



Intro to **Data Science**



Presented by

Muhammad Sifa'u'l Rizky

Data Analyst at **Astra Graphia IT**
Former Lead of Data Science Instructor.

Brought to you by **gratisinbelajar.com**

Kelas **Gratis** ini akan kita adakan pada hari **Sabtu, 31 Oktober 2020** **Pukul 15:30 WIB**

Link Pendaftaran
bit.ly/daftarkelasgb25

*Pendaftaran Kelas terbatas

zoom YouTube

About Me:

Data Scientist and Data Science Instructor

Data Analyst at Astra Graphia IT

Mathematics ITS Surabaya 2013-2017

Experience:

- Lead Data Science Instructor Make.AI (2018-2020)
- Reviewer Udacity (2019-now)
- Bertelsmann Data Science Scholarship from Udacity (2018-2019)
- IBM Data Science Professional Certificate Course Coursera 2019

Project:

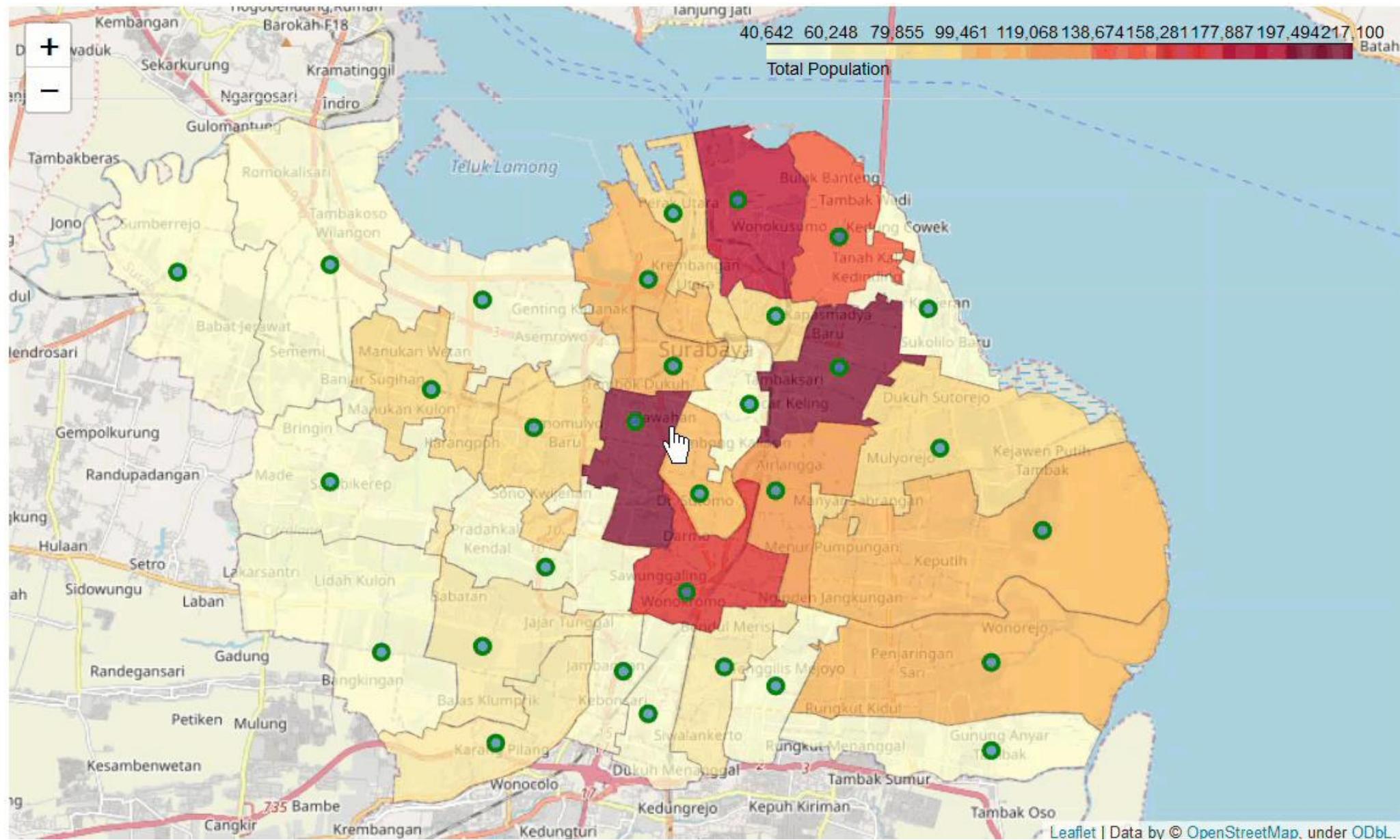
- Exploring Places and Area Regions in Surabaya (2019)



Muhammad Sifa'ul Rizky
rizkysifaul



Exploring Places and Area Regions in Surabaya (2019)



Python Conference 2019 Indonesia

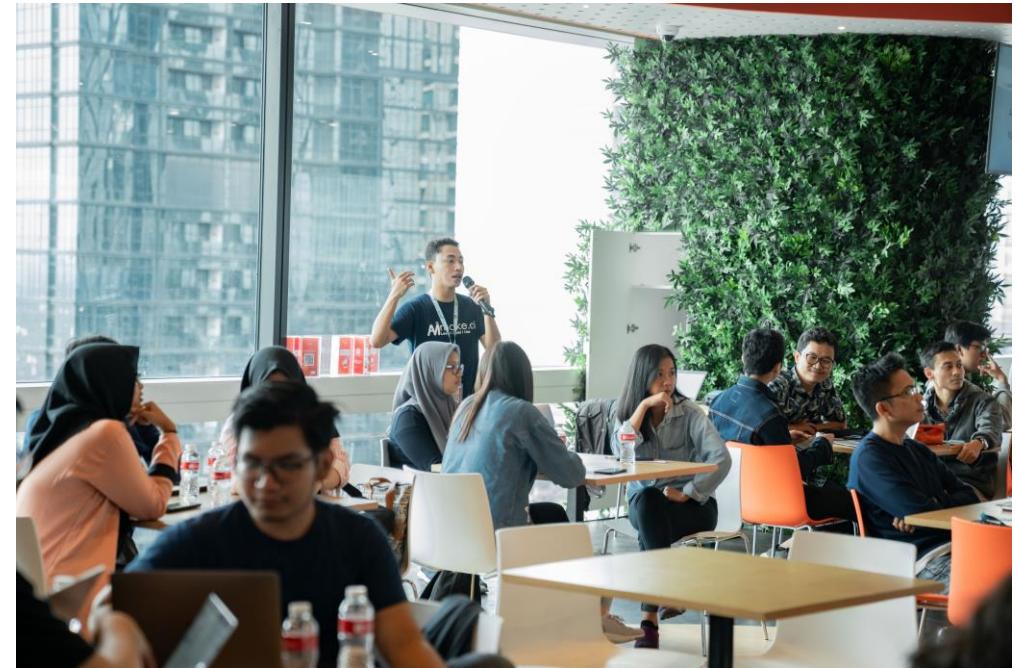


[Youtube Videos](#)

