Machine Learning Engineer Course
Day4

- Exploratory Data Analysis （EDA) -

DIVE INTO CODE

Thursday April 1st, 2021

DIOP Mouhamed

Agenda

1. Check-in
2. How to proceed
3. Quick Review
4. Pandas
5. EDA
6. Important to Know
7. Assignments – Sample Code
8. Check-out

© 2020 DIVE INTO CODE Corp.

Check-in

3 minutes Please post the following point to Zoom chat.

Q. What is your main purpose in this course (your goal) ?

(Anything is fine.)

© 2020 DIVE INTO CODE Corp.

How to proceed - Objective

Purpose of learning. Purpose clarifies a person's role and the learning required. Clear learning leads to a sense of growth and confidence.

	Objective	NOT Objective
1	**Learn how to think about the program with your peers**	Memorize lots of functions
2	**Use the basic elements of the program**	Complete assignments quickly

© 2020 DIVE INTO CODE Corp.

How to proceed - Objective

Use the basic elements of analytical tools.

What does it mean to be able to use analytical tools in the first place? It's not about knowing how to do a lot of plotting.

- **Need to be able to plot to solve the problem**

© 2020 DIVE INTO CODE Corp.

Quick Review (Python)

Functions
(using lambda or def)

Variables, data types and operations

Control Statements
(if/else, For and while loops)

Data structures
(Lists and tuples)

Dictionaries

List comprehension
listtt = [i**2 for i in range(10)]

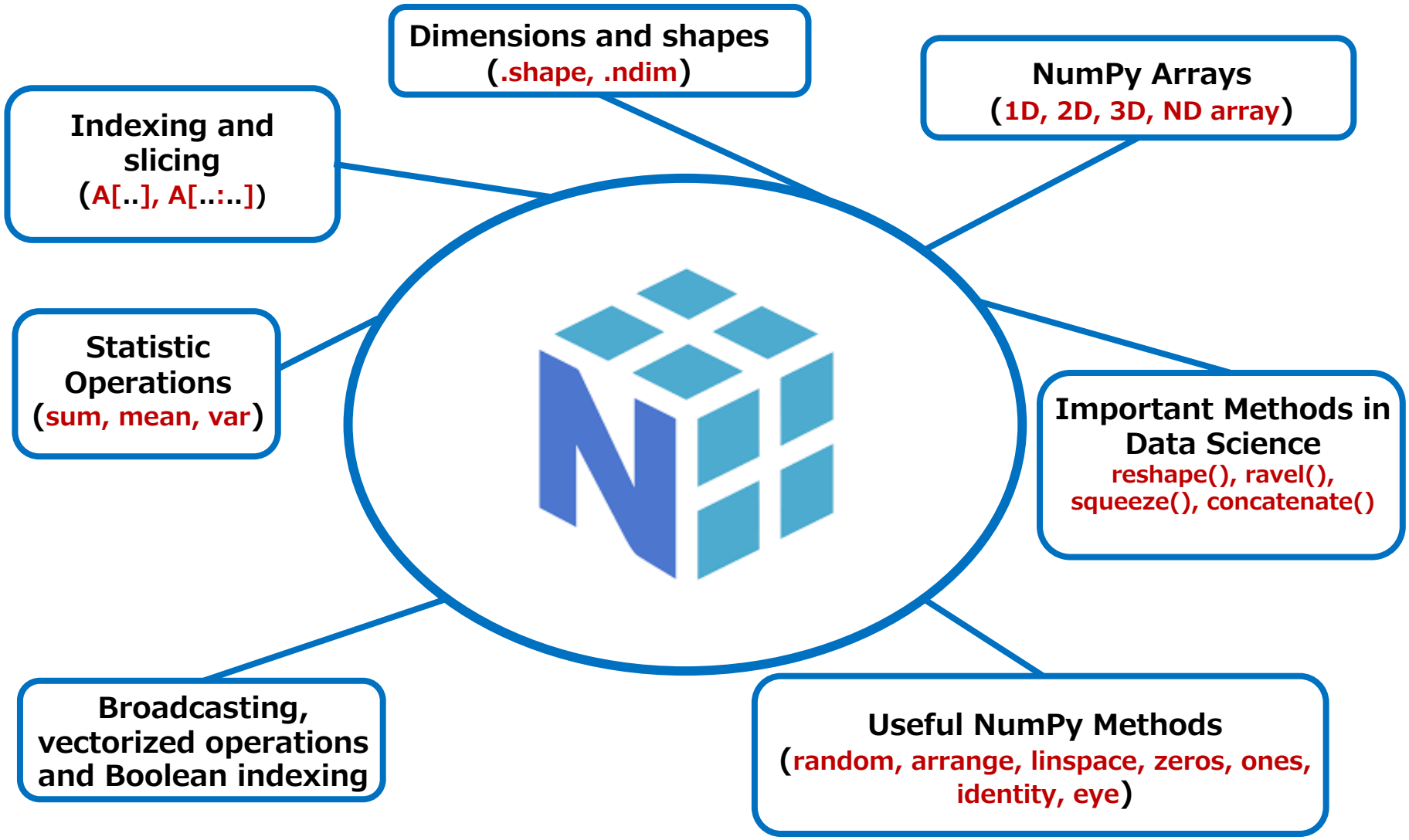
Built-in Functions
Basics & conversion functions, input(), format(), open()

Basic Python Modules
Math, statistics, & random modules

Object Oriented Programming with Python
Objects, attributes, actions and classes

© 2020 DIVE INTO CODE Corp.

