INSEA IT

INTRODUCTION TO DATA SCIENCE

A DATA SCIENTIST GUIDELINE

Full Course with code here :
https://github.com/INSEAIT/Data-Science

A Brief History

BIG DATA AND THE DATA SCIENCE HYPE

CAREER PROJECTION IN MOROCCO

A DATA SCIENTIST GUIDELINE

THE DATA SCIENCE TOOLS

PYTHON

R

JUPYTER NOTEBOOKS

Use model_selection.train_test_split from sklearn to split the data into training and testing sets. Set test_size=0.3 and random_state=101

In [12]:
```
from sklearn.model_selection import train_test_split
```

In [13]:
```
X_train,X_test,Y_train,Y_test = train_test_split(X,Y,test_size=0.3,random_state=101)
```

Training the Model

Now its time to train our model on our training data!

Import LinearRegression from sklearn.linear_model

In [14]:
```
from sklearn.linear_model import LinearRegression
```

Create an instance of a LinearRegression() model named lm.

In [15]:
```
lm = LinearRegression()
```

Train/fit lm on the training data.

In [16]:
```
lm.fit(X_train,Y_train)
```

Out[16]:
```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)
```

Print out the coefficients of the model

In [17]:
```
lm.coef_
```

Out[17]:
```
array([25.98154972, 38.59015875,  0.19040528, 61.27909654])
```

THE DATA SCIENCE ENVIRONNEMENT

ANACONDA

COLAB.GOOGLE.COM

THE DATA SCIENCE PROCESS

Things to keep in mind

EXTRACT DATA

Extract data from multiple sources, and formally organize it to have an overview of an event through multiple perspectives.

ANALYZE DATA

Find hidden correlations between multiple data, in order to understand and interpret an event, and come up with better processes.

PREDICT USING DATA

Use machine learning and statistical algorithms to make accurate future predictions using the old data and the correlations between them.

A DATA SCIENTIST GUIDELINE

The Data Science Process

EXTRACT DATA FROM MULTIPLE SOURCES

Full code : https://github.com/INSEAIT/Data-Science/tree/master/Exctraction

A DATA SCIENTIST GUIDELINE

PROBLEM ANALYSIS

You are tasked with figuring out how to increase your sales team’s performance. In our hypothetical situation, potential customers have rather spontaneous demand. When this happens, your sales team puts an order lead in the system. Your sales reps then try to get set up a meeting that occurs around the time the order lead was noticed. Sometimes before, sometimes after. Your sales reps have an expense budget and always combine the meeting with a meal for which they pay. The sales reps expense their costs and hand the invoices to the accounting team for processing. After the potential customer has decided whether or not they want to go with your offer, your diligent sales reps track whether or not the order lead converted into a sale.

TOOLS OF DATA EXTRACTION

BEAUTIFUL SOUP

PANDAS

OPENPYXL

A DATA SCIENTIST GUIDELINE

INSTALLATION

Of the python libraries

Pip install Pandas

Pip install openpyxl

Pip install beautifulsoup

Download Sheets

In [32]:
```
from oauth2client.service_account import ServiceAccountCredentials
import gspread
import pandas as pd
```

In [45]:
```
scope = [
    'https://www.googleapis.com/auth/spreadsheets',
]

GOOGLE_KEY_FILE = 'Medium_Data_Extraction_Key.json'

credentials = ServiceAccountCredentials.from_json_keyfile_name(GOOGLE_KEY_FILE, scope)
gc = gspread.authorize(credentials)

wokbook_key = '10HX66PbcGDvx6QKM8DC9_zCGp1TD_CZhovGUbtu_M6Y'
workbook = gc.open_by_key(wokbook_key)
sheet = workbook.get_worksheet(0)
values = sheet.get_all_values()
sales_data = pd.DataFrame(values[1:],columns=values[0])
```

In [36]:
```
credentials.get_access_token()
```

Out[36]:
```
AccessTokenInfo(access_token='ya29.c.ElpvBwlwBQLXzFUw97ASCsodQurmzrnxJGS0-hQCYCMOporyQ4FBZiDQJk2RkzPT
fQ_uhqattZqMKeYVuTLmtTyi69JfzT3jK5Rz8P4fdX6hwGllpoG6Qk1i6bQ', expires_in=3538)
```

Download csv

In [46]:
```
import requests
url = 'https://raw.githubusercontent.com/FBosler/Medium-Data-Extraction/master/sales_team.csv'
res = requests.get(url, allow_redirects=True)
with open('sales_team.csv','wb') as file:
    file.write(res.content)
sales_team = pd.read_csv('sales_team.csv')
```

