Intro to Pandas and Data Science

SHWETA SHARMA

SENIOR SOFTWARE ENGINEER

WWW.AFFABLE.AI

AFFABLE INTRODUCES DATA ANALYTICS IN INFLUENCER MARKETING

Demographics

We have profiled 1 million users for their age and gender.

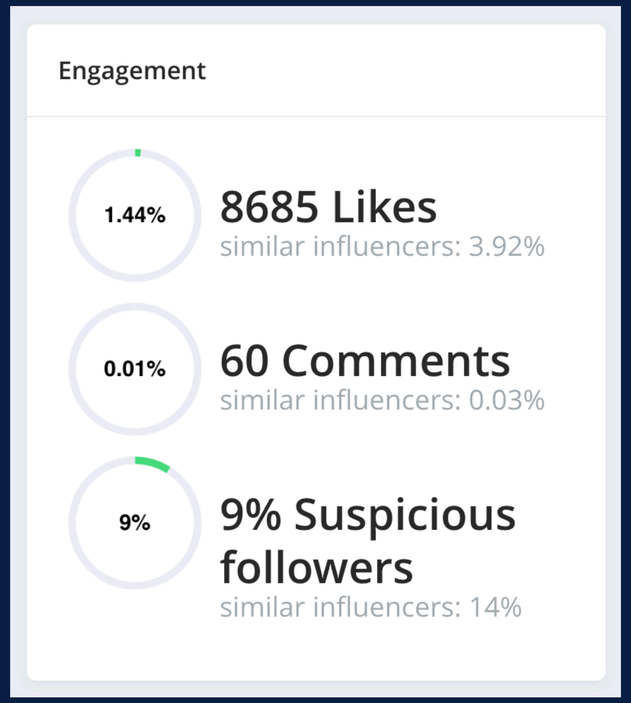
Interests

Discover authentic influencers based on the interest of their audience.

Brand Affinity

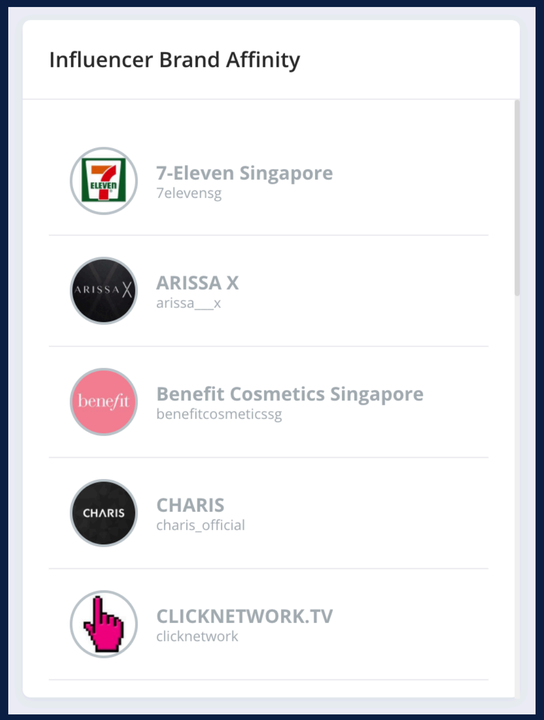
Find an influencer via the brands their audience follows.

MACHINE LEARNING BASED SUSPICIOUS ACCOUNT DETECTION



Suspicious Followers

For any particular influencer, you can check the % of suspicious accounts following them



Past Sponsored Posts

Look at the brands that the influencer has worked in the past for competitive analysis