Tech Talk at Startup



Tokopedia Tech-a-Break
2019



Shopee for National Data Science
Competition Workshop
2020



Workshop at STT-PLN



Seminar at Angkasa Pura

Sharing with Many More



Seminar at UMN

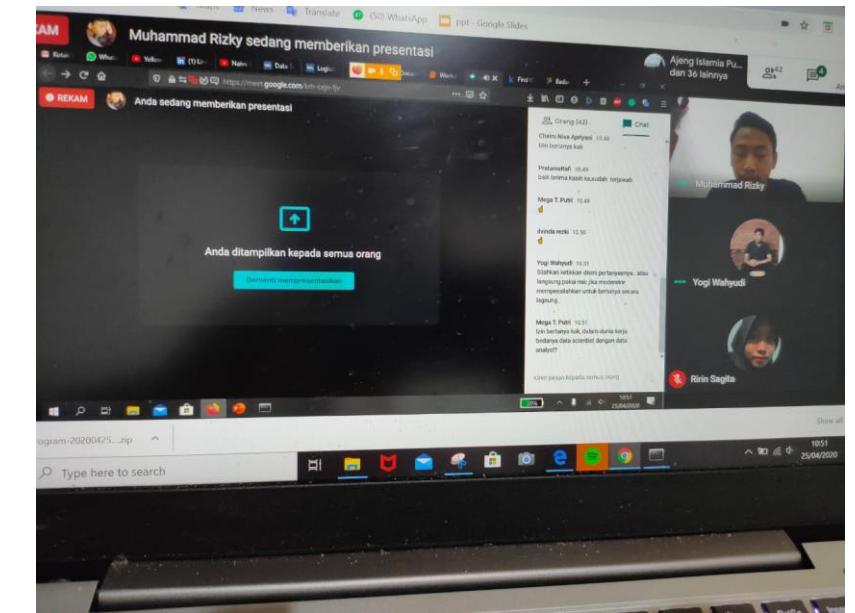
Sharing with Many More



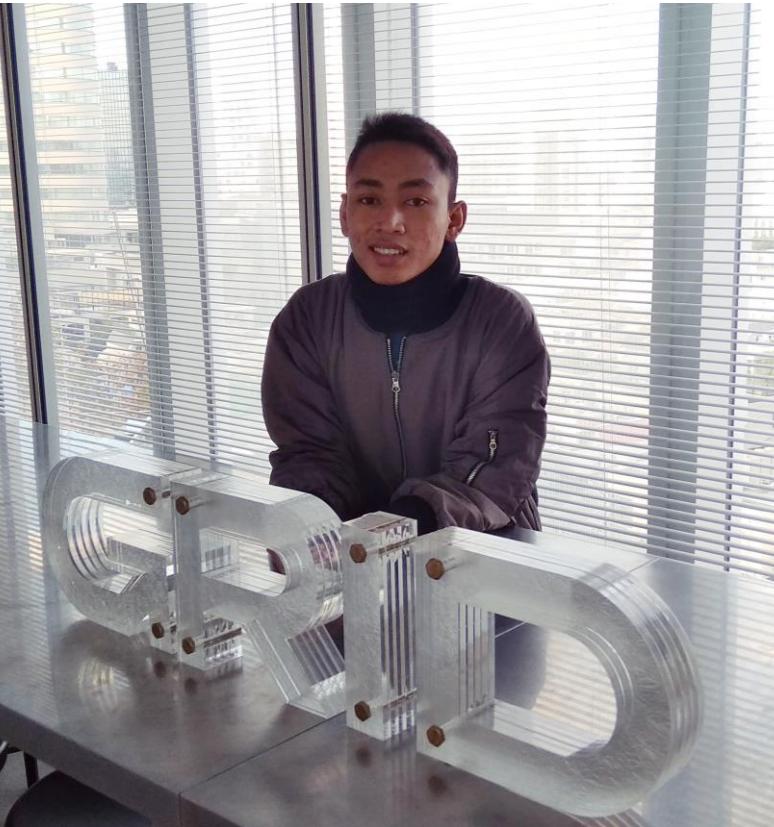
Workshop at
STMIK Nusa Mandiri



Seminar at ITS Surabaya
(My Hometown!!!)



Webinar with Sriwijaya University



Japan



Muhammad Sifa'ul Rizky



[rizkysifaul](#)



@rizkyfaul



msifaulkiki@gmail.com

Content Creator

Muhammad Sifa'ul Rizky 178 Followers · About



Published in Data Driven Investor · Sep 30

Recreating My Project Using Folium and Streamlit Part 2



ABOUT

Muhammad Sifa'ul Rizky

Dreamer, football,
and code. Loving
data science.

[Medium](#)



Coming soon!

The image shows a Spotify podcast page for 'Curious with Data'. The main title 'Curious with Data' is displayed in large white letters on a dark blue background. Below it, the subtitle 'Curious with Data' is shown. A green play button icon and a 'IKUTI' (Follow) button are visible. The page features a section titled 'Semua Episode' (All Episodes) with a preview of the first episode: 'Eps. 1 Perjalanan Data Science ala Anak Sistem Informasi'. The episode details mention it was published on 21 Sep and is 30 menit tersisa (30 minutes left). A 'Tentang' (About) section is also present, describing the podcast as discussing data and its various aspects without getting into technical details. The overall theme is data science and its applications.

[Spotify Podcast](#)

Data Science?



Harvard
Business
Review

Data | **Data Scientist: The Sexiest Job of the 21st Century**

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

Intro to Data Science

Do you know what is data science is all about?

