Capstone Project 2: NLP Track - Youtube – Finding the most applicable content for a specific topic – Tom Preston

Study Background:

Millions of people use Youtube daily for learning how to do things. Diverse topics from how to replace a watch battery to how to learn Python is available however you often have to sift through lots of videos to find the “really good ones”. I personally use Youtube (YT) daily for finding training materials on Python and Data Science. My challenge is that YT’s search is helpful but the search options are based on “relevance“ (proprietary algorithm determined by YT), view count, and ratings. I would like a better way to assess and prioritize the content I am looking for.

My goal for this Capstone is to use the YT API to read in the video title, summary description, comments and likes / dislikes to find relevant content more quickly. I will use NLP techniques to review the document titles, summaries and comments to rank these videos to determine their relevance to my search criteria. I tested this on a couple subjects like python and golf instruction (subjects I personally spend a lot of time in YT looking at) to see if this approach improves upon my manual searches.

I defined a set of queries below that are meant to be broad topics around two different topic areas, python and golf. Below is the initial set of draft queries.

Python Queries	Golf Queries
Python tutorial	Hitting the Driver
Python reading CSV files	Bunker Shots
Python pandas DataFrames	Fairway Bunkers
Python lists	Putting tips
Python dictionaries	Stop hooking your driver
Python sort functions	Pitch shots
Python for loops	Chipping
Python tuples	Flop shots

Data Source

YT provides an API to search videos and extract statistics. The API accepts different types of queries and provides a nested dictionary of the results. There are a few challenges when using the YT API. First, you have to go to developers.google.com and register your email address to get an API key. This API key has to be part of the data query to YT. Second, the YT API query and related error messages are not very intuitive on how to use or troubleshoot.

YT API Method

```
api_key = 'xxxxxxxxxxxxxxxxxxxxxx'  # tom's API key
from apiclient.discovery import build
```

```
youtube = build('youtube', 'v3', developerKey=api_key)
```

This YT method creates a googleapiclient.discovery.Resource type when must be instantiated before running queries.

Data Wrangling:

Data Setup and Cleaning

The three main YT queries are video searches (using a search phrase), statistics query, and the comments query.

Video Searches: A call to the YT API with a search query (see below) returns a standard python dictionary of search requests.

```
query_results = youtube.search().list(
        part = 'snippet',
        q = ‘python tutorial’,
        order = 'relevance', # You can consider using viewCount
        maxResults = 50,    # max of 50 results returned
        type = 'video',         # Channels might appear in search results
        relevanceLanguage = 'en',
        safeSearch = 'moderate',
        ).execute()
```

The query above returns up to 50 videos based on relevance to the query. The “relevance” is a proprietary YT algorithm. The “query_results” response has a nested dictionary for each video (see below). Key items (shown below) are the video Id, title, and description. Also, this query has over 1,000,000 results and YT only allows you to request 50 results per query. A function using the nextPageToken allow a user to request a series of continuous results.

Query_results – first nested dictionary entry (key items bolded):

```
{'kind': 'youtube#searchListResponse',
 'etag': '"p4VTdlkQv3HQeTEaXgvLePAydmU/XzqcriHV8BFFo0Vqw6zP2YmSLYc"',
 'nextPageToken': 'CDIQAA',
 'regionCode': 'US',
 'pageInfo': {'totalResults': 1000000, 'resultsPerPage': 50},
 'items': [{'kind': 'youtube#searchResult',
   'etag': '"p4VTdlkQv3HQeTEaXgvLePAydmU/z11WbGTNgEvyDRL_DFX3UN5ZeyQ"',
   'id': {'kind': 'youtube#video', 'videoId': 'rfscVS0vtbw'},
   'snippet': {'publishedAt': '2018-07-11T18:00:42.000Z',
    'channelId': 'UC8butISFwT-Wl7EV0hUK0BQ',
    'title': 'Learn Python - Full Course for Beginners [Tutorial]',
    'description': "This course will give you a full introduction into all of the core concepts in
python. Follow along with the videos and you'll be a python programmer in no time!",
    'thumbnails': {'default': {'url': 'https://i.ytimg.com/vi/rfscVS0vtbw/default.jpg',
      'width': 120,
      'height': 90},
     'medium': {'url': 'https://i.ytimg.com/vi/rfscVS0vtbw/mqdefault.jpg',
      'width': 320,
      'height': 180},
     'high': {'url': 'https://i.ytimg.com/vi/rfscVS0vtbw/hqdefault.jpg',
      'width': 480,
      'height': 360}},
    'channelTitle': 'freeCodeCamp.org',
    'liveBroadcastContent': 'none'}},
```

Video Requests Data Wrangling. I initially queried the top 50 results for each search query. After reviewing the results, analyzing the first 25 queries is reasonable. As shown in appendix A, for all eight python queries, the first 5 – 10 typically have significantly more views. In my analysis, the queries 25 – 50 often do not have enough views or comments to provide much insights. The results 11 – 25 typically do have enough information (views, comments, etc) to be useful in the analysis.

The age of each video is calculated in weeks using the datetime module. It's assumed that newer videos will probably not be the most viewed however their ratio of likes to views might identify promising new videos which are exactly the ones I am trying to find by adjusting the search criteria. The video request information is exacted into a list and then converted into a pandas DataFrame to be merged with the statistics information.

Video Statistics: The second query is the video statistics query where a video id (each YT video has a unique alpha-numeric video id) is required to retrieve the video statistics (view count, like count, dislike count and comment count). The video IDs are exacted from the initial video search query results and then combined into a single string. The YT statistics query allows a user to query video statistics either via single or group requests. To be a "good" user, I ask for the video statistics in one group request instead of 25 individual requests for each of the 8 python queries.