In [ ]:
```
sales_team.head()
```

Time to install the openpyxl and xlrd libraries (uncomment the following)

In [52]:
```
#!pip install openpyxl
#!pip install xlrd
```

Download Excel Data

In [48]:
```
url = 'https://github.com/FBosler/Medium-Data-Extraction/blob/master/invoices.xlsx?raw=true'
res = requests.get(url, allow_redirects=True)
with open('invoices.xlsx','wb') as file:
    file.write(res.content)
invoices = pd.read_excel('invoices.xlsx')
```

In [51]:
```
invoices.head()
```

Out[51]:

	Meal Id	Company Id	Date of Meal	Participants	Meal Price
0	DUBEVHKJL58CD5RL	UJWAWDP4T329H04P	2018-07-14 12:00:00	['Lucy Salas']	147
1	6K4UB1OB146J0SDP	UJWAWDP4T329H04P	2014-05-28 09:00:00	['Melissa Rawlings' 'Cedric Pena']	456
2	L7PFX1ZO5XTO1G3O	UJWAWDP4T329H04P	2017-01-14 13:00:00	['Cedric Pena']	16
3	5K7IQ569G82RJ23Y	UJWAWDP4T329H04P	2018-10-04 13:00:00	['Melissa Rawlings']	595
4	BVRM8W1S0RPJ99VC	UJWAWDP4T329H04P	2018-03-31 07:00:00	['Melissa Rawlings']	359