Definition

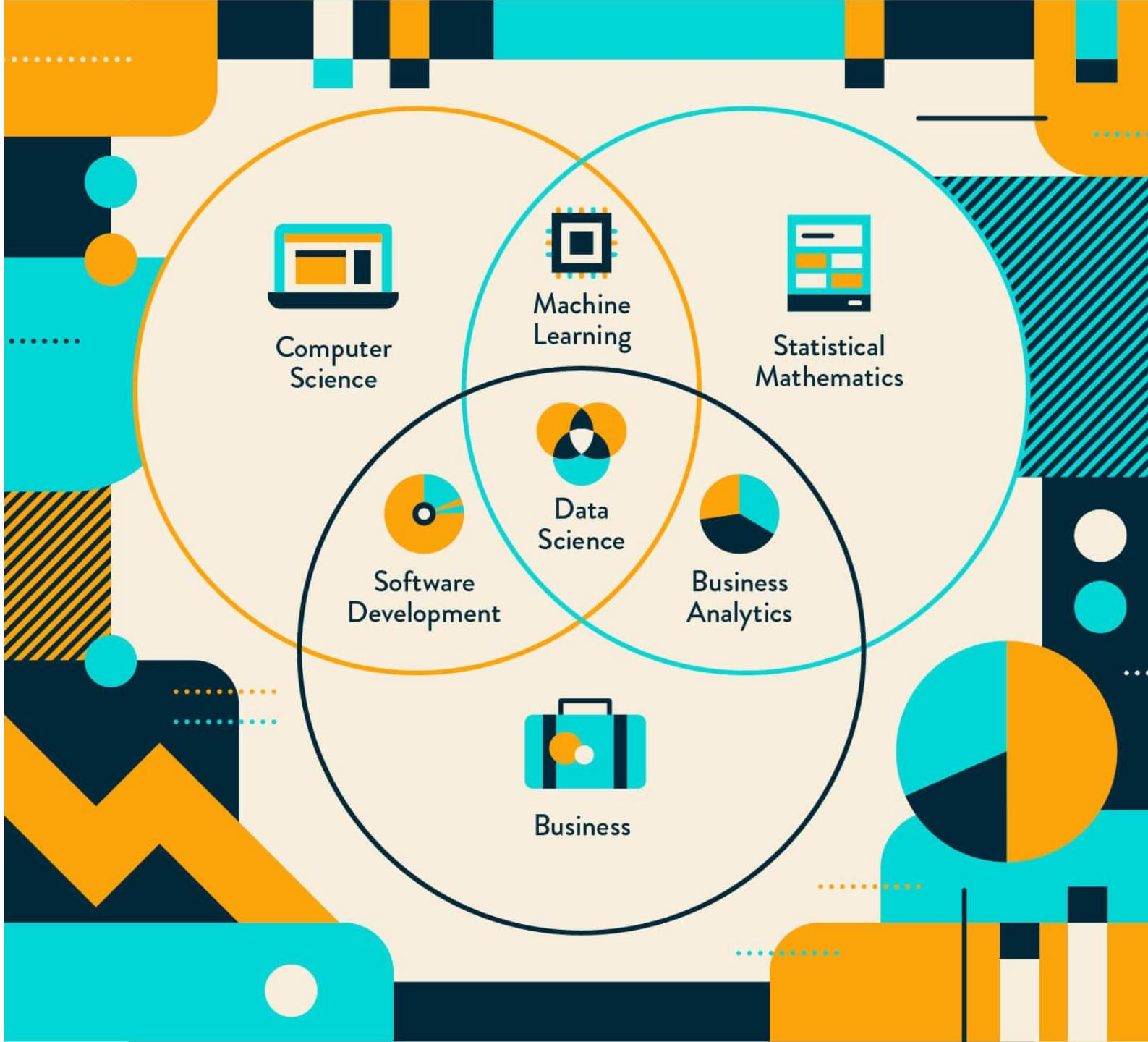
Data science is the study of data. It involves developing methods of recording, storing, and analyzing data to effectively extract useful information. ([Tech Terms](#))

Data scientist analyze, process, and model data then interpret the results to create actionable plans for companies and other organizations. ([Masters Data Science](#))

Data Science != Data Scientist

Why?

- Need variety skills
- From multiple subjects



source :
[Clevertap](#)

DIGITAL AROUND THE WORLD IN 2020

THE ESSENTIAL HEADLINE DATA YOU NEED TO UNDERSTAND MOBILE, INTERNET, AND SOCIAL MEDIA USE

TOTAL
POPULATION



7.75
BILLION

URBANISATION:

55%

UNIQUE MOBILE
PHONE USERS



5.19
BILLION

PENETRATION:

67%

INTERNET
USERS



4.54
BILLION

PENETRATION:

59%

ACTIVE SOCIAL
MEDIA USERS



3.80
BILLION

PENETRATION:

49%

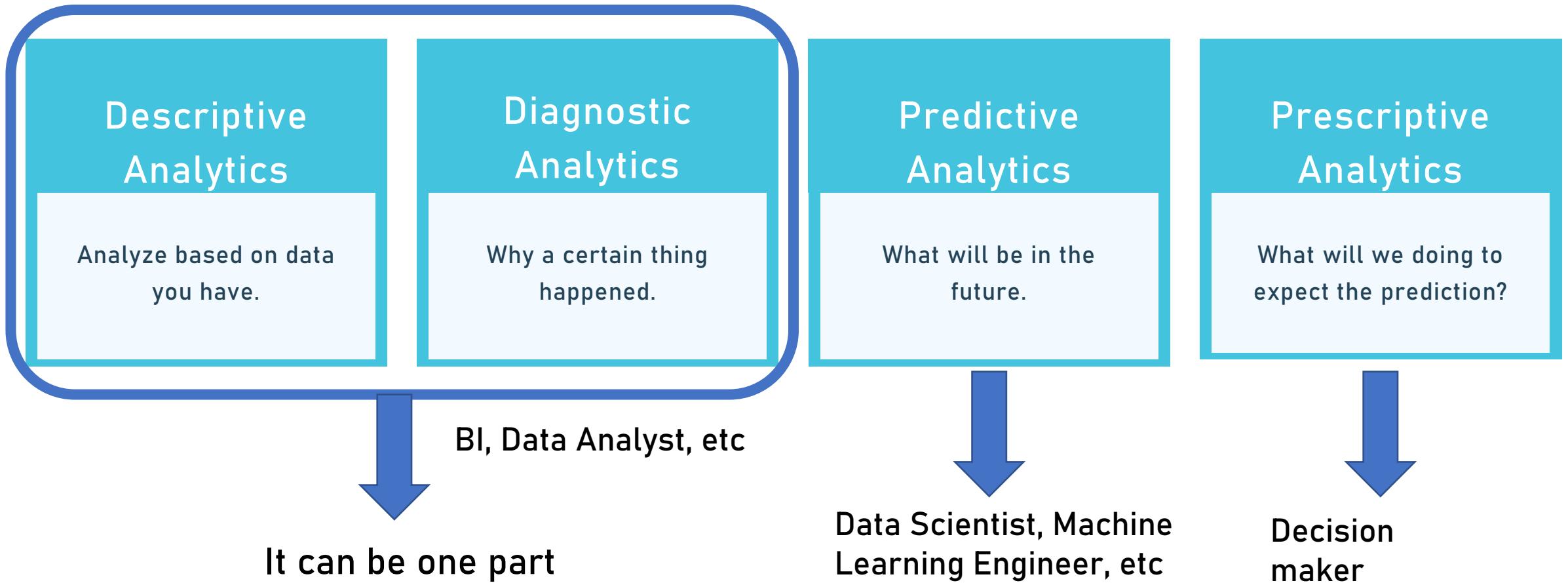


we
are
social

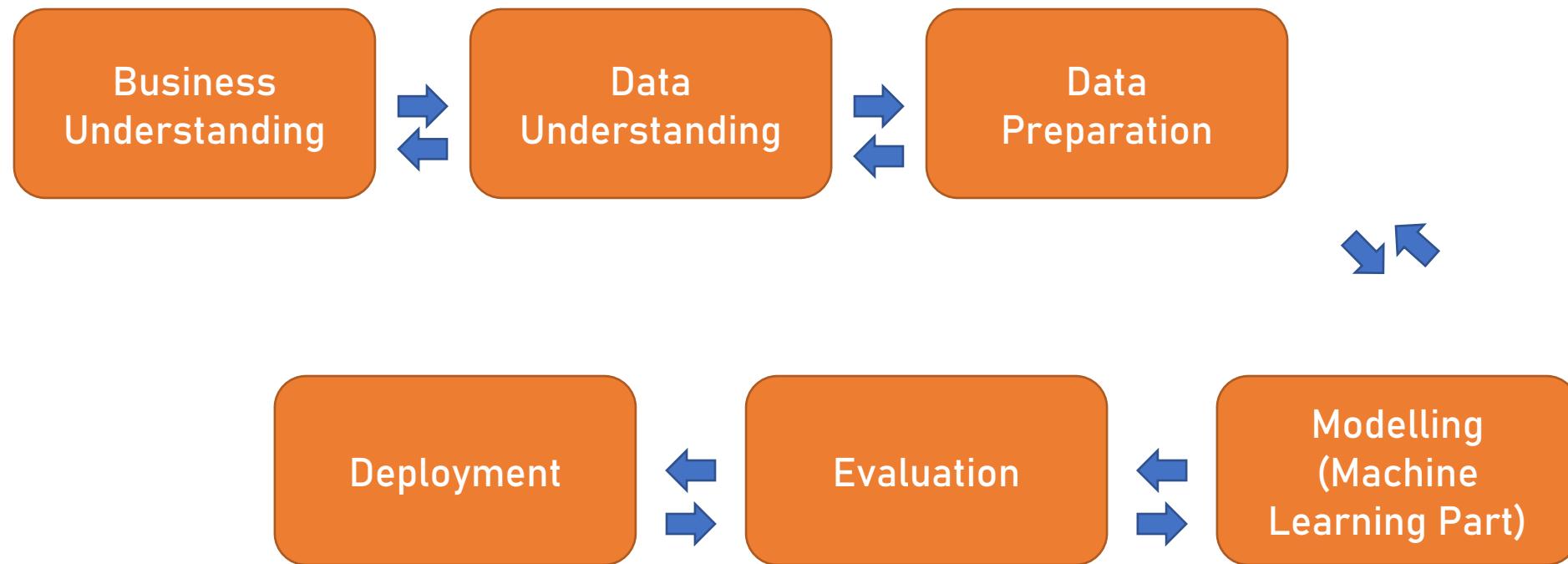


KEPIOS

Type of Data Analytics

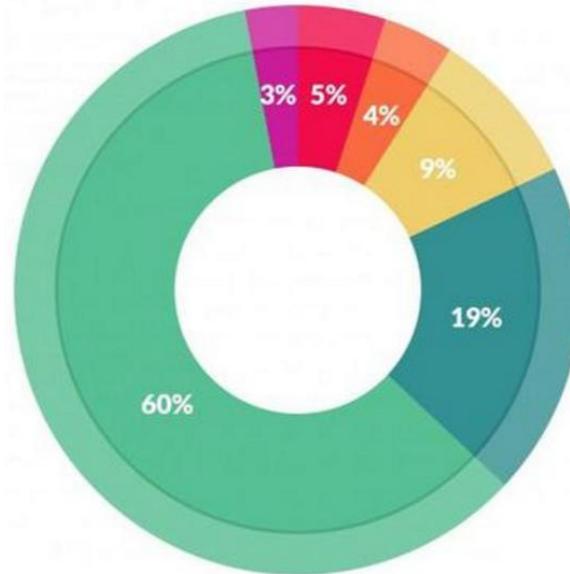


The Process of Data Science



Fun Fact

WHAT A DATA SCIENTISTS SPEND THE MOST TIME DOING?