Video Statistics Query:

```
# create 1 request string with all video Ids
for i in range(0,len(video_id),50):
        video_id_request = ','.join(video_id[i:i+50])

#
# request stats for all video ids
#
```

```
res_stats = youtube.videos().list(id=video_id_request, part='statistics').execute()
```

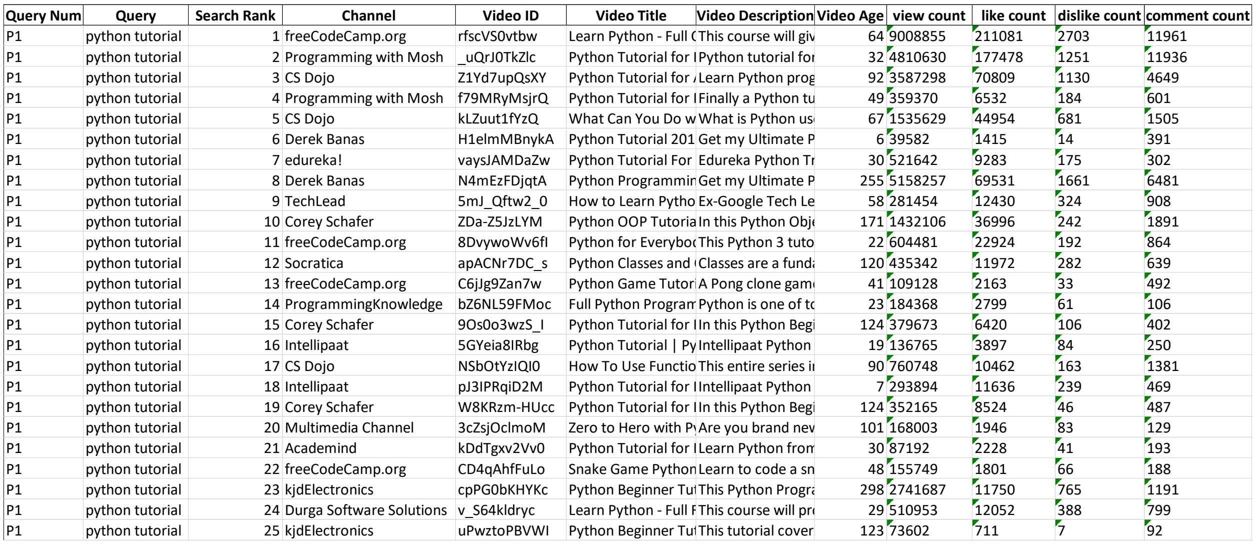
Video Statistics Result (key items bolded):

```
{'kind': 'youtube#videoListResponse',
 'etag': '"p4VTdlkQv3HQeTEaXgvLePAydmU/qiy6k0yl5Y3g3e2QvCZEX05VHlk"',
 'pageInfo': {'totalResults': 50, 'resultsPerPage': 50},
 'items': [{'kind': 'youtube#video',
   'etag': '"p4VTdlkQv3HQeTEaXgvLePAydmU/0UAojIcSxzPMfuQXkS_PxoeHKmc"',
   'id': 'q5uM4VKywbA',
   'statistics': {'viewCount': '327156',
    'likeCount': '4973',
    'dislikeCount': '51',
    'favoriteCount': '0',
    'commentCount': '364'}},
  {'kind': 'youtube#video',
   'etag': '"p4VTdlkQv3HQeTEaXgvLePAydmU/0--4nW9re9Qy8_Cjvdu-hVyrET8"',
   'id': 'Xi52tx6phRU',
   'statistics': {'viewCount': '259361',
    'likeCount': '4764',
    'dislikeCount': '126',
    'favoriteCount': '0',
    'commentCount': '301'}},
```

Video Statistics Data Wrangling: As shown above, the YT video statistics query returns a nested dictionary entry for each video. The information is exacted into a list and then converted into a pandas DataFrame. This information is then merged with the video statistics to create one DataFrame per query.

The query below (Python query 1 – “python tutorial”) is broad query. As shown by the 9M views for the first result, it’s by far the most viewed. After the top 5, the view count drops off steadily. Interestingly, as a user scroll down the results list, YT will keep showing the video results but after about 30 – 40 results, the results are not as relevant to the search criteria.

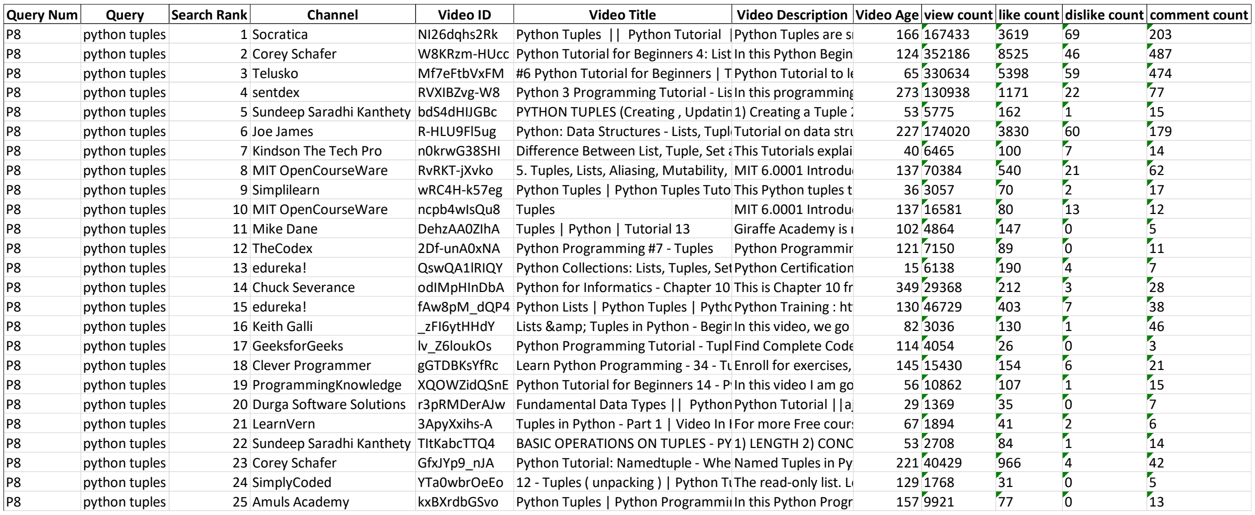
Python query 1 – “python tutorial” Top 25 results:



Query Num	Query	Search Rank	Channel	Video ID	Video Title	Video Description	Video Age	view count	like count	dislike count	comment count
P1	python tutorial	1	freeCodeCamp.org	rfscVS0vtbw	Learn Python - Full (	This course will giv	64	9008855	211081	2703	11961
P1	python tutorial	2	Programming with Mosh	_uQrJ0TkZlc	Python Tutorial for I	Python tutorial for	32	4810630	177478	1251	11936
P1	python tutorial	3	CS Dojo	Z1Yd7upQsXY	Python Tutorial for /	Learn Python prog	92	3587298	70809	1130	4649
P1	python tutorial	4	Programming with Mosh	f79MRyMsjrQ	Python Tutorial for I	Finally a Python tu	49	359370	6532	184	601
P1	python tutorial	5	CS Dojo	kLZuut1fYzQ	What Can You Do w	What is Python us	67	1535629	44954	681	1505
P1	python tutorial	6	Derek Banas	H1elmMBnykA	Python Tutorial 201	Get my Ultimate P	6	39582	1415	14	391
P1	python tutorial	7	edureka!	vaysJAMDaZw	Python Tutorial For	Edureka Python Tr	30	521642	9283	175	302
P1	python tutorial	8	Derek Banas	N4mEzFDjqtA	Python Programmin	Get my Ultimate P	255	5158257	69531	1661	6481
P1	python tutorial	9	TechLead	5mJ_Qftw2_0	How to Learn Pytho	Ex-Google Tech Le	58	281454	12430	324	908
P1	python tutorial	10	Corey Schafer	ZDa-Z5JzLYM	Python OOP Tutoria	In this Python Obj	171	1432106	36996	242	1891
P1	python tutorial	11	freeCodeCamp.org	8DvywoWv6fI	Python for Everyboc	This Python 3 tuto	22	604481	22924	192	864
P1	python tutorial	12	Socratica	apACNr7DC_s	Python Classes and (	Classes are a funda	120	435342	11972	282	639
P1	python tutorial	13	freeCodeCamp.org	C6jJg9Zan7w	Python Game Tutori	A Pong clone gam	41	109128	2163	33	492
P1	python tutorial	14	ProgrammingKnowledge	bZ6NL59FMoc	Full Python Progran	Python is one of tc	23	184368	2799	61	106
P1	python tutorial	15	Corey Schafer	9Os0o3wzS_I	Python Tutorial for I	In this Python Begi	124	379673	6420	106	402
P1	python tutorial	16	Intellipaat	5GYeia8IRbg	Python Tutorial | Py	Intellipaat Python	19	136765	3897	84	250
P1	python tutorial	17	CS Dojo	NSbOtYzIQI0	How To Use Functio	This entire series i	90	760748	10462	163	1381
P1	python tutorial	18	Intellipaat	pJ3IPRqiD2M	Python Tutorial for I	Intellipaat Python	7	293894	11636	239	469
P1	python tutorial	19	Corey Schafer	W8KRzm-HUcc	Python Tutorial for I	In this Python Begi	124	352165	8524	46	487
P1	python tutorial	20	Multimedia Channel	3cZsjOclmoM	Zero to Hero with Py	Are you brand nev	101	168003	1946	83	129
P1	python tutorial	21	Academind	kDdTgxv2Vv0	Python Tutorial for I	Learn Python from	30	87192	2228	41	193
P1	python tutorial	22	freeCodeCamp.org	CD4qAhfFuLo	Snake Game Python	Learn to code a sn	48	155749	1801	66	188
P1	python tutorial	23	kjdElectronics	cpPG0bKHYKc	Python Beginner Tut	This Python Progra	298	2741687	11750	765	1191
P1	python tutorial	24	Durga Software Solutions	v_S64kldryc	Learn Python - Full F	This course will pr	29	510953	12052	388	799
P1	python tutorial	25	kjdElectronics	uPwztoPBVWI	Python Beginner Tut	This tutorial cover	123	73602	711	7	92

The next query below (Python query 8 – “python tuples”) is a more specific python related query. The first result only have 167K views and after the seventh query, the number of view drops off steady. These are the type of focused queries where other ways to rank the video could move it up on the results list.

Python query 8 – “python tuples” Top 25 results:



Query Num	Query	Search Rank	Channel	Video ID	Video Title	Video Description	Video Age	view count	like count	dislike count	comment count
P8	python tuples	1	Socratica	NI26dqhs2Rk	Python Tuples || Python Tutorial |	Python Tuples are s	166	167433	3619	69	203
P8	python tuples	2	Corey Schafer	W8KRzm-HUcc	Python Tutorial for Beginners 4: List	In this Python Begin	124	352186	8525	46	487
P8	python tuples	3	Telusko	Mf7eFtbVxFM	#6 Python Tutorial for Beginners | T	Python Tutorial to le	65	330634	5398	59	474
P8	python tuples	4	sentdex	RVXIBZvg-W8	Python 3 Programming Tutorial - Lis	In this programming	273	130938	1171	22	77
P8	python tuples	5	Sundeep Saradhi Kanthety	bdS4dHIJGBc	PYTHON TUPLES (Creating , Updatin	1) Creating a Tuple 2	53	5775	162	1	15
P8	python tuples	6	Joe James	R-HLU9Fl5ug	Python: Data Structures - Lists, Tupl	Tutorial on data stru	227	174020	3830	60	179
P8	python tuples	7	Kindson The Tech Pro	n0krwG38SHI	Difference Between List, Tuple, Set a	This Tutorials explai	40	6465	100	7	14
P8	python tuples	8	MIT OpenCourseWare	RvRKT-jXvko	5. Tuples, Lists, Aliasing, Mutability,	MIT 6.0001 Introdu	137	70384	540	21	62
P8	python tuples	9	Simplilearn	wRC4H-k57eg	Python Tuples | Python Tuples Tuto	This Python tuples t	36	3057	70	2	17
P8	python tuples	10	MIT OpenCourseWare	ncpb4wIsQu8	Tuples	MIT 6.0001 Introdu	137	16581	80	13	12
P8	python tuples	11	Mike Dane	DehzAA0ZIhA	Tuples | Python | Tutorial 13	Giraffe Academy is r	102	4864	147	0	5
P8	python tuples	12	TheCodex	2Df-unA0xNA	Python Programming #7 - Tuples	Python Programmin	121	7150	89	0	11
P8	python tuples	13	edureka!	QswQA1lRIQY	Python Collections: Lists, Tuples, Set	Python Certification	15	6138	190	4	7
P8	python tuples	14	Chuck Severance	odIMpHInDbA	Python for Informatics - Chapter 10	This is Chapter 10 fr	349	29368	212	3	28
P8	python tuples	15	edureka!	fAw8pM_dQP4	Python Lists | Python Tuples | Pytho	Python Training : ht	130	46729	403	7	38
P8	python tuples	16	Keith Galli	_zFI6ytHHdY	Lists &amp; Tuples in Python - Begin	In this video, we go	82	3036	130	1	46
P8	python tuples	17	GeeksforGeeks	lv_Z6loukOs	Python Programming Tutorial - Tupl	Find Complete Code	114	4054	26	0	3
P8	python tuples	18	Clever Programmer	gGTDBKsYfRc	Learn Python Programming - 34 - Tu	Enroll for exercises,	145	15430	154	6	21
P8	python tuples	19	ProgrammingKnowledge	XQOWZidQSnE	Python Tutorial for Beginners 14 - P	In this video I am go	56	10862	107	1	15
P8	python tuples	20	Durga Software Solutions	r3pRMDerAJw	Fundamental Data Types || Python	Python Tutorial ||aj	29	1369	35	0	7
P8	python tuples	21	LearnVern	3ApyXxihs-A	Tuples in Python - Part 1 | Video In I	For more Free cours	67	1894	41	2	6
P8	python tuples	22	Sundeep Saradhi Kanthety	TItKabcTTQ4	BASIC OPERATIONS ON TUPLES - PY	1) LENGTH 2) CONC	53	2708	84	1	14
P8	python tuples	23	Corey Schafer	GfxJYp9_nJA	Python Tutorial: Namedtuple - Whe	Named Tuples in Py	221	40429	966	4	42
P8	python tuples	24	SimplyCoded	YTa0wbrOeEo	12 - Tuples (unpacking) | Python Tu	The read-only list. L	129	1768	31	0	5
P8	python tuples	25	Amuls Academy	kxBXrdbGSvo	Python Tuples | Python Programmi	In this Python Progr	157	9921	77	0	13

Extracting the dictionary data into dataframes took some effort to optimize the code and the queries. One interesting challenge is the fact the video owners can turn off comments on their videos. When this happens, there are no statistics for likes, dislikes, or comments. YT just skips that information in the dictionary response, and I had to check for that condition and put zeros in the impacted fields.

Video Comments: The third query is the video comments query. This exacts the first 100 comments (selecting by relevance) for each video, populates lists to hold the comments, and then creates a dataframe. The dataframe is exported to excel.

Video Comments Search Query:

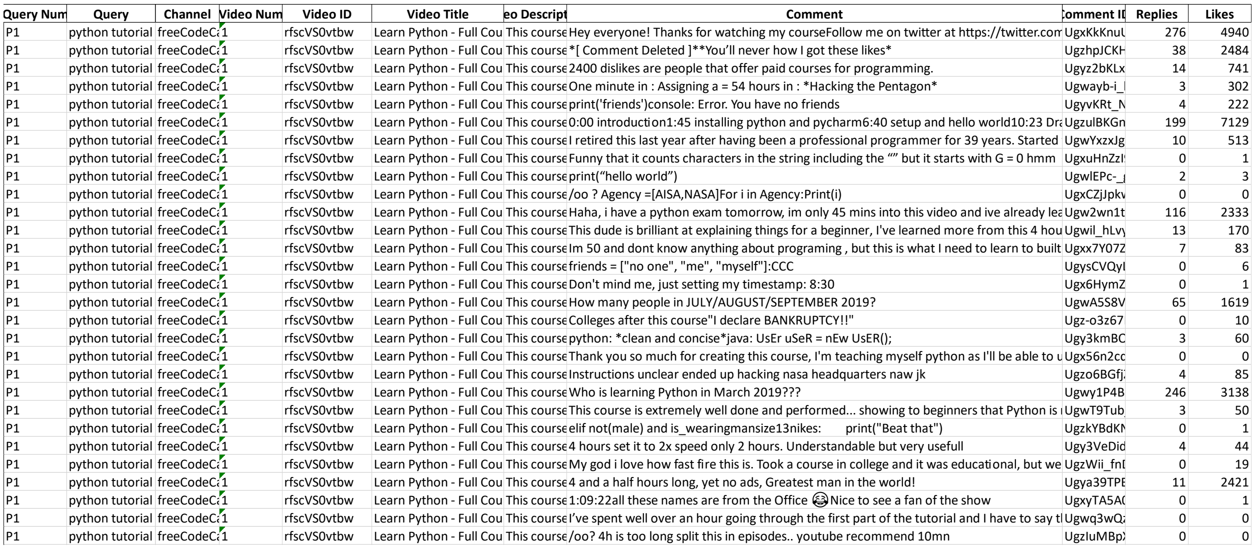
```
from tqdm import tqdm
for i, video in enumerate(tqdm(video_id, ncols = 100)):
    response = youtube.commentThreads().list(
                    part = 'snippet',
                    videoId = video,
                    maxResults = 100, # Only take top 100 comments...
                    order = 'relevance', #... ranked on relevance
                    textFormat = 'plainText',
                    ).execute()
```