Quick Review (NumPy)

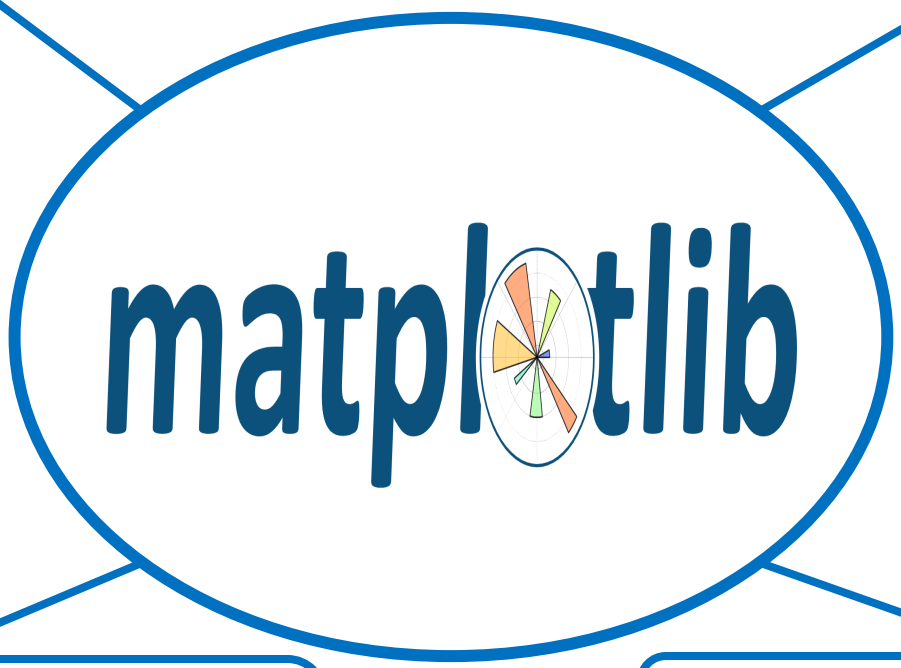


© 2020 DIVE INTO CODE Corp.

Quick Review (matplotlib)

Website

A library for plotting Data



Life Cycle of a Figure

1. plt.figure(figsize=())
2. plt.plot()
4. Extras (title, axes, legend)
5. plt.show()

Quick example

© 2020 DIVE INTO CODE Corp.

Quick Review (seaborn)

