HW#1: Data Processing (13 points)

Cloud Computing and Big Data Analytics

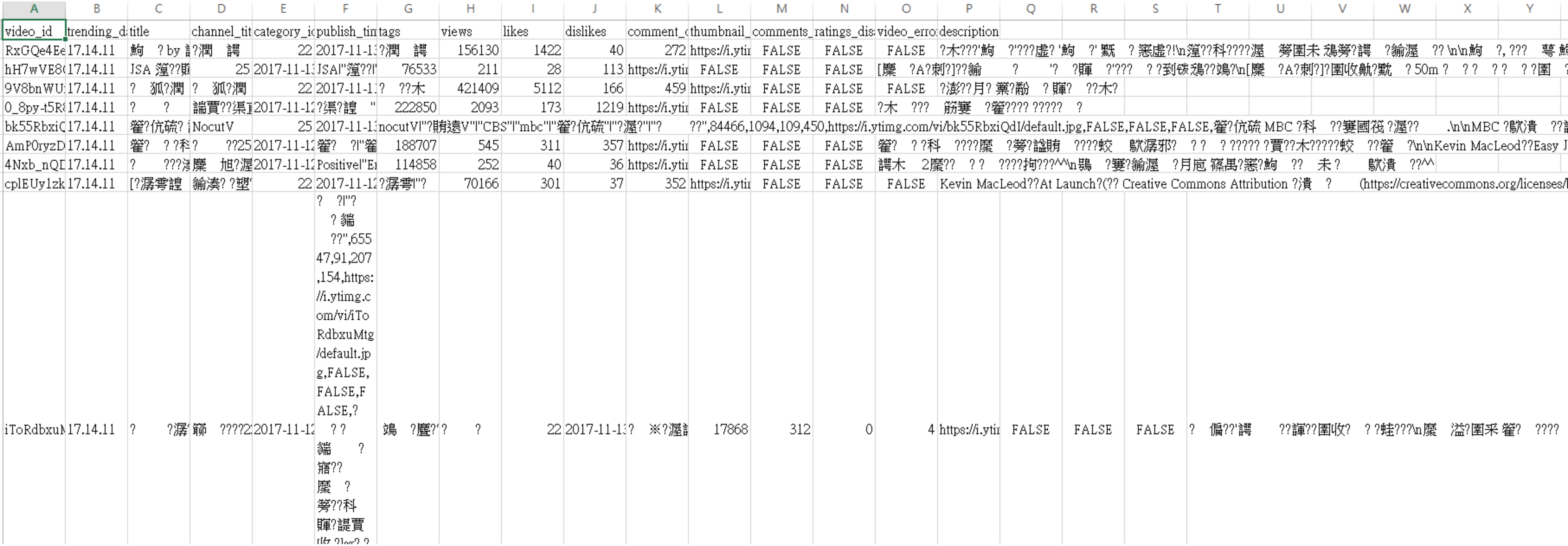
Dataset

- **Statistics for trending YouTube videos**

USvideos.csv

- video_id, trending_date, title, channel_title, category_id, publish_time, tags, views, likes, dislikes, comment_count, thumbnail_link, comments_disabled, ratings_disabled, video_error_or_removed, description

If you open with Excel, you will find a new world…



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
	video_id	trending_d	title	channel_tit	category_ic	publish_tin	tags	views	likes	dislikes	comment_c	thumbnail_	comments_	ratings_dis	video_erro	description
	RxGQe4Ee	17.14.11	鮈 ? by 訁	?潤 譁	22	2017-11-1:	?潤 譁	156130	1422	40	272	https://i.ytir	FALSE	FALSE	FALSE	?朩???'鮈 ?'???虛? '鮈 ?' 黖 ? 罻虛?!\n湟??科????渥 蒡圉未 鳰蒡?譁 ?緰渥 ?? \n\n鮈 ?, ??? 萼
	hH7wVE8(	17.14.11	JSA 湟??賟	25	2017-11-1:	JSAl"湟??l'	76533	211	28	113	https://i.ytir	FALSE	FALSE	FALSE	[縻 ?A?剌?]??緰 ? '? ?賱 ?'??? ??到铍鳰??鴳?\n[縻 ?A?剌?]?圉收魷?黖 ? 50m ? ?? ?? ??圉
	9V8bnWU:	17.14.11	? 狐?潤	? 狐?潤	22	2017-11-1:	? ??朩	421409	5112	166	459	https://i.ytir	FALSE	FALSE	FALSE	?澎??月? 蕖?豂 ? 賱? ??朩?
	0_8py-t5R8	17.14.11	? ?	諯賈??渠	2017-11-1:	?渠?諻 "	222850	2093	173	1219	https://i.ytir	FALSE	FALSE	FALSE	?朩 ??? 筋蔞 ?鶱???? ????? ?
	bk55RbxiC	17.14.11	鶱?伉硫?	NocutV	25	2017-11-1:	nocutVl"?賟遠V"l"CBS"l"mbc"l"鶱?伉硫"l"?渥?"l"? ??",84466,1094,109,450,https://i.ytimg.com/vi/bk55RbxiQdI/default.jpg,FALSE,FALSE,FALSE,鶱?伉硫 MBC ?科 ??蔞國筏 ?渥?? .\n\nMBC ?馱潰 ??諯									
	AmP0ryzD	17.14.11	鶱? ??科	? ??25	2017-11-1:	鶱? ?l"鶱	188707	545	311	357	https://i.ytir	FALSE	FALSE	FALSE	鶱? ??科 ????縻 ?蒡?謚賟 ????蛟 馱潺邪? ?? ????? ?賈??朩?????蛟 ??鶱 ?\n\nKevin MacLeod??Easy J
	4Nxb_nQI	17.14.11	? ???淺	縻 旭?渥	2017-11-1:	Positivel"Ei	114858	252	40	36	https://i.ytir	FALSE	FALSE	FALSE	譁朩 2縻?? ?? ????拘???^^\n鶚 ?蔞?緰渥 ?月庖 豯禺?罻?鮈 ?? 未? 馱潰 ??^^
	cplEUy1zk	17.14.11	[?潺雩諻	緰湊? ?塑'	22	2017-11-1:	?潺雩l"?	70166	301	37	352	https://i.ytir	FALSE	FALSE	FALSE	Kevin MacLeod??At Launch?(?? Creative Commons Attribution ?潰 ? (https://creativecommons.org/licenses/
	iToRdbxuM	17.14.11	? ?潺'	篩 ????2	2017-11-1:	? ?l"? ? 鍽 ??",655 47,91,207 ,154,https://i.ytimg.com/vi/iToRdbxuMtg/default.jpg,FALSE,FALSE,FALSE,? ?? 鍽 ? 寤?? 縻 ? 蒡??科 賱?諟賈	鴳 ?廱?'	? ?	22	2017-11-1:	? ※?渥	17868	312	0	4	https://i.ytir FALSE FALSE FALSE ? 倔??'譁 ??譁??圉收? ??蛙???\n縻 溢?圉釆 鶱? ????