Video Comments Search Query Results (key items bolded:

```
{'kind': 'youtube#commentThread',
   'etag': '"p4VTdlkQv3HQeTEaXgvLePAydmU/vWpuLBC_C_QJf9bmvK4Vq9nNUJo"',
   'id': 'UgzDzCdMIO235N77cK54AaABAg',
   'snippet': {'videoId': 'N0lxfilGfak',
    'topLevelComment': {'kind': 'youtube#comment',
     'etag': '"p4VTdlkQv3HQeTEaXgvLePAydmU/12KavxCw51GxXPaWHn7IEtIrfkk"',
     'id': 'UgzDzCdMIO235N77cK54AaABAg',
     'snippet': {'authorDisplayName': 'Triple Jay',
      'authorProfileImageUrl': 'https://yt3.ggpht.com/-CpH8J3FO_1Y/AAAAAAAAAAI/AAAAAAAAAAA/t7ljbaa_tZc/s28-c-k-no-mo-rj-c0xffffff/photo.jpg',
      'authorChannelUrl': 'http://www.youtube.com/channel/UC5Smaf9QTcCYaUfR9pxf98Q',
      'authorChannelId': {'value': 'UC5Smaf9QTcCYaUfR9pxf98Q'},
      'videoId': 'N0lxfilGfak',
      'textDisplay': "I'm new to python, but with this tutorial, i have by-passed my fears.. Thank you very much.",
      'textOriginal': "I'm new to python, but with this tutorial, i have by-passed my fears.. Thank you very much.",
      'canRate': True,
      'viewerRating': 'none',
      'likeCount': 1,
      'publishedAt': '2017-12-14T08:36:02.000Z',
      'updatedAt': '2017-12-14T08:36:02.000Z'}},
    'canReply': True,
    'totalReplyCount': 1,
    'isPublic': True}},
```

As shown in the query statistics above, each video can have as many as 11K comments (line 1 of the Python query 1) or as few as 3 comments (line 18 of the Python query 8). For this study, I am just extracting up to the first 100 ‘relevant” comments. If I took the most recent comments by date posted, that might provide a different view of the comments. However, given the relatively low numbers of dislikes across the video sample Isee video statistics on the next page), I observed most comments are either mostly positive or a question / comment (i.e. – “can you do a video on X topic…..”, “can you explain your comment about Y at Z:ZZ in the video). Most people won’t write too negative of a comment, especially for python coding topics.

Python query 1 – “python tutorial” Comments extract:



Query Num	Query	Channel	Video Num	Video ID	Video Title	eo Descript	Comment	Comment ID	Replies	Likes
P1	python tutorial	freeCodeCa	1	rfscVS0vtbw	Learn Python - Full Cou	This course	Hey everyone! Thanks for watching my courseFollow me on twitter at https://twitter.com	UgxKkKnuU	276	4940
P1	python tutorial	freeCodeCa	1	rfscVS0vtbw	Learn Python - Full Cou	This course	*[Comment Deleted]**You’ll never how I got these likes*	UgzhpJCKH	38	2484
P1	python tutorial	freeCodeCa	1	rfscVS0vtbw	Learn Python - Full Cou	This course	2400 dislikes are people that offer paid courses for programming.	Ugyz2bKLx	14	741
P1	python tutorial	freeCodeCa	1	rfscVS0vtbw	Learn Python - Full Cou	This course	One minute in : Assigning a = 54 hours in : *Hacking the Pentagon*	Ugwayb-i_	3	302
P1	python tutorial	freeCodeCa	1	rfscVS0vtbw	Learn Python - Full Cou	This course	print('friends')console: Error. You have no friends	UgyvKRt_N	4	222
P1	python tutorial	freeCodeCa	1	rfscVS0vtbw	Learn Python - Full Cou	This course	0:00 introduction1:45 installing python and pycharm6:40 setup and hello world10:23 Dra	UgzulBKGn	199	7129
P1	python tutorial	freeCodeCa	1	rfscVS0vtbw	Learn Python - Full Cou	This course	I retired this last year after having been a professional programmer for 39 years. Started	UgwYxzxJg	10	513
P1	python tutorial	freeCodeCa	1	rfscVS0vtbw	Learn Python - Full Cou	This course	Funny that it counts characters in the string including the “” but it starts with G = 0 hmm	UgxuHnZzI	0	1
P1	python tutorial	freeCodeCa	1	rfscVS0vtbw	Learn Python - Full Cou	This course	print(“hello world”)	UgwlEPc-_	2	3
P1	python tutorial	freeCodeCa	1	rfscVS0vtbw	Learn Python - Full Cou	This course	/oo ? Agency =[AISA,NASA]For i in Agency:Print(i)	UgxCZjJpkv	0	0
P1	python tutorial	freeCodeCa	1	rfscVS0vtbw	Learn Python - Full Cou	This course	Haha, i have a python exam tomorrow, im only 45 mins into this video and ive already lea	Ugw2wn1t	116	2333
P1	python tutorial	freeCodeCa	1	rfscVS0vtbw	Learn Python - Full Cou	This course	This dude is brilliant at explaining things for a beginner, I've learned more from this 4 hou	Ugwil_hLvy	13	170
P1	python tutorial	freeCodeCa	1	rfscVS0vtbw	Learn Python - Full Cou	This course	Im 50 and dont know anything about programing , but this is what I need to learn to built	Ugxx7Y07Z	7	83
P1	python tutorial	freeCodeCa	1	rfscVS0vtbw	Learn Python - Full Cou	This course	friends = ["no one", "me", "myself"]:CCC	UgysCVQyl	0	6
P1	python tutorial	freeCodeCa	1	rfscVS0vtbw	Learn Python - Full Cou	This course	Don't mind me, just setting my timestamp: 8:30	Ugx6HymZ	0	1
P1	python tutorial	freeCodeCa	1	rfscVS0vtbw	Learn Python - Full Cou	This course	How many people in JULY/AUGUST/SEPTEMBER 2019?	UgwA5S8V	65	1619
P1	python tutorial	freeCodeCa	1	rfscVS0vtbw	Learn Python - Full Cou	This course	Colleges after this course"I declare BANKRUPTCY!!"	Ugz-o3z67	0	10
P1	python tutorial	freeCodeCa	1	rfscVS0vtbw	Learn Python - Full Cou	This course	python: *clean and concise*java: UsEr uSeR = nEw UsER();	Ugy3kmBC	3	60
P1	python tutorial	freeCodeCa	1	rfscVS0vtbw	Learn Python - Full Cou	This course	Thank you so much for creating this course, I'm teaching myself python as I'll be able to u	Ugx56n2cc	0	0
P1	python tutorial	freeCodeCa	1	rfscVS0vtbw	Learn Python - Full Cou	This course	Instructions unclear ended up hacking nasa headquarters naw jk	Ugzo6BGfj	4	85
P1	python tutorial	freeCodeCa	1	rfscVS0vtbw	Learn Python - Full Cou	This course	Who is learning Python in March 2019???	Ugwy1P4B	246	3138
P1	python tutorial	freeCodeCa	1	rfscVS0vtbw	Learn Python - Full Cou	This course	This course is extremely well done and performed... showing to beginners that Python is	UgwT9Tub	3	50
P1	python tutorial	freeCodeCa	1	rfscVS0vtbw	Learn Python - Full Cou	This course	elif not(male) and is_wearingmansize13nikes: print("Beat that")	UgzkYBdKN	0	1
P1	python tutorial	freeCodeCa	1	rfscVS0vtbw	Learn Python - Full Cou	This course	4 hours set it to 2x speed only 2 hours. Understandable but very usefull	Ugy3VeDid	4	44
P1	python tutorial	freeCodeCa	1	rfscVS0vtbw	Learn Python - Full Cou	This course	My god i love how fast fire this is. Took a course in college and it was educational, but we	UgzWii_fnI	0	19
P1	python tutorial	freeCodeCa	1	rfscVS0vtbw	Learn Python - Full Cou	This course	4 and a half hours long, yet no ads, Greatest man in the world!	Ugya39TPE	11	2421
P1	python tutorial	freeCodeCa	1	rfscVS0vtbw	Learn Python - Full Cou	This course	1:09:22all these names are from the Office 😂Nice to see a fan of the show	UgxyTA5A	0	1
P1	python tutorial	freeCodeCa	1	rfscVS0vtbw	Learn Python - Full Cou	This course	I’ve spent well over an hour going through the first part of the tutorial and I have to say t	Ugwq3wQ	0	0
P1	python tutorial	freeCodeCa	1	rfscVS0vtbw	Learn Python - Full Cou	This course	/oo? 4h is too long split this in episodes.. youtube recommend 10mn	UgzIuMBp	0	0

The comments extract above highlights the challenges with reviewing the comments. A majority will be generally positive and the others typically fall into 2 categories (a question or a comment). The key is to focus on the positive comments (ie. – “great video”, “very clear instruction”, etc). It’s easy for a YT viewer to just click the Like button but taking the time to type in a positive comment shows an extra level of praise for the video content.

Data EDA & Analysis:

Python Query Summary Video Statistics: The Python queries video statistics table below shows the total number of views (46.8M), likes (947K) and dislikes (14K) for the 8 python related queries. In aggregate, the videos only averaged getting 1.6% of likes compared to the total views. Since the total number of dislikes is also very small, the ratio of dislikes to likes is just 1.8% (meaning for 1,000 likes, on average there is only 18 dislikes for the top 25 videos).

The comments numbers are even smaller. Across the whole population, there was only 65K comments vs. 46.8M views (0.1%). If people are going to take the time to write a comment, they should have extra weighting in the search results. I extracted the top 100 “relevant” comments from each of the top 25 videos from the eight python queries for a total of 8,492 (13%) comments across 200 videos (65K total comments). For the top viewed videos, it’s a pretty small subset of the comments but for the top 10 – 25 videos for each query, often, they had less than 100 total comments for the less viewed videos.