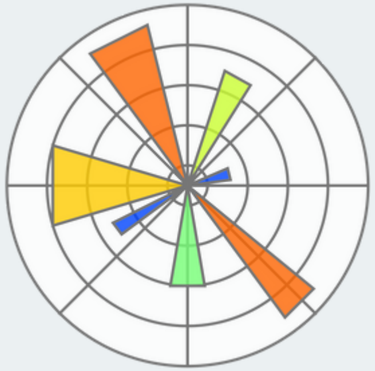
The Data Science Process

VISUALIZE & ANALYZE THE DATA

Full code : https://github.com/INSEAIT/Data-Science/tree/master/Analysis

A DATA SCIENTIST GUIDELINE

TOOLS OF DATA ANALYSIS



MATPLOTLIB

Pandas

PANDAS

Seaborn

SEABORN

A DATA SCIENTIST GUIDELINE

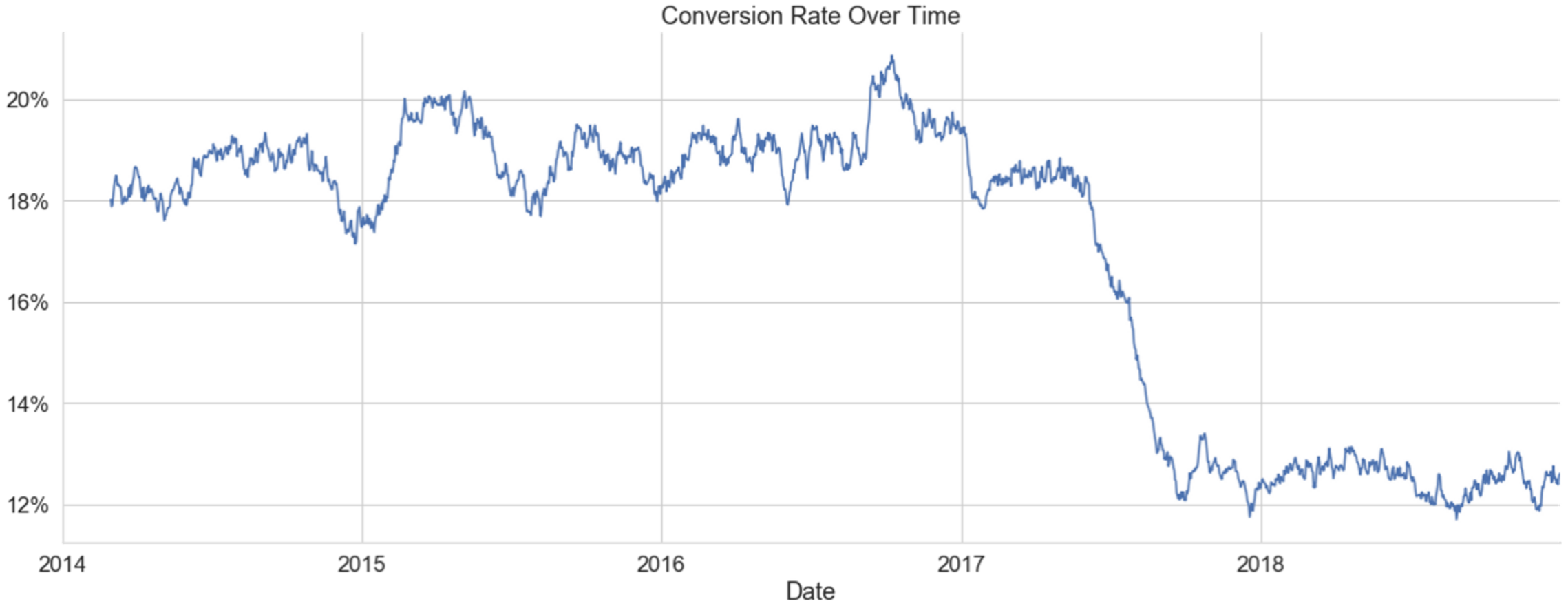
INSTALLATION

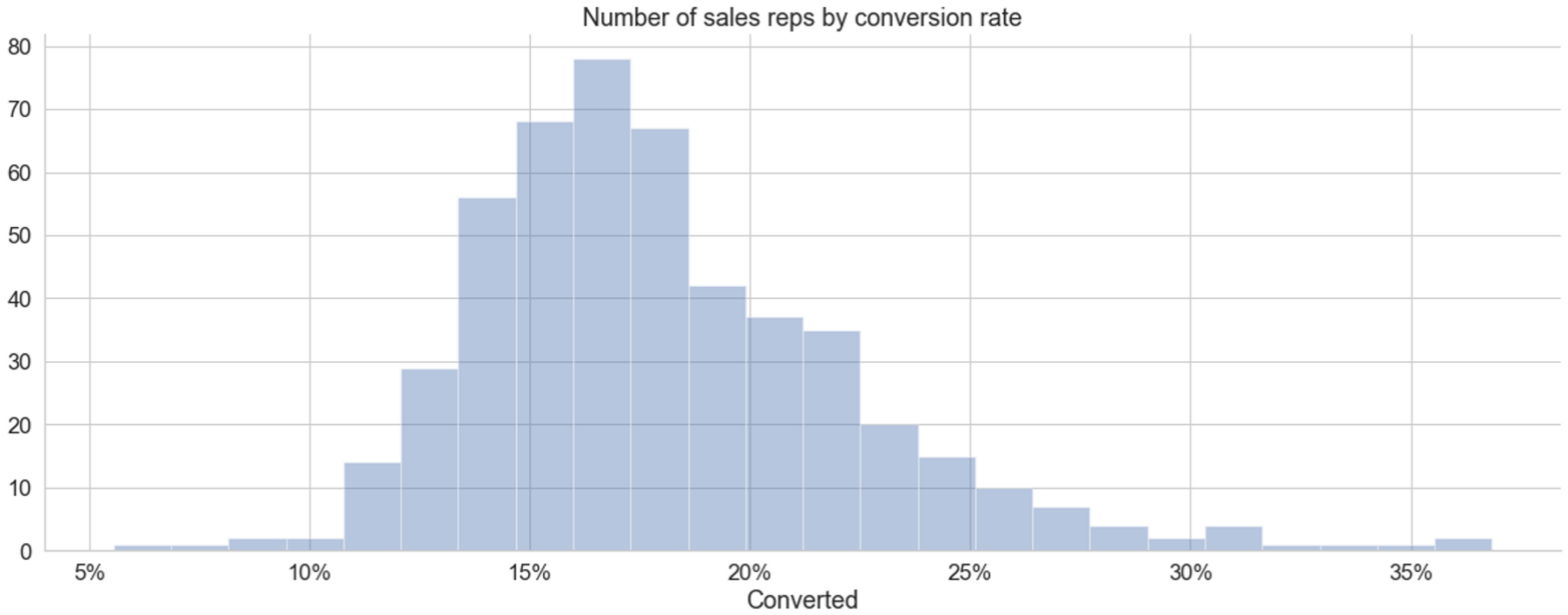
Of the python libraries

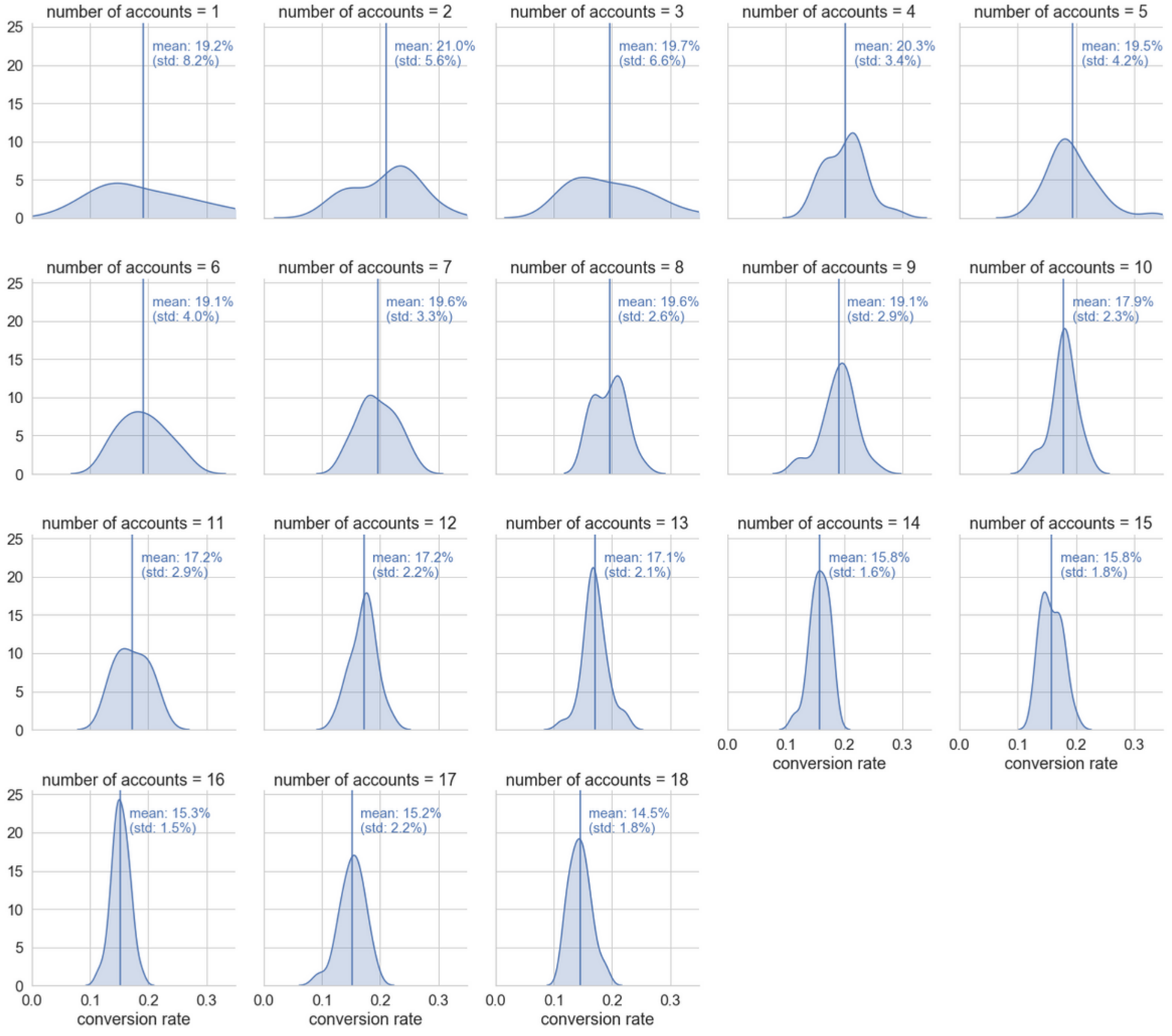
Pip install Pandas

Pip install seaborn

Pip install matplotlib


Conversion Rate Over Time
20%
18%
16%
14%
12%
2014
2015
2016
2017
2018
Date


Number of sales reps by conversion rate
80
70
60
50
40
30
20
10
0
5%
10%
15%
20%
25%
30%
35%
Converted


number of accounts = 1
mean: 19.2%
(std: 8.2%)
number of accounts = 2
mean: 21.0%
(std: 5.6%)
number of accounts = 3
mean: 19.7%
(std: 6.6%)
number of accounts = 4
mean: 20.3%
(std: 3.4%)
number of accounts = 5
mean: 19.5%
(std: 4.2%)
number of accounts = 6
mean: 19.1%
(std: 4.0%)
number of accounts = 7
mean: 19.6%
(std: 3.3%)
number of accounts = 8
mean: 19.6%
(std: 2.6%)
number of accounts = 9
mean: 19.1%
(std: 2.9%)
number of accounts = 10
mean: 17.9%
(std: 2.3%)
number of accounts = 11
mean: 17.2%
(std: 2.9%)
number of accounts = 12
mean: 17.2%
(std: 2.2%)
number of accounts = 13
mean: 17.1%
(std: 2.1%)
number of accounts = 14
mean: 15.8%
(std: 1.6%)
number of accounts = 15
mean: 15.8%
(std: 1.8%)
number of accounts = 16
mean: 15.3%
(std: 1.5%)
number of accounts = 17
mean: 15.2%
(std: 2.2%)
number of accounts = 18
mean: 14.5%
(std: 1.8%)
0 5 10 15 20 25
0.0 0.1 0.2 0.3
conversion rate