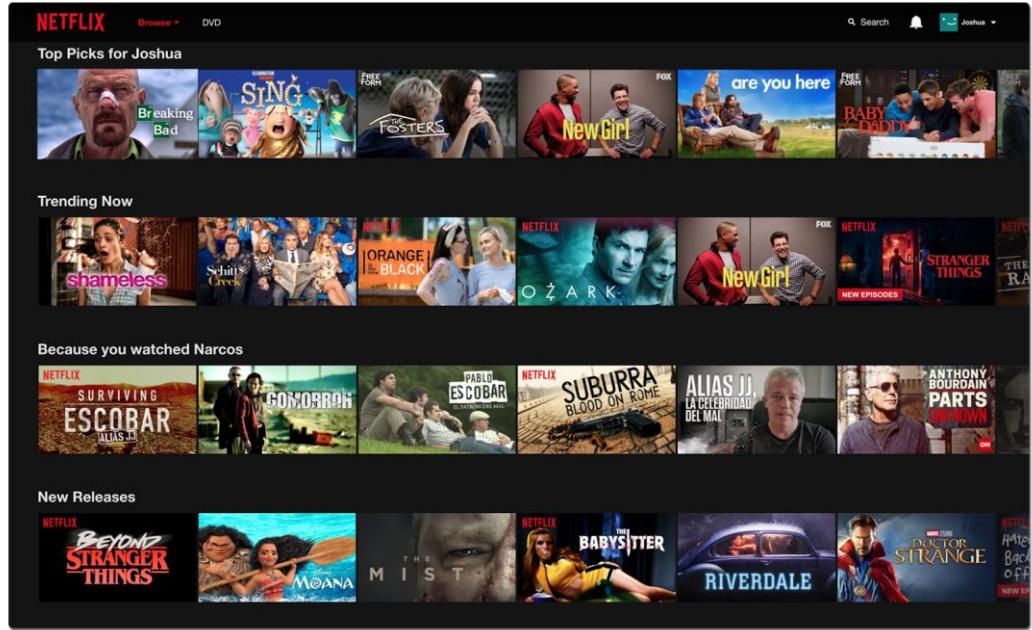


- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets; 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

REBBIS

source Forbes

What can you do?



Series/Video Recommendation



Fintech Credit Scoring

Data
Scientist

Data
Analyst

Data
Engineer

Machine
Learning
Engineer

Data
Science

Business
Analyst

Business
Intelligence

Who
knows?

Job Description

Roles/ Responsibilities	Data Scientist	Data Analyst	Data Engineer	ML Engineer	Business Analyst/ BI
Programming Tools	Very important	Somewhat important	Very important	Very important	Somewhat important
Data Visualization and Communication	Very important	Very important	Somewhat important	Somewhat important	Very important
Data Intuition	Very important	Very important	Very important	Very important	Somewhat important
Math & Statistics	Very important	Very important	Not that important	Very important	Somewhat important
Business Understanding	Somewhat important	Somewhat important	Not that important	Not that important	Very important
Application-based	Somewhat important	Somewhat important	Not that important	Not that important	Very important



Not that
important



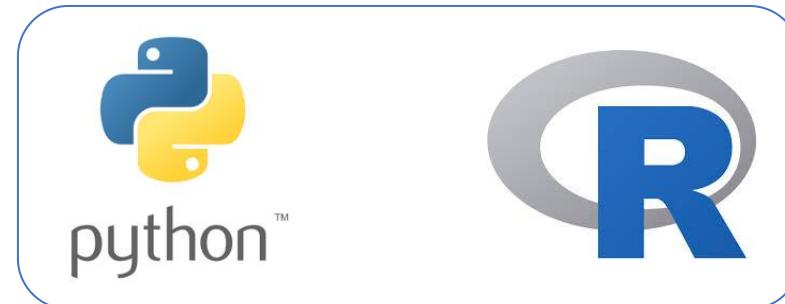
Somewhat
important



Very
important

Tools for your life

Data Analyst and Data Scientist



Business Analyst/ Intelligence



Data Scientist/ ML Engineer



Data Engineer/ ML Engineer/ Data Scientist



Google Cloud Platform





Study Case

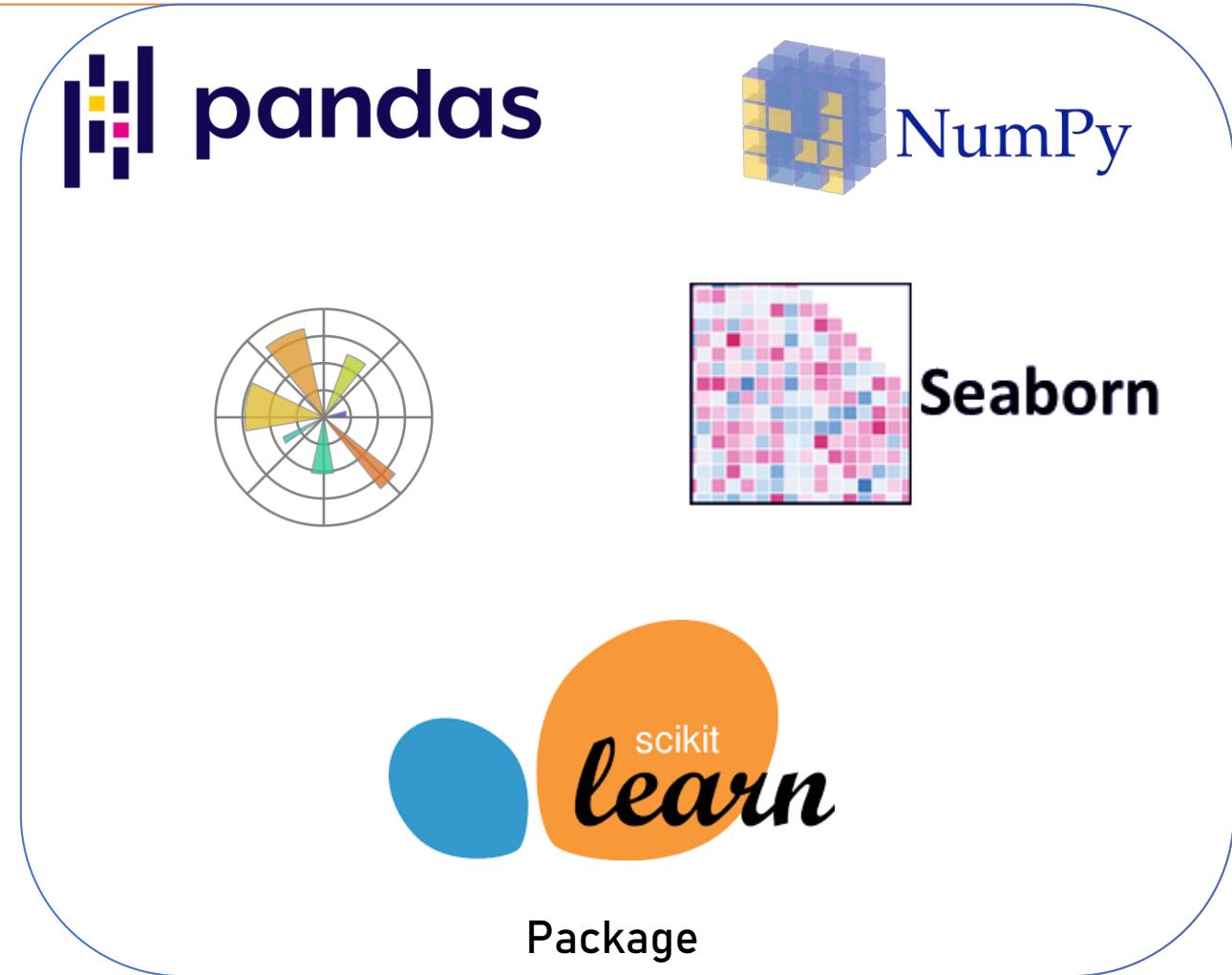
House Prices Dataset

Ps. You can see this file project [here](#)

Tools



Anaconda Navigator



Business Understanding

- We want to know how to predicting house prices and what important factor that makes prices more or less expensive.