Python Query Summary Video Statistics:

Python Query	Total Views for top 25 results	Total Likes for top 25 results	Total Dislikes for top 25 results	Total Comments	Total Extracted Comments	Hand Tagged Comments
1) Python tutorial	33,728,573	751,794 (2.2%)	10,921 (1.5%)	48,307	2,366 (5%)	595
2) Python reading CSV files	2,231,893	27,462 (1.2%)	616 (2.2%)	2,263	808 (36%)	264
3) Python pandas DataFrames	1,214,931	18,761 (1.5%)	378 (2.0%)	1,912	721 (38%)	111
4) Python lists	2,956,419	50,627 (1.7%)	616 (1.2%)	3,869	1,282 (33%)	50
5) Python dictionaries	1,650,572	28,577 (1.7%)	472 (1.7%)	2,170	1,016 (47%)	25
6) Python sort functions	419,081	7,567 (1.8%)	185 (2.4%)	678	432 (64%)	25
7) Python for loops	3,195,745	36,107 (1.1%)	836 (2.3%)	4,358	1,184 (27%)	25
8) Python tuples	1,443,193	26,187 (1.8%)	329 (1.3%)	1,801	683 (38%)	25
Totals:	**46,840,407**	**947,082 (1.6%)**	**14,353 (1.8%)**	**65,358 (0.1%)**	**8,492 (13%)**	**1,120 (13%)**

I hand tagged 1,120 comments (1 = positive, 0 = other) to develop a train-test data set to classify the other videos comments. I will classify the remaining 7,372 comments to see if these classifications will help determine if the percentage of positive comments might be used to alter the YT video search rankings.

Python Query Analysis: The video statistics data from each of the 8 separate python queries is graphed in Appendix A. The following metrics were gathered for each of the top 25 videos: age (calculated in weeks), total views, total likes (i.e. – a user clicks on the thumbs up icon), total dislikes (i.e. – a user clicks on the thumbs down icon), and total comments (a user posts a comment and his YT user id is also posted).

The chart for Python Query 2 – “python reading CSV files” is indicative of the 8 charts in Appendix A. The key observations are:

Age: The top 5- 10 videos are typically 1 – 2 years old. The top 11 – 25 videos age can vary widely (brand new to several years old.

Like Count: As shown in the video statistics summary table above, each video receives on average 1.6% likes per view ratio (for 1,000 views, it will get 16 likes). The 5-8 videos get the most likes and it then steadily drops from there. When a user search YT, the first few videos show up in the search results. Once a user selects 1 video, another 5 – 6 show up in a sidebar view. Most users probably just view the videos where they see the video listed on the main landing page and they do not seem to search manually through the results list.

The significance of this result is that unless a video makes the top 5 – 8 results, it will be difficult for it to make it to the top of the search results.

Dislike Count: While the total like count is approximately 1.6% of the total view count, the total dislike count is only 1.8% of the like count. This number is so small, it's off limited value.

Comments Count: Users comment on only 0.1% of all views (1 comment per 1,000 views). After personally reviewing and tagging 1,120 python query comments, there is a definitely pattern to the comments. A majority (estimate 75%) are positive comments and another 20% are questions / comments. Certain videos that provide unclear python coding demonstrations which might be confusing or have errors generate a large number of questions in the comments section.

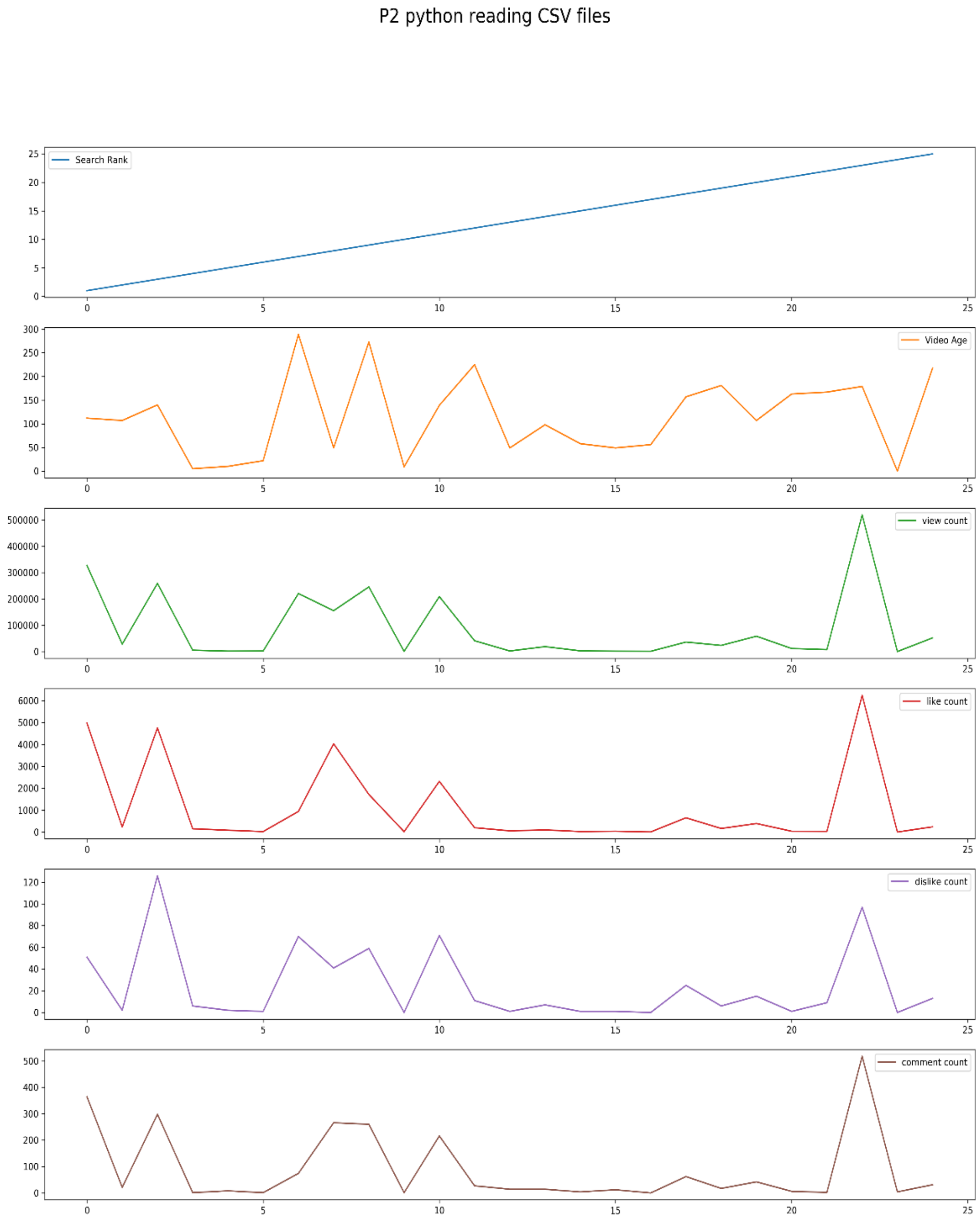
The key here is seeing which comments have a vast majority of positive comments. Certain videos channels (Corey Schafer, Socratica, etc.) get very positive reviews for their valuable content.