Website

Visualization library based on matplotlib

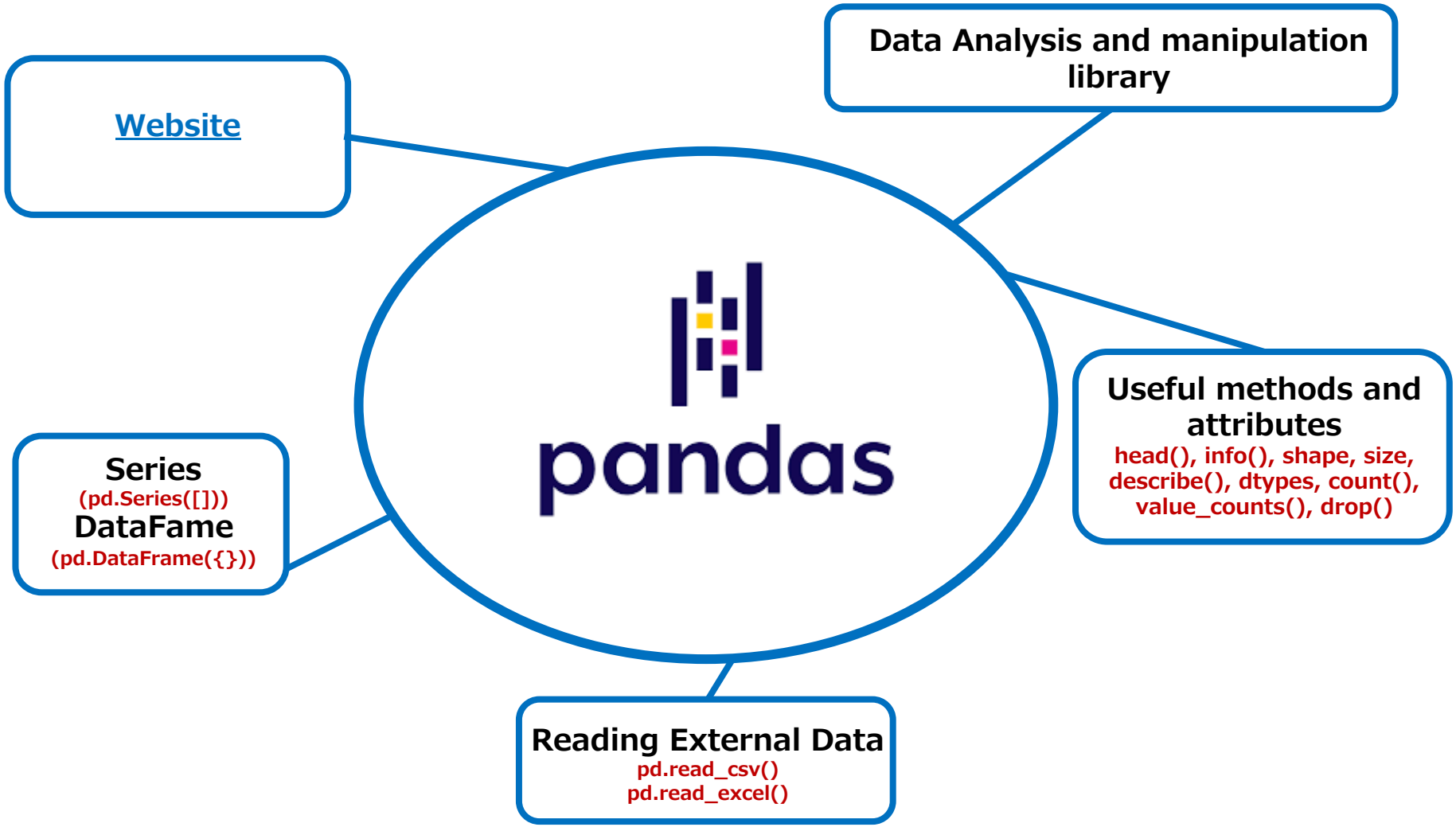
Famous methods

1. pairplot()
2. scatterplot()
4. displot()
5. boxplot() / violinplot()

IMPORTANT

© 2020 DIVE INTO CODE Corp.

PANDAS

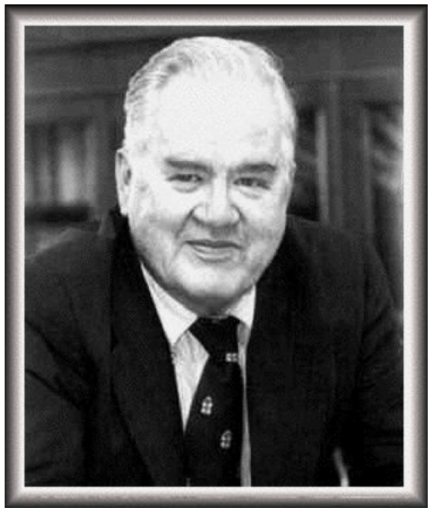


© 2020 DIVE INTO CODE Corp.

EDA

What is EDA (Exploratory Data Analysis)?

A data analysis method proposed by statistician John W. Tukey (1915-2000).
Tukey criticized traditional mathematical statistics and stressed the importance of data visualization in statistics. He is also the inventor of the "box-and-whisker diagram.



© 2020 DIVE INTO CODE Corp.

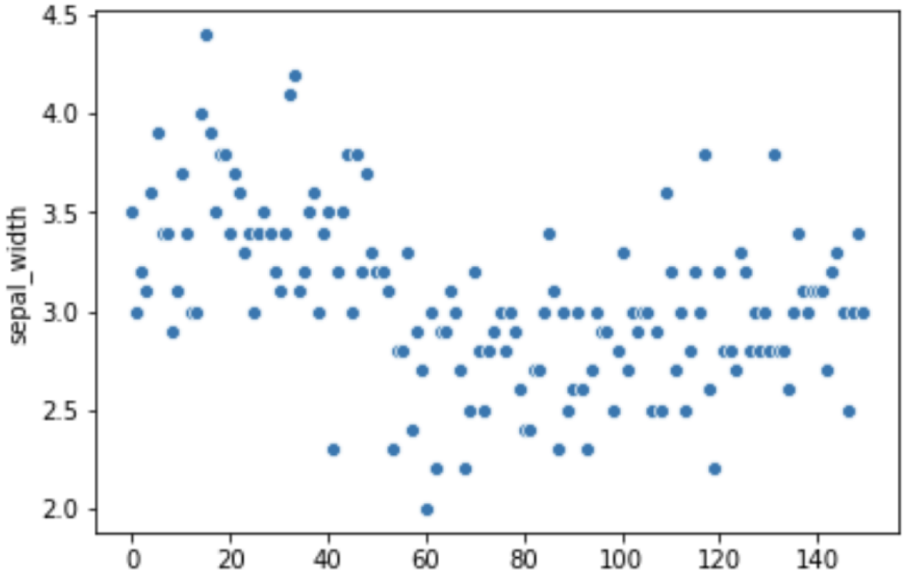
EDA

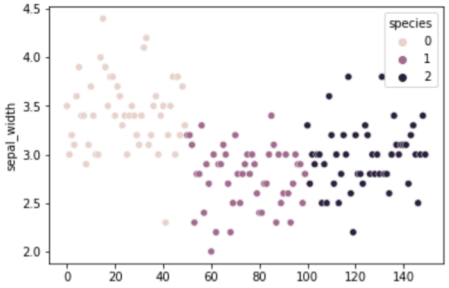
Examples of Visualization (Statistical Graphics)