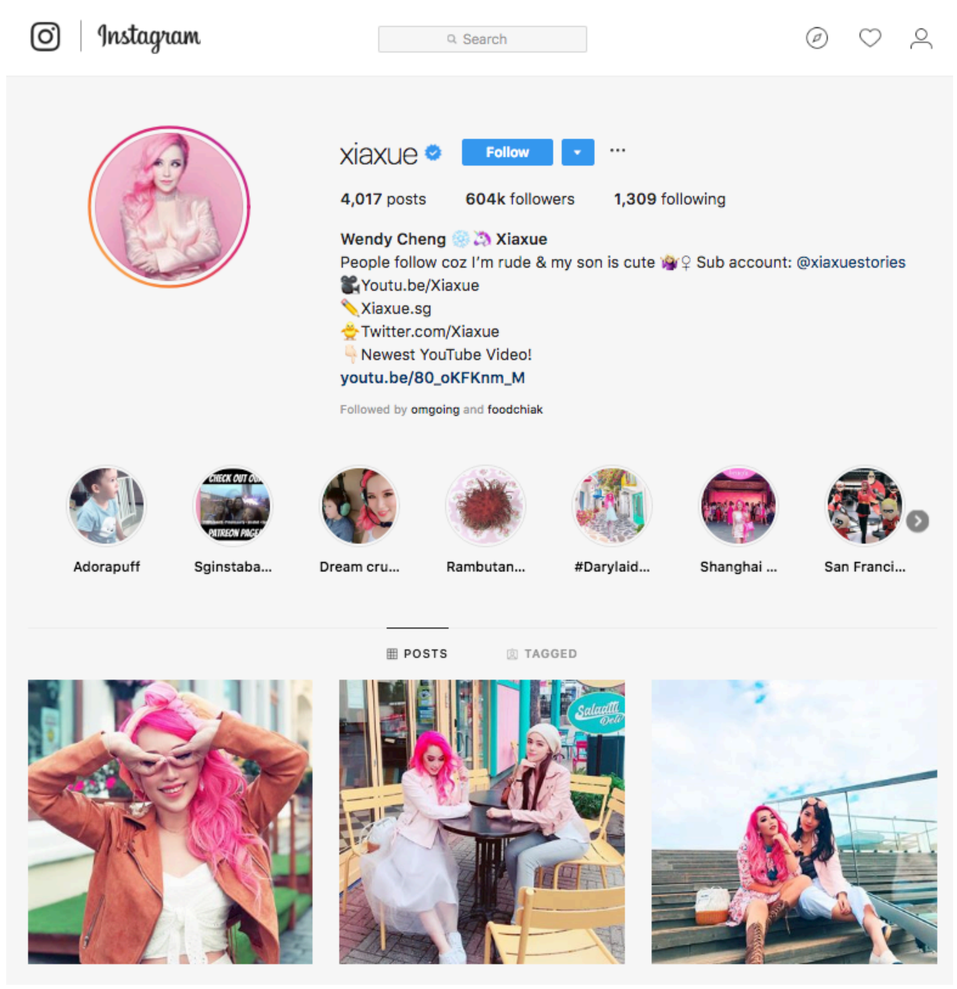
Guess who uploaded these images

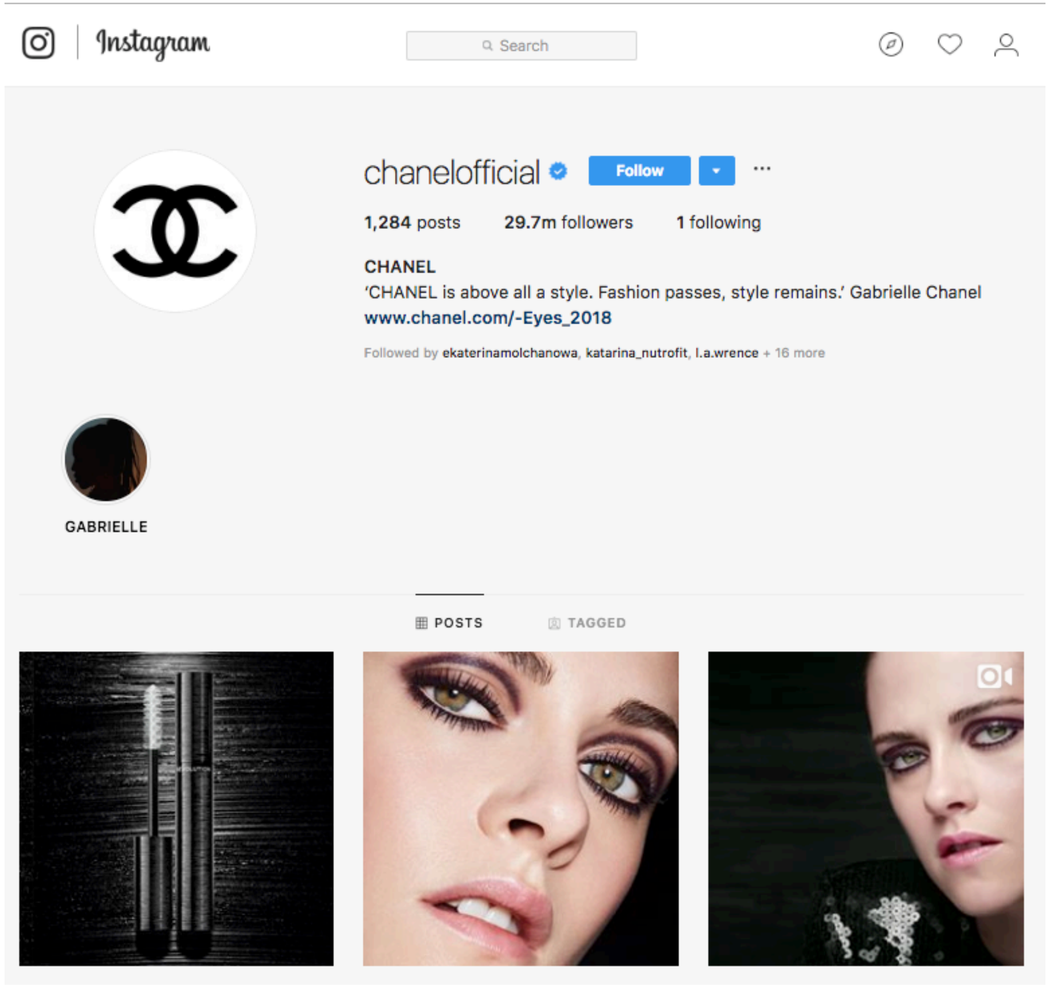
Guess who uploaded these images

Gigi Hadid

Tommy Hilfiger

One of the biggest Challenges we face?





AGENDA

1. Pandas

2. Preparing Data

3. Data Science

4. Types of Machine Learning

5. Data Modelling & Visualisation

6. Day in the life of a Data Scientist

🐼 PANDAS

Data Analysis Library in Python

Used for Manipulating and Analysing variety of data - CSVs, SQL DB, Dicts, Arrays

Integrates well with the sci-kit learn libraries of Python

IMPORTING DATA

```
import pandas as pd
users = pd.DataFrame('users.csv')
users.head()
```

```
users = pd.read_csv('users-new.csv', index_col='username')
users.head()
```

username	biography	isVerified	publicPhoneNumber	isBusiness	ui.shoppablePostsCount	addressStreet	picture	is
angelaxchen	!!aGLOBE TROTTER✈ \n♡ I Taipei x Singapore I ♡"	False	NaN	True	0.0	NaN	https://scontent-atl3-1.cdninstagram.com/vp/da...	
moxideofficial	"Nothing but Heavy Beats and Bass ☠ \n▲ BORN T...	False	9.126827e+07	True	0.0	NaN	https://scontent-atl3-1.cdninstagram.com/vp/2a...	
lianafinck	"email lianafinck@gmail.com to buy a redrawn o...	False	NaN	True	0.0	NaN	https://scontent-atl3-1.cdninstagram.com/vp/1c...	