IDR 300 Million



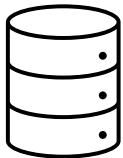
?

Data Understanding

- There are so many features (column) on this data like:
 - How many bedroom
 - How many bathroom
 - Area of living room
 - Year built and renovated (if any)
 - and many more...
- Based on our knowledge, price is depends by how big the house is, more bedroom means more pricey.

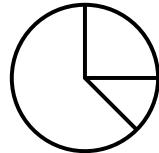
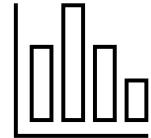
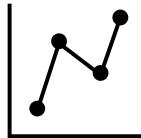
Data Preparation

- Data Gathering

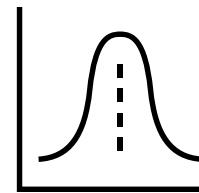
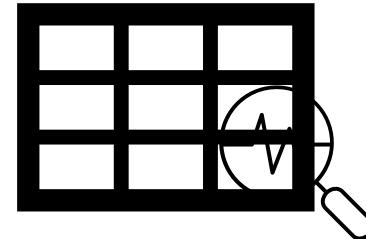


many more

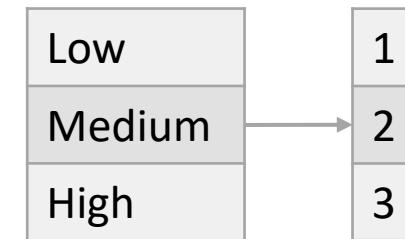
- Data Visualization



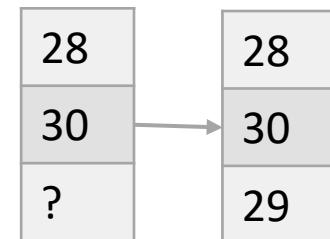
- Exploratory Data Analysis



- Data Cleansing/Preprocessing



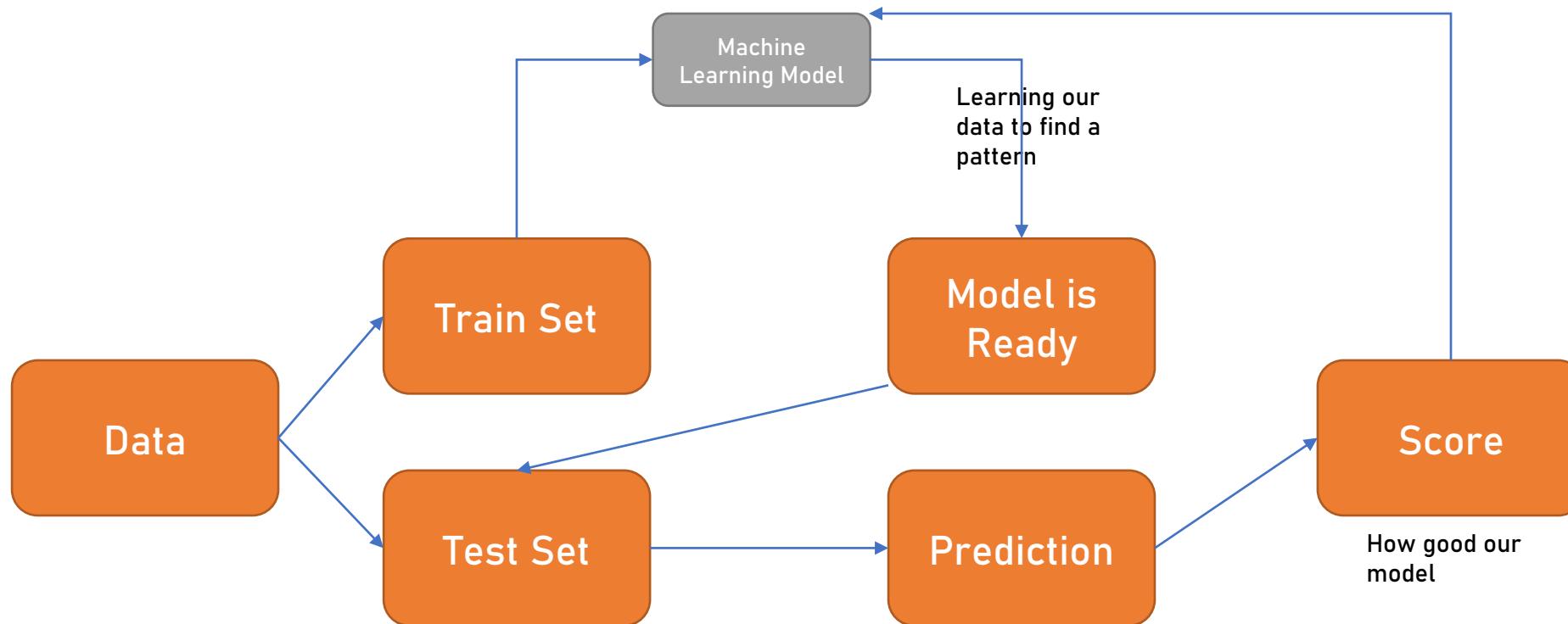
quality



age

Modelling (Machine Learning)

- Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.



Evaluation

- When we think our model is good enough, now do some real test



Sqft_living	Bedroom	Bathrooms	Floors	Year Built	Price
2000	2	3	2	1990	?

Deployment

- Show your projects!

Related article:

Here and here

Welcome to the House Price Prediction App

This is an application for knowing how much range of house prices you choose using machine learning. Let's try and see!

See the simple data

Now let's find out how much the prices when we choosing some parameters.

What is your square feet of living room?

2030

370

7320

How many bathrooms?

2

9

5

How many bedrooms?

3

9

9

How many floor do you want?