(1) index plot

: Just numerical data (sepal_width) arranged according to the index number

	sepal_width
0	3.5
1	3.0
2	3.2
3	3.1
4	3.6
...	...
145	3.0
146	2.5
147	3.0
148	3.4
149	3.0





© 2020 DIVE INTO CODE Corp.

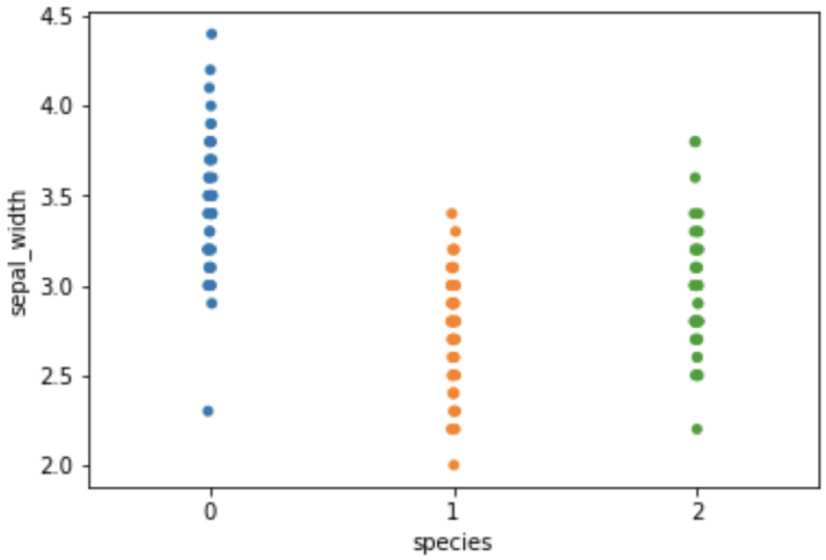
EDA

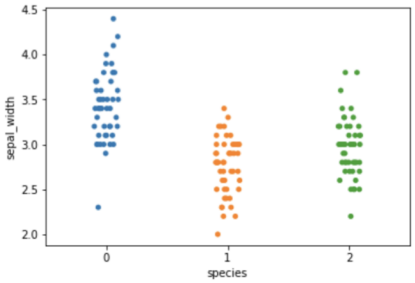
Examples of Visualization (Statistical Graphics)

(2) dot chart

: The vertical axis is the numerical data (sepal_width), and the horizontal axis is the samples grouped by type and arranged in a row.

sepal_width	species
3.5	0
3.0	0
3.2	0
3.1	0
3.6	0
...	...
3.0	2
2.5	2
3.0	2
3.4	2
3.0	2





© 2020 DIVE INTO CODE Corp.

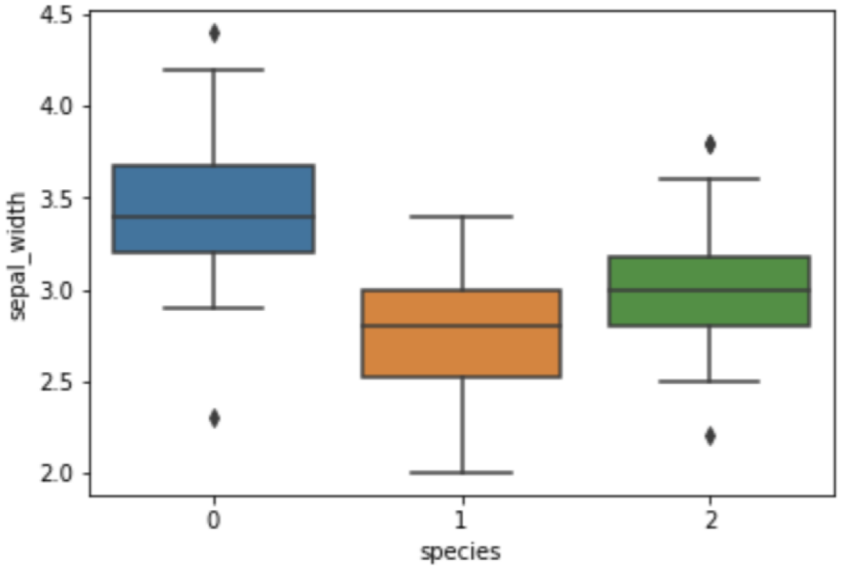
EDA

Examples of Visualization (Statistical Graphics)