Modules we can use…

```
import os
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

Answer the following questions

1. How many empty entries we have for each attributes (5%)?

- Hint: Google-isnull()

2. What are the average values of "likes","dislikes","views","comment_count“ in 2017? (10%)

- Hint: Google-groupby

3. Plot the boxplot of #dislikes for each month in 2017. (10%)

- Hint: Google-sns.boxplot

4. Plot the histogram of #views for each category in 2017 and 2018 in one figure. (10%)

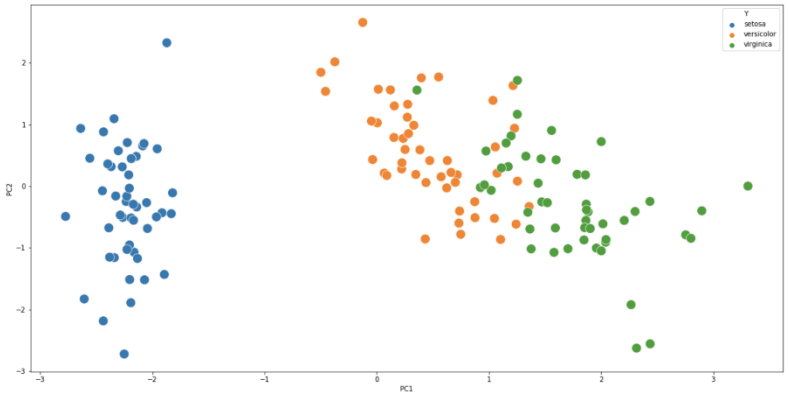
- Hint: Google-sns.barplot

5. Write a myPCA.py from scratch. (50%)

- Hint: https://towardsdatascience.com/principal-component-analysis-pca-from-scratch-in-python-7f3e2a540c51

Details

- myPCA.py is the function for PCA
  - Input: dataset path, target #dim Output: new data object
- In the main jupyter notebook,
  - select 50 samples from each category as the dataset, use “views, likes, dislikes, comment_count, comments_disabled, ratings_disabled, video_error_or_removed” as the features
  - output visualization of data points in 2-dim with color representing the category_id



6. Plot the word cloud of “title”. (15%)

Word Segmentation > Count > Cutoff > Visualization

- Hint: You can use “wordcloud”
- https://pypi.org/project/wordcloud/

Submission

- File Name: [studentID].ipynb, myPCA.py
- Deadline: 23:55. April 3rd, 2021

Lab 1: Data Cleaning

Cloud Computing and Big Data Analytics

Jupyter Notebook

- Blog
  - https://www.dataquest.io/blog/jupyter-notebook-tutorial/
- Tutorial Video
  - https://www.youtube.com/watch?v=HW29067qVWk
- Cloud Services
  - https://www.dataschool.io/cloud-services-for-jupyter-notebook/

Search in Dr

Google

My Drive

New folder

Upload files

Upload folder

Google Docs

Google Sheets

Google Slides

Google Forms

More

Google Drawings

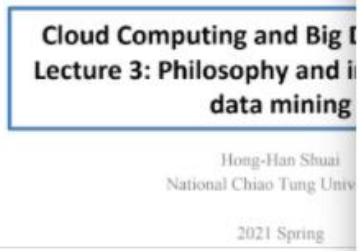
Google My Maps

Google Sites

Google Apps Script

Google Colaboratory

Google Jamboard



Lecture3.pptx

You opened today

BASIC LAB

Edited in the past month by Tam

Name

Owner

MAC_Backup me

data science 2020 me

Colab Test me

Lab1.ipynb

- https://colab.research.google.com/drive/1J9NfyYvM0Ti4TEojGQtX8dmjsz_As3RR?usp=sharing