CONCLUSION

- Don’t give your sales reps more than 9 accounts (as conversion rate drops off rapidly)
- Make sure that every order lead is accompanied by a meeting/meal (as this more than doubles conversion rate
- Dinners are the most effective when there is only one employee from the customer
- Timing is key, ideally, your sales reps know as early as possible that a deal is coming up.

The Data Science Process

MAKE PREDICTIONS USING THE DATA

Full code : https://github.com/INSEAIT/Data-Science/tree/master/Prediction

A DATA SCIENTIST GUIDELINE

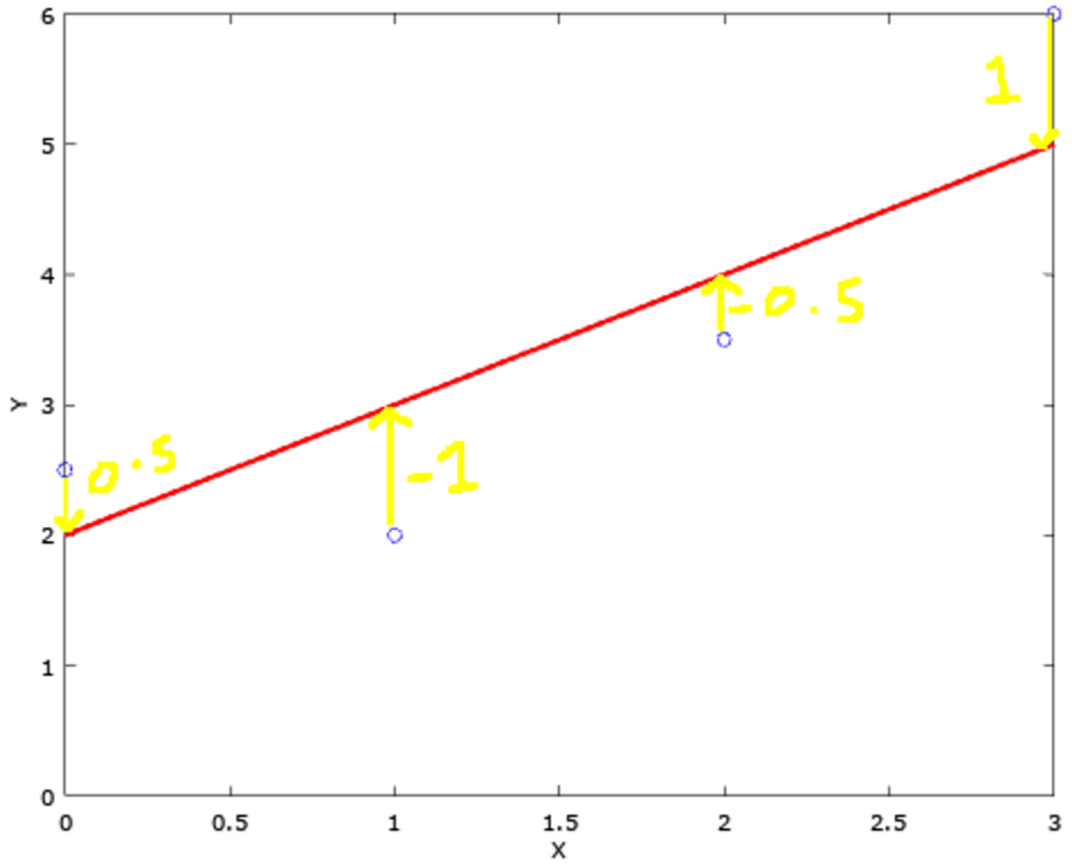
MACHINE LEARNING

LINEAR REGRESSION

A simple example of this could be a program that predicts a country's average yearly temperature from its proximity to the equator. In this case we have our X data (the proximity to equator), which should show some correlation with our Y data (the average yearly temperature). If we train our model with enough data, we can build a hypothesis that can estimate what the Y value (output) would likely be for any given value of X (input).