(3) box-and-whisker diagram

: From the dot chart, we extracted the middle value (median) and variation for each species.

sepal_width	species
3.5	0
3.0	0
3.2	0
3.1	0
3.6	0
...	...
3.0	2
2.5	2
3.0	2
3.4	2
3.0	2



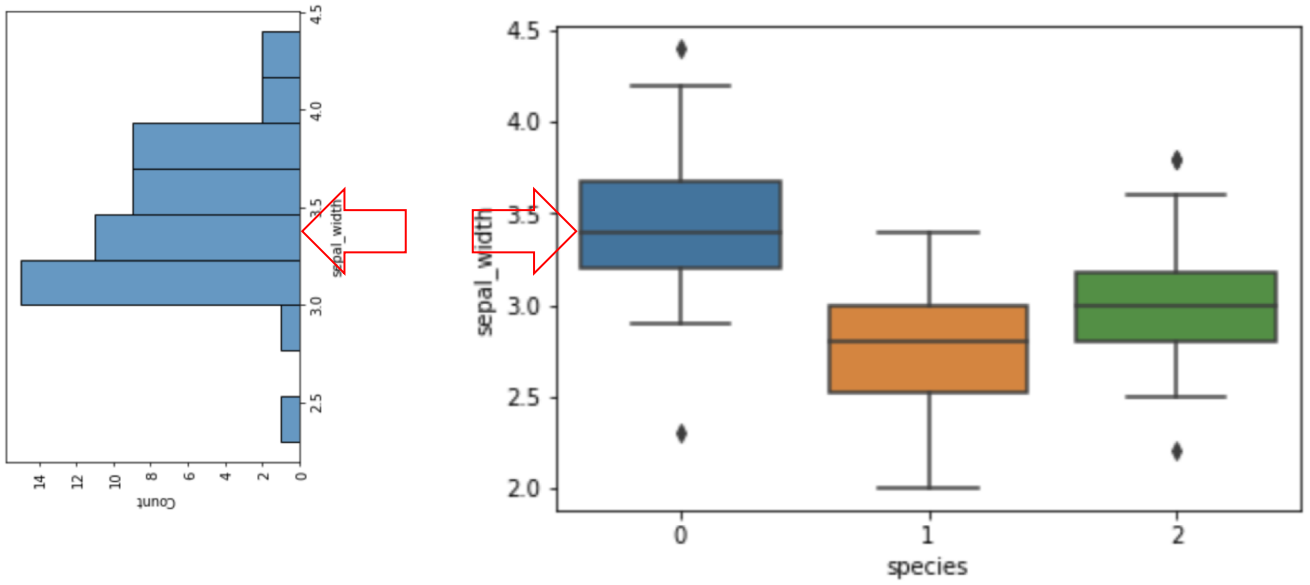
© 2020 DIVE INTO CODE Corp.

EDA

Examples of Visualization (Statistical Graphics)

(4) compare with histogram

: Compare the median value (the value in the middle of a finite number of data arranged in decreasing order)

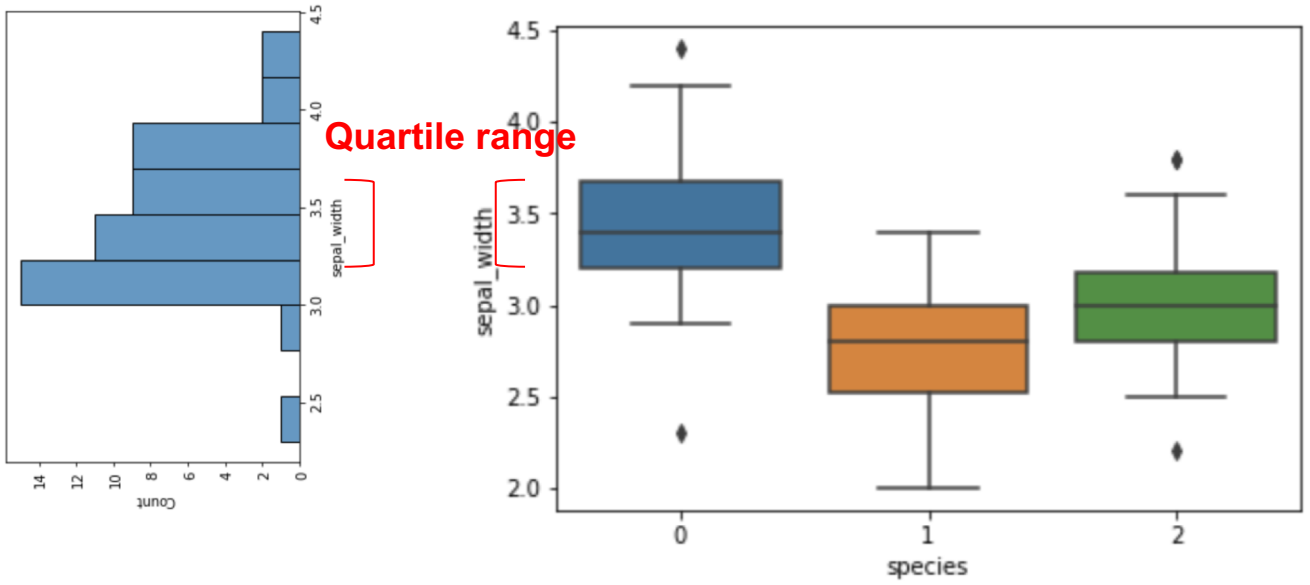


© 2020 DIVE INTO CODE Corp.