leovieirabjj	"just Leo\n• @checkmathq \n• @dojosealbeach"	False	7.143626e+09	True	0.0	3290 E 19th St	https://scontent-atl3-1.cdninstagram.com/vp/1c...	
bpucci	"👊 MMA-@onechampionship \n🥋 BJJ blackbelt\n🏆 2x...	True	9.018596e+07	True	0.0	NaN	https://scontent-atl3-1.cdninstagram.com/vp/db...	

5 rows × 25 columns

DATA OVERVIEW

```
# Explore the users data statistically
users.describe()
```

```
users.describe()
```

	publicPhoneNumber	ui.shoppablePostsCount	latitude	longitude	mediaCount	contactPhoneNumber	followerCount	usertagsCount	geoM
count	2.900000e+02	822.000000	836.000000	835.000000	994.000000	2.900000e+02	9.940000e+02	9.940000e+02	
mean	6.571163e+09	9.537713	3.902322	20.466765	2176.989940	5.310285e+11	1.338838e+06	1.043603e+05	
std	1.899497e+10	53.718900	13.384915	54.121690	3761.219039	1.606560e+12	9.146392e+06	4.895629e+05	
min	1.316690e+05	0.000000	-37.885240	-122.828440	0.000000	6.113167e+07	4.000000e+01	0.000000e+00	
25%	6.839682e+07	0.000000	0.000000	0.000000	531.250000	6.572141e+09	1.004675e+04	4.375000e+02	
50%	1.283176e+08	0.000000	0.000000	0.000000	1173.500000	1.800401e+10	3.464900e+04	1.519500e+03	
75%	3.101603e+09	0.000000	1.293060	10.407229	2543.500000	6.681431e+10	1.712270e+05	1.196150e+04	
max	8.963707e+10	821.000000	55.676100	174.762590	60519.000000	6.518007e+12	2.490919e+08	8.465223e+06	