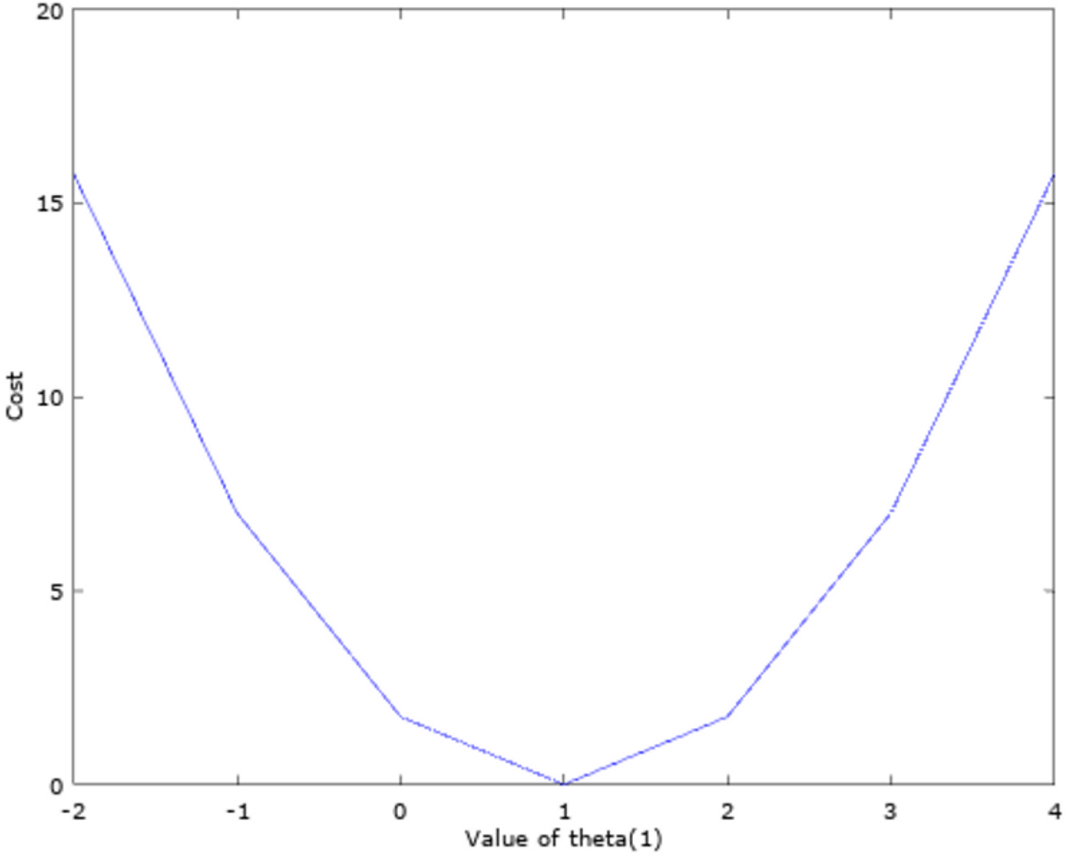
$$h_\theta(x) = \theta_0 + \theta_1 x$$

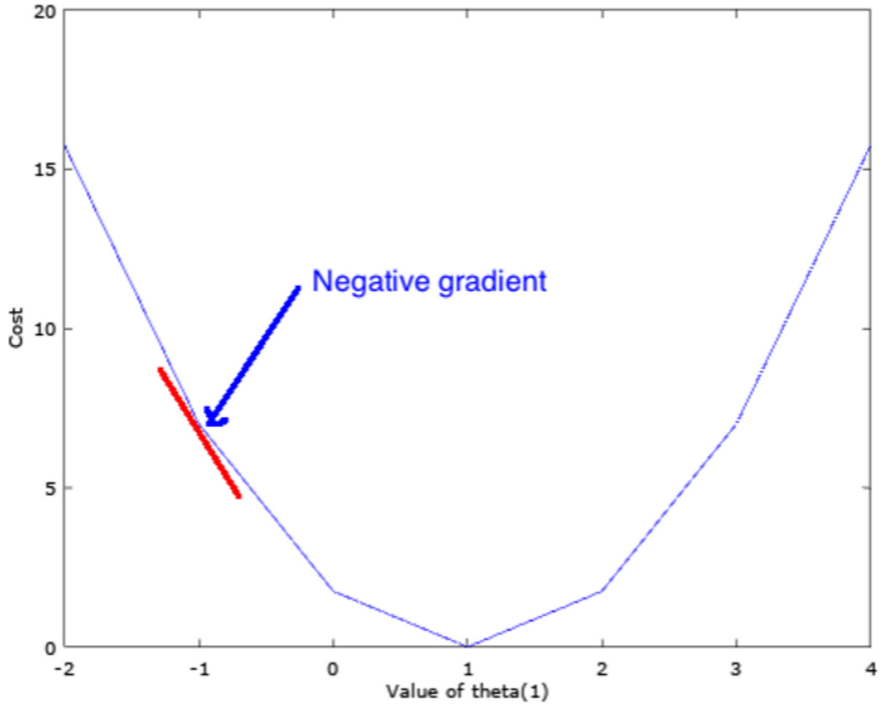
θ0 is the y intercept and θ1 is the gradient, while hθ(x) simply refers to the predicted output.