Examples of Visualization (Statistical Graphics)

(4) compare with histogram

: The width from the top to the bottom of the box is the quartile range (the range containing 25% of the sample above and below the median). In other words, it corresponds to the width where "half" of the data fits.

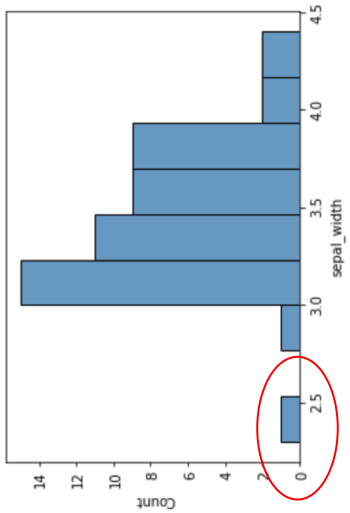


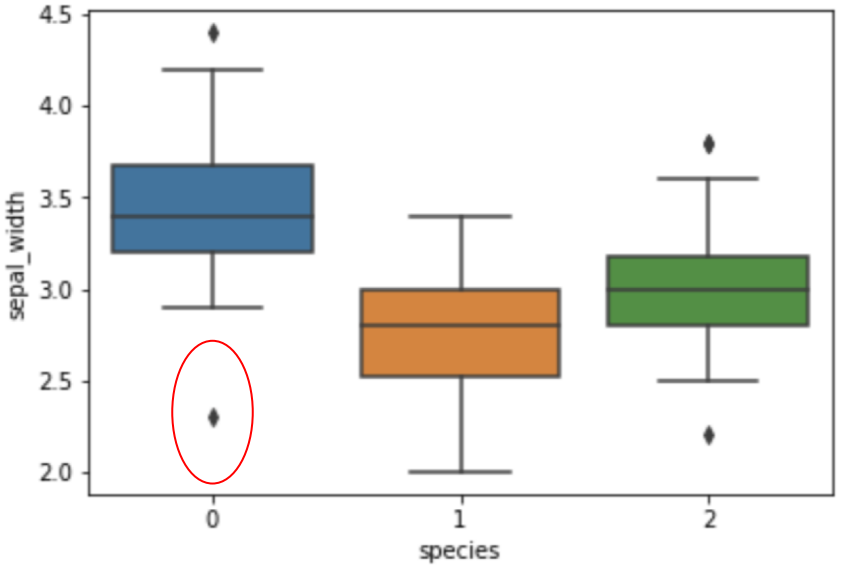
© 2020 DIVE INTO CODE Corp.

Examples of Visualization (Statistical Graphics)

