end
- Negative skew means there are more data points on the low end

- False Positive means prediction is positive, but reality is negative (Type-1 error)
- False Negative means prediction is negative, but reality is positive (Type-2 error)

- Model is built from picking a training set, it can be tailored made for that particular training data resulting in “overfit”, so it may not generalize well into the testing data set (This is low bias, but high variance case)
- Model can be too simple, resulting in high bias (ie inherently wrong), but it will have low variance when applying to various other testing data set.