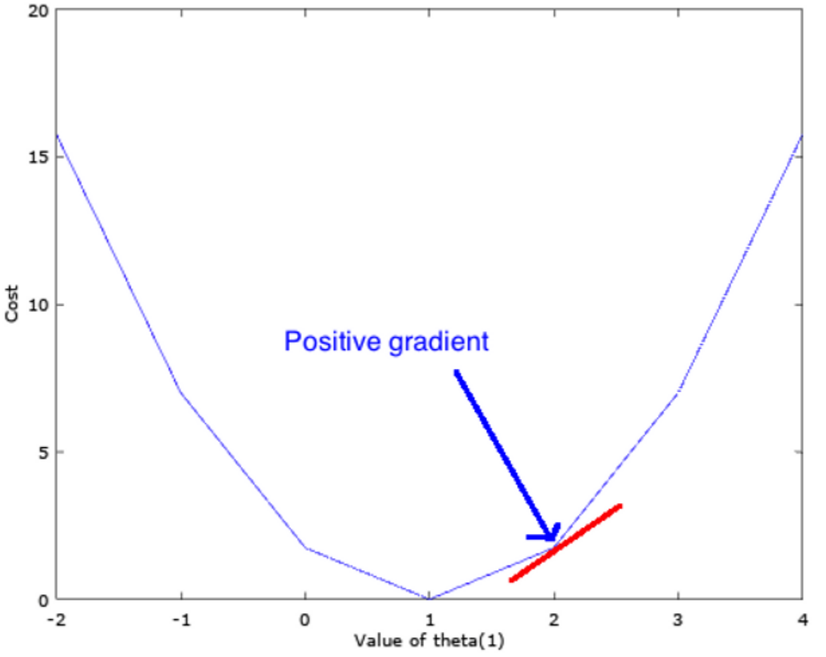
$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

J(θ) is the cost, worked out using this equation. hθ(x(i)) gives you the results predicted by your straight line, and then you subtract from them the actual values (y(i)) to get the costs, shown in yellow on the figure





If θ1 is -1, the cost is 7, and the gradient at this point would be negative. Therefore if we subtract this negative value from θ1, θ1 will get bigger (subtracting a negative makes a plus!), and move towards the vertex.



Conversely if θ1 is 2, the cost is 1.75, and the gradient at this point would be positive. Therefore if we subtract the value of the gradient from θ1, θ1 will get smaller, and move toward the vertex.

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$$

From our θ value we subtract the derived cost function (which tells us the gradient) multiplied by alpha. This takes the theta value closer to the vertex. In our example, we would carry out this process with j as 0 and then 1.