(4) compare with histogram

: Comparison of outliers

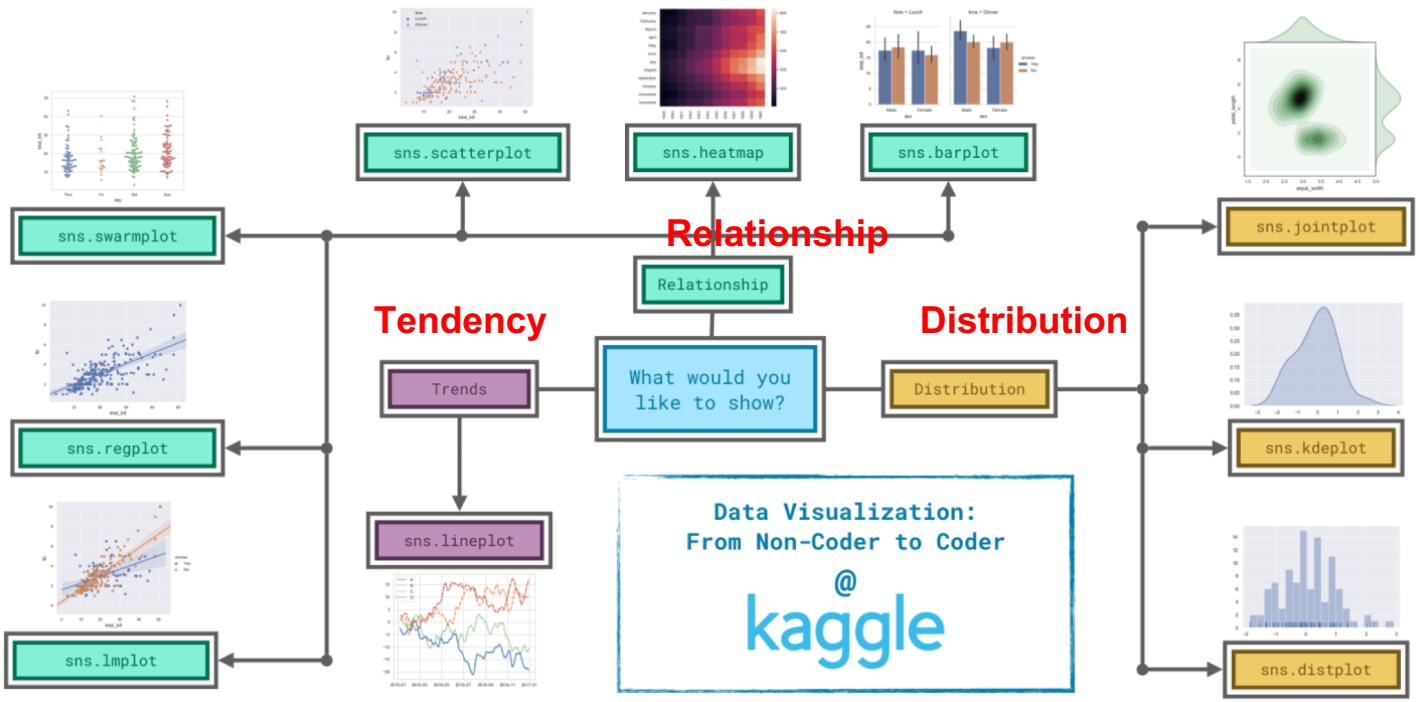




© 2020 DIVE INTO CODE Corp.

How to proceed with EDA?

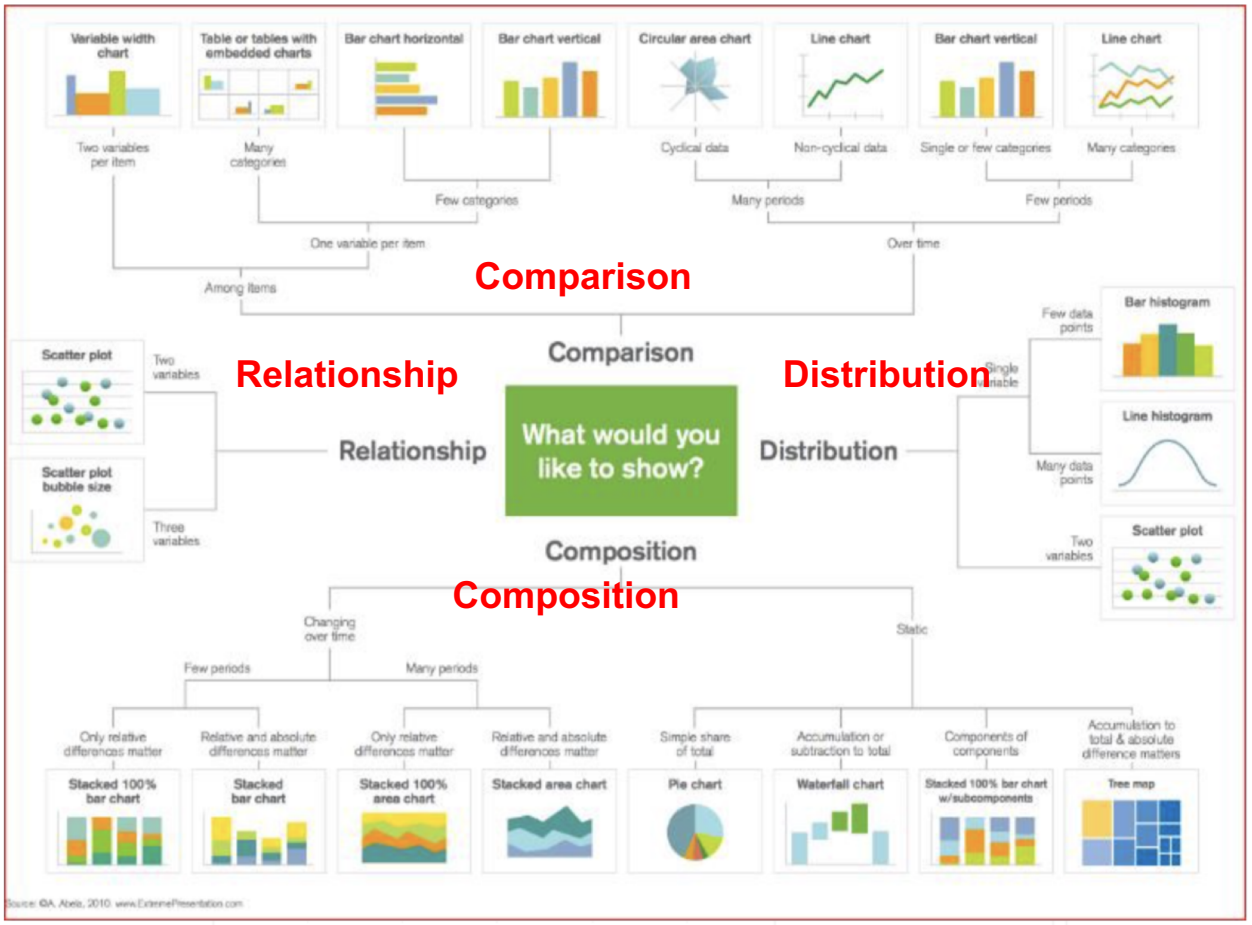
How to choose a plot



© 2020 DIVE INTO CODE Corp.