1

1

Bunuel

Mentoring (Private/ Groups) Class

Data Science with Python

- It can be for private or groups who want to learn data science.
- 2 hours/ session
- Discuss about what do you want to learn, or maybe skill up about data science project, and many more.

More information contact me on  @rizkyfaul

*“Without big data analytics,
companies are blind and deaf,
wandering out onto the web like deer
on a freeway.” (Gary King-Harvard)*

Danke schoen!

Terima kasih!

Python 101

How many people that still struggling with this language?

Do you think you are **beginner**, **intermediate**, or even **advanced**?

Starting with a simple things



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

Know your variable



```
#input  
money = 100000  
cash = 50000.0  
money + cash
```

```
#output  
150000.0
```



```
#input  
type(money)  
print("----")  
type(cash)  
-----  
  
#output  
int  
---  
float
```

Data type is important too

Containers is needed!

```
● ● ●  
#input  
name_single = "Faul"  
name_multi = ["Faul", "Arga", "Clara"]  
type(name_single)  
print("---")  
type(name_multi)  
-----  
#output  
str  
---  
list
```

Using **index** for access



```
#input  
print(name_multi[0])  
print(name_multi[1])  
print(name_multi[2])  
-----  
#output  
"Faul"  
"Arga"  
"Clara"
```

So many things inside

```
● ● ●  
  
#input  
gpa = 3.01  
if gpa > 3.5:  
    print("Great result!")  
elif gpa > 3:  
    print("Good one")  
else:  
    print("Don't give up, get better!")  
-----  
#output  
"Good one"
```

Iterative? No Problem



```
#input  
#for  
for number in range(5):  
    print(number)  
  
#while  
number = 0  
while number < 5:  
    print(number)  
    number = number + 1  
-----  
#output(for and while return same output)  
0  
1  
2  
3  
4
```

**More details?
Function is your
solution**



```
#input
def hitung_bulan(tahun):
    return tahun * 12

hitung_bulan(25)
-----
#output
300
```

Sneak Peek of Machine Learning

So, actually after we learn some basics of Python, now how would we applied machine learning to this?

Predicting House Prices (Simple but Annoying)



IDR 300 Million



IDR 450 Million



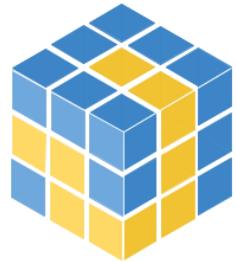
?

First, always
import your
package



```
#input
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
-----
#output
#nothing, it just imported your package
```

What is that?



NumPy



pandas



matplotlib



Your data
should be ready



```
#input
data = pd.read_csv('data.csv')
data.head()
-----
#output
#hmm what's is the output?
```

Output

◆	date ◆	price ◆	bedrooms ◆	bathrooms ◆	sqft_living ◆	sqft_lot ◆	floors ◆	waterfront ◆	view ◆	condition ◆	sqft_above ◆	sqft_basement ◆
0	2014-05-02 00:00:00	313000.0	3.0	1.50	1340	7912	1.5	0	0	3	1340	0
1	2014-05-02 00:00:00	2384000.0	5.0	2.50	3650	9050	2.0	0	4	5	3370	280
2	2014-05-02 00:00:00	342000.0	3.0	2.00	1930	11947	1.0	0	0	4	1930	0
3	2014-05-02 00:00:00	420000.0	3.0	2.25	2000	8030	1.0	0	0	4	1000	1000
4	2014-05-02 00:00:00	550000.0	4.0	2.50	1940	10500	1.0	0	0	4	1140	800



```
#checking name of columns  
print(data.columns)  
#checking how long column we have  
print(len(data.columns))
```

Check your columns

Output

- `Index(['date', 'price', 'bedrooms',
'bathrooms', 'sqft_living',
'sqft_lot', 'floors', 'waterfront', 'view',
'condition', 'sqft_above', 'sqft_basement',
'yr_built', 'yr_renovated', 'street', 'city',
'statezip', 'country'], dtype='object')`
- 18

How much is the price range?



```
data[['price']].describe()
```

Name	Price
count	4600.00
mean	551962.99
std	563834.70
min	0.00
25%	322875.00
50%	460943.46
75%	654962.50
max	26590000.00

Seems strange!

Anomalies(?)



```
#check what house that dont have price (price=0)
data[data['price'] == 0]
```

Price = 0?

So much!

Out [14]:

	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition
4354	2014-05-05 00:00:00	0.00	3.00	1.75	1490	10125	1.00	0	0	4
4356	2014-05-05 00:00:00	0.00	4.00	2.75	2600	5390	1.00	0	0	4
4357	2014-05-05 00:00:00	0.00	6.00	2.75	3200	9200	1.00	0	2	4
4358	2014-05-06 00:00:00	0.00	5.00	3.50	3480	36615	2.00	0	0	4
4361	2014-05-07 00:00:00	0.00	5.00	1.50	1500	7112	1.00	0	0	5
4362	2014-05-07 00:00:00	0.00	4.00	4.00	3680	18804	2.00	0	0	3

49 rows



```
#check what house that dont have price (price=26590000)
data[data['price'] == 26590000]
```

Price = 26590000

So few!

Out[16]:

	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition
	2014-07-03	26590000.00	3.00	2.00	1180	7793	1.00	0	0	4
4350	00:00:00									

1 row

What should you do?

Just go on

Do something

Analyze another row,
maybe ignoring or
do at the end of project

Delete price rows, or
just delete missing
value etc

I can't do this!
Why everything is so
heavy.

Do something

- Looking for upper and lower bounds

$$\text{Lower bound} = Q1 - 1.5 * (Q3 - Q1)$$

$$\text{Upper bound} = Q3 + 1.5 * (Q3 - Q1)$$



```
#what should you do?  
#in this case I will do something, drop value that not on range  
iqr = data['price'].describe()['75%'] - data['price'].describe()['25%']  
lower_bound = data['price'].describe()['25%'] - (1.5*iqr)  
upper_bound = data['price'].describe()['75%'] + (1.5*iqr)  
  
print("IQR equals {}".format(iqr))  
print("Lower bound of price is {}".format(lower_bound))  
print("Upper bound of price is {}".format(upper_bound))
```

How-to

Output

- IQR equals 332087.5
- Lower bound of price is -175256.25
- Upper bound of price is 1153093.75

How if Lower bound is minus?