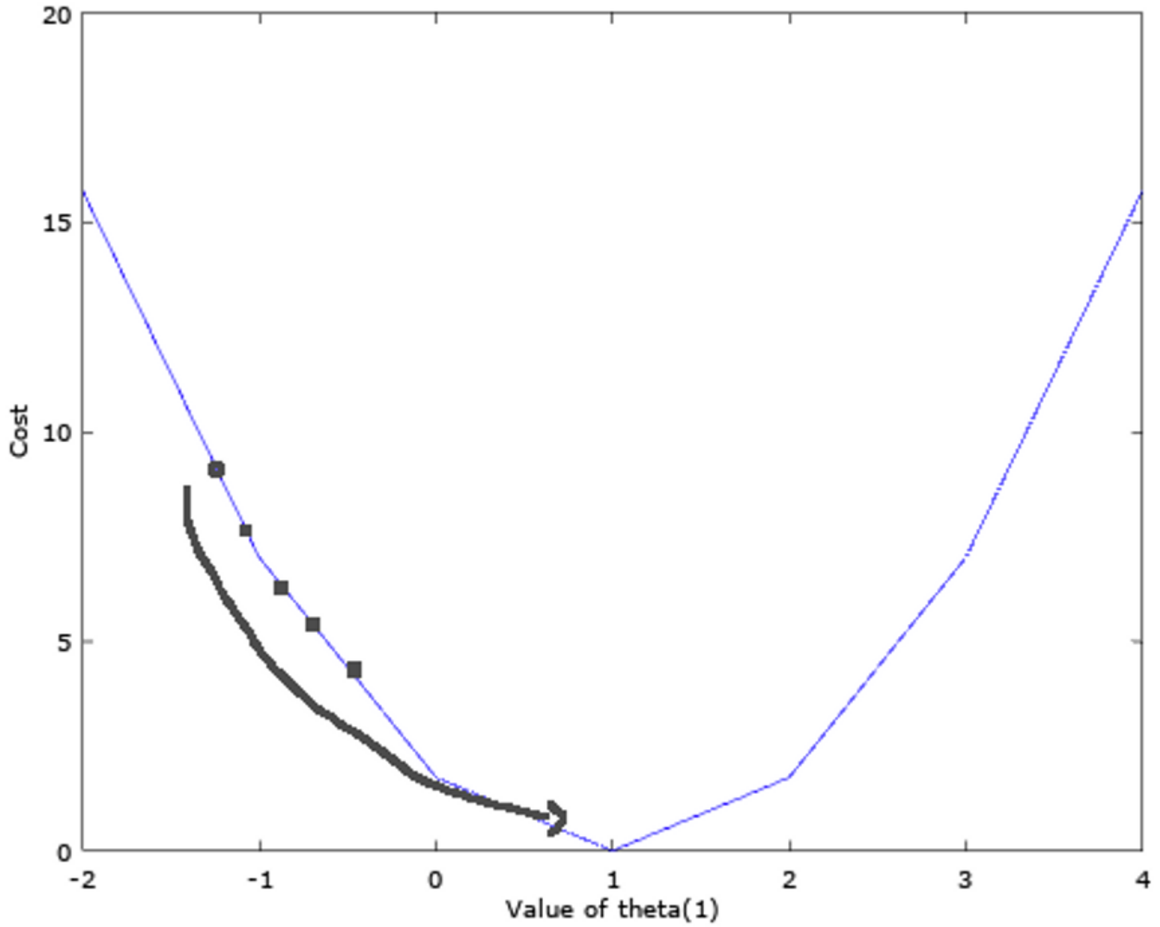
$$\theta_0 \; j = 0 : \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)$$

$$\theta_1 \; j = 1 : \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) \cdot x^{(i)}$$

From Andrew Ng machine learning Coursera course. In order to get the derived cost function (and so the gradient) for a value of θ0, you sum the cost, and multiply it by 1/the sample size. For θ1, you do the same, but multiply the costs by the x data, before summing it.

GRADIENT DESCENT ALGORITHM

$$
\begin{aligned}
&\text{repeat until convergence } \{ \\
&\quad \theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) \\
&\quad \theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) \cdot x^{(i)} \\
&\}
\end{aligned}
$$

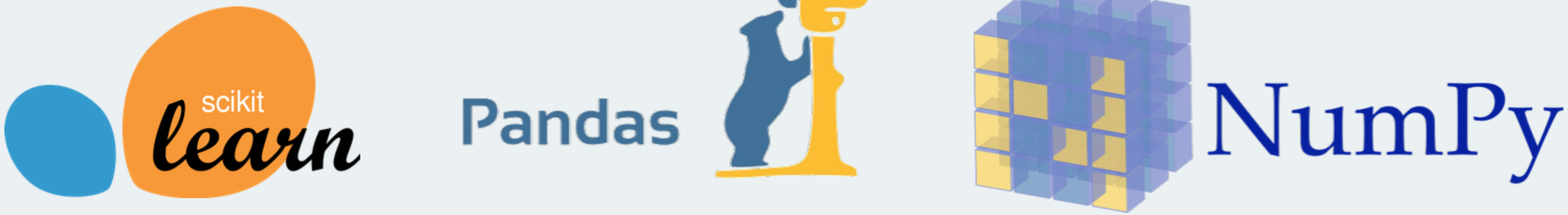


With each loop, the value of theta gets closer to the vertex.

$$h_\theta(x) = \theta_0 + \theta_1 x$$

θ0 is the y intercept and θ1 is the gradient, while hθ(x) simply refers to the predicted output.

TOOLS OF MACHINE LEARNING



SCIKIT-LEARN PANDAS NUMPY

A DATA SCIENTIST GUIDELINE

MACHINE LEARNING PRACTICE

Full code : https://github.com/INSEAIT/Data-Science/tree/master/Prediction

A DATA SCIENTIST GUIDELINE

INSEA IT

For more news and trainings

WEBSITE

https://inseait.com/

BLOG

https://medium.com/insea-it-blog

FACEBOOK PAGE

INSEA IT