TERMINOLOGY

DataFrame is a 2-D structure with labelled columns.	Think Excel spreadsheet!
Series is a one-dimensional labeled array capable of holding any data type.	Think Excel column but with a labelled index.
Numpy is a Python library that is used for efficient computation on arrays.	Series values are Numpy arrays.

LET’S GET STARTED

1. Open google colab

2. Download and upload the jupyter notebook

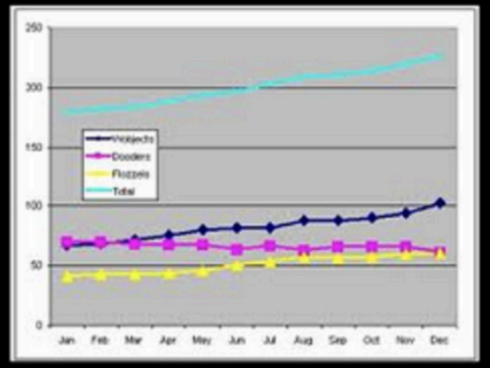
3. Upload the data file - users.csv

DATA SCIENTIST

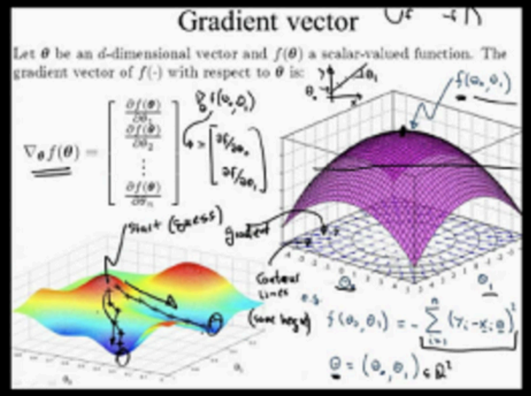
What my friends think I do

What my mom thinks I do

What society thinks I do



What my boss thinks I do



What I think I do

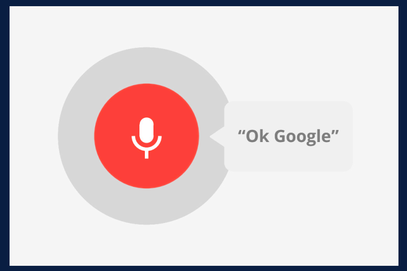
What I actually do

DATA SCIENCE - WHAT, WHEN, WHY

Making decisions and predictions from data

Involves Data Exploration, Data Modelling, Data Product Engineering

NETFLIX



TYPES OF MACHINE LEARNING - SUPERVISED

Learning by examples.

Observe what others have done in the past while making your next decision.

Example
What is the price of the house in 2019, if you know the prices of this house for the last 50 years.

TYPES OF MACHINE LEARNING - UNSUPERVISED

Observing a situation and trying to come up with best possible logic on the spot to decide.

Unsupervised is where we ask the computer to take decisions based on raw data attributes and a set of measurable quantities.

Example

Where should I open my 3 pizza joints if I know the locations of all the houses in the city.

TYPES OF MACHINE LEARNING - REINFORCEMENT

Learning from previous mistakes / success.

Computer starts with making random decisions, and then learns based on errors it makes and successes it encounters as it goes.

EX IN DATA MODELLING + VISUALISING THE RESULTS

Jupyter Notebook Exercise

THE DAY IN THE LIFE OF A DATA SCIENTIST

“Identify brand profiles on Instagram”

1. Frame the problem
2. Collect the raw data needed to solve the problem
3. Process the data
4. Explore the data
5. Perform in-depth analysis (machine learning, statistical models, algorithms)
6. Communicate results of the analysis

Q n A?

WWW.AFFABLE.AI

SHWETA@AFFABLE.AI