So that all the values on lower should be count,
starting from > 0.



```
#just go on with data itself
data_clean = data.copy()
data_clean = data[(data.price > 0) & (data.price <= upper_bound)]
data_clean.shape
```

```
#output
(4311, 18)
```

Drop data price

col

date



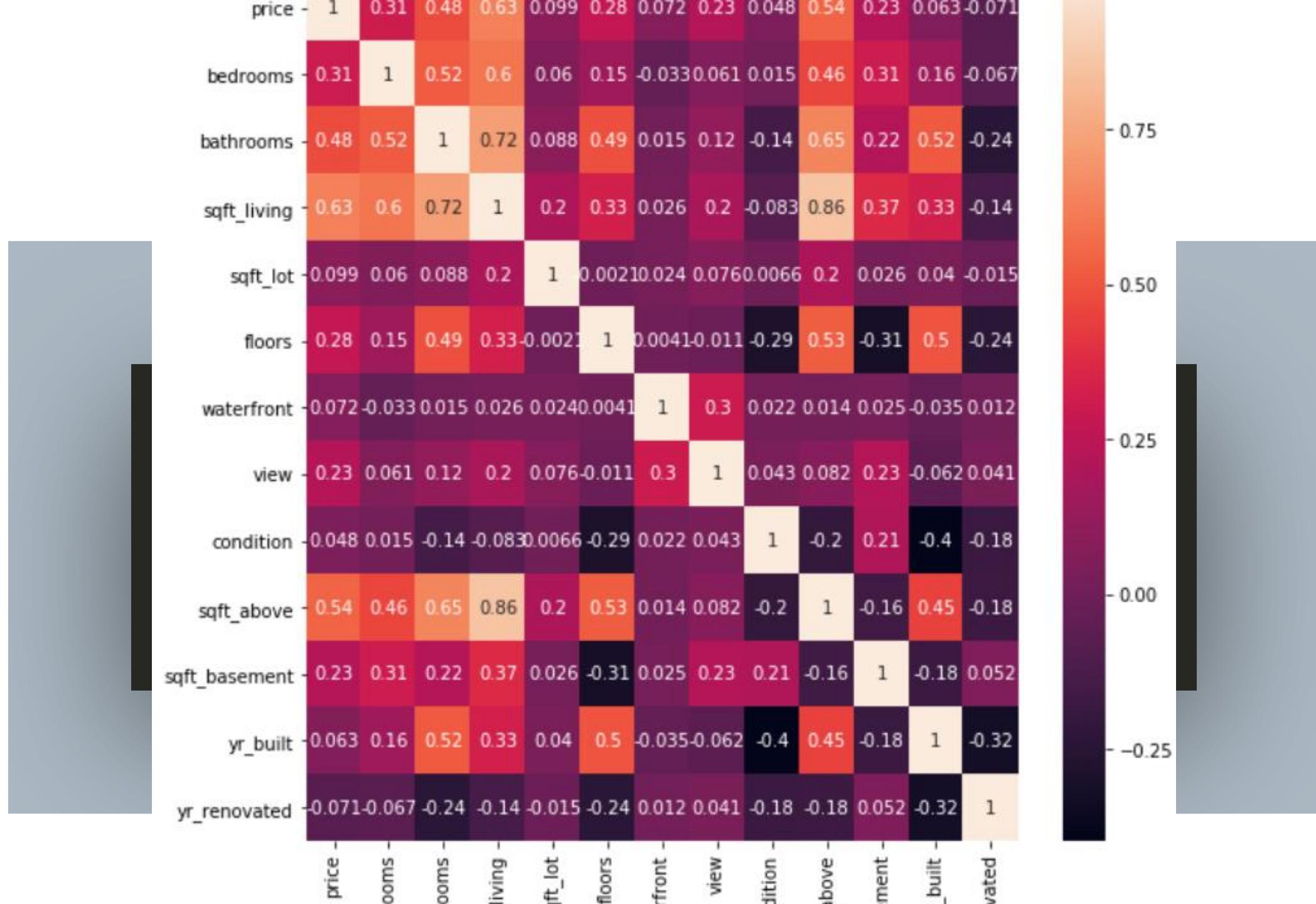
Run Interact



@int
def

(col))

false);



Simple selecting



```
#selecting bigger and/or smaller correlation plot  
data_clean = data_clean[['sqft_living','bathrooms','bedrooms','floors','price']]  
data_clean.head()
```

sqft_living	bathrooms	bedrooms	floors	price
1340	1.5	3.00	1.5	313,000.00
1930	2.00	3.00	1.00	342,000.00
2000	2.25	3.00	1.00	420,000.00
1940	2.50	4.00	1.00	550,000.00
880	1.00	2.00	1.00	490,000.00



```
#Import addionpackage
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split

X = data_clean.drop('price', axis=1)
y = data_clean['price']
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=.2, random_state=45)
```

Split your data

Modelling your data



```
#import your model  
model=LinearRegression()  
  
#fitting and predict your model  
model.fit(X_train, y_train)  
model.predict(X_test)
```

```
array([ 485354.69817553,  
       441835.3076801 ,  
       453789.98520846,  
              415049.11850156,  
       473551.53971655,  
       550880.38328669,  
              413405.93671759,  
       743722.90407821,  
       484311.20643497,  
              376271.3433367 ,  
       761396.97759918,  
       303086.79083996,  
              328005.09496555,  
       813657.50395897,  
       303050.37988799,...
```

Weight of each columns

```
● ● ●  
#coeficient for each column  
print(model.coef_)  
#bias of model  
print(model.intercept_)
```

$$\begin{aligned} price \\ = 179.3 * \text{sqft}_{\text{living}} \\ + 5527.7 * \text{bathrooms} \\ - 28018.6 * \text{bedrooms} \\ + 23832.5 * \text{floors} \\ + 170132 \end{aligned}$$

Scoring



```
#metrics scoring
from sklearn.metrics import mean_squared_error, r2_score
print("Root MSE of this model is {}".format(np.sqrt(mean_squared_error(y_test,model.predict(X_test)))))
print("R2 of this model is {}".format((r2_score(y_test,model.predict(X_test)))))
```

Root MSE of this model is 161854.5292971274
R2 of this model is 0.3675004503026602

How to understand

ID	Y_test	prediction	Error
1591	350,000.00	485,354.00	Error: -135,534
2164	240,000.00	441,835.31	Error: -201,835.31
1890	532,000.00 On the range	453,789.99	Error: 78,210.01

Range of price: 291,935.4 - 615,644.5

RMSE makes range between your prediction and error itself.