How to proceed with EDA?

How to choose a plot



EDA

A problem appears that requires students to apply the skills they have acquired in the DIVER pre-class assignment to larger-scale data.

1. Analyze your credit information

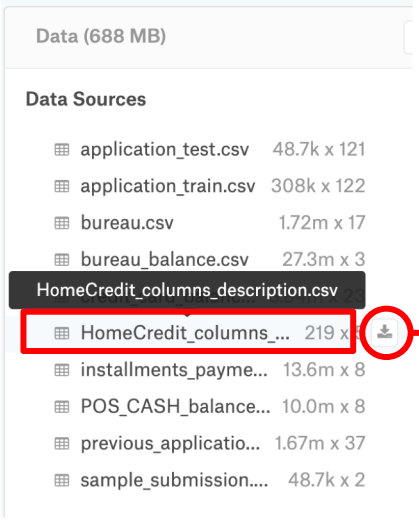
a. Understanding the Competition
b. Understanding the Data Overview
c. Assignment Setting
d. Exploratory Data Analysis (EDA)
e. (Advanced Assignment) Posting to Notebook

© 2020 DIVE INTO CODE Corp.

EDA

https://www.kaggle.com/c/home-credit-default-risk/data

[Note] Take a look at HomeCredit_columns_description.csv to see the description of the columns in the dataset.



	A	B	C	D
1		Table	Row	Description
2	1	application_{trai	SK_ID_CURR	ID of loan in our sample
3	2	application_{trai	TARGET	Target variable (1 - client with payment difficulties
4	5	application_{trai	NAME_CONTRACT_TYPE	Identification if loan is cash or revolving
5	6	application_{trai	CODE_GENDER	Gender of the client
6	7	application_{trai	FLAG_OWN_CAR	Flag if the client owns a car
7	8	application_{trai	FLAG_OWN_REALTY	Flag if client owns a house or flat
8	9	application_{trai	CNT_CHILDREN	Number of children the client has
9	10	application_{trai	AMT_INCOME_TOTAL	Income of the client
10	11	application_{trai	AMT_CREDIT	Credit amount of the loan
11	12	application_{trai	AMT_ANNUITY	Loan annuity
12	13	application_{trai	AMT_GOODS_PRICE	For consumer loans it is the price of the goods fo
13	14	application_{trai	NAME_TYPE_SUITE	Who was accompanying client when he was appl
14	15	application_{trai	NAME_INCOME_TYPE	Clients income type (businessman, working, mat
15	16	application_{trai	NAME_EDUCATION_TYPE	Level of highest education the client achieved
16	17	application_{trai	NAME_FAMILY_STATUS	Family status of the client
17	18	application_{trai	NAME_HOUSING_TYPE	What is the housing situation of the client (renting
18	19	application_{trai	REGION_POPULATION_RELATIVE	Normalized population of region where client lives
19	20	application_{trai	DAYS_BIRTH	Client's age in days at the time of application
20	21	application_{trai	DAYS_EMPLOYED	How many days before the application the perso
21	22	application_{trai	DAYS_REGISTRATION	How many days before the application did client
22	23	application_{trai	DAYS_ID_PUBLISH	How many days before the application did client
23	24	application_{trai	OWN_CAR_AGE	Age of client's car
24	25	application_{trai	FLAG_MOBIL	Did client provide mobile phone (1=YES, 0=NO)

© 2020 DIVE INTO CODE Corp.

Important to Know

Kaggle: a must-known platform.

© 2020 DIVE INTO CODE Corp.

Week 3 Assignments

Explanations about each one of them will be given but please try them on your own first.

1. **EDA - Pre-Class Assignment 1 Analysis of Irises**
2. **EDA - Pre-Class Assignment 2 Analysis of Housing Information**
3. **EDA - Class Assignment Credit Information Analysis**

Please work on your own after class and submit your assignments on DIVER.

© 2020 DIVE INTO CODE Corp.

ToDo by next class

Next class will be Zoom : Thursday 8 April 2021

🚨ToDo: Classification of irises

© 2020 DIVE INTO CODE Corp.

Check-out

3 minutes Please post the following point to Zoom chat.

Q. Current feelings and reflections
(joy, anger, sorrow, anticipation, nervousness, etc.)

© 2020 DIVE INTO CODE Corp.

Thank You For Your Attention

© 2020 DIVE INTO CODE Corp.