Overall, the charting of these metrics across the 4 key data points (views, likes, dislikes, comments) all follow a similar pattern – the views set the pattern and the likes, dislikes, and comments data is in proportion to the views.

The Python Query 2 (reading CSV files) does show an interesting data spike in video 23. The video is listed in the 23 query result yet it has the highest views (200K more than the number 1 video) as well as more likes, more comments and less dislikes. Upon reviewing the data, the video in question is from the python exper Corey Schafer. I have personally viewed many of his videos and they are always excellent.

The highlights a key issue around video title and video description relevance in the YT search query. Interestingly Corey Schafer is the author of the number 1 ranked video for reading CSV files. His video title and video description both mention "reading CSV files". His 23rd ranked video is called "reading files in python". It does not mention CSV files in either the video title or description. This shows that the video title and description relevance is key to video ranking even when a similar video has literally 200K more views.

Python Query 2: “Python reading CSV files”



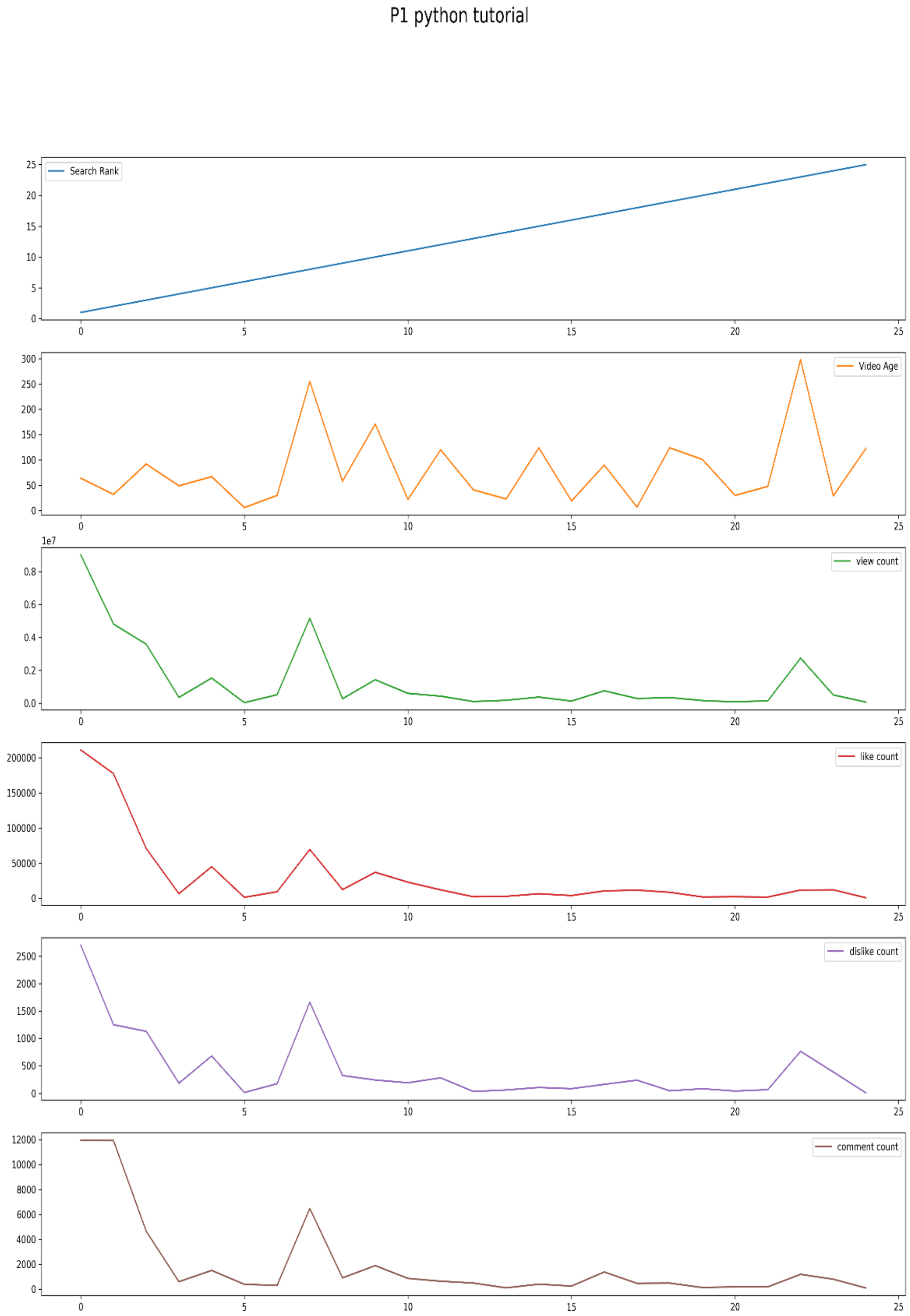
P2 python reading CSV files

Machine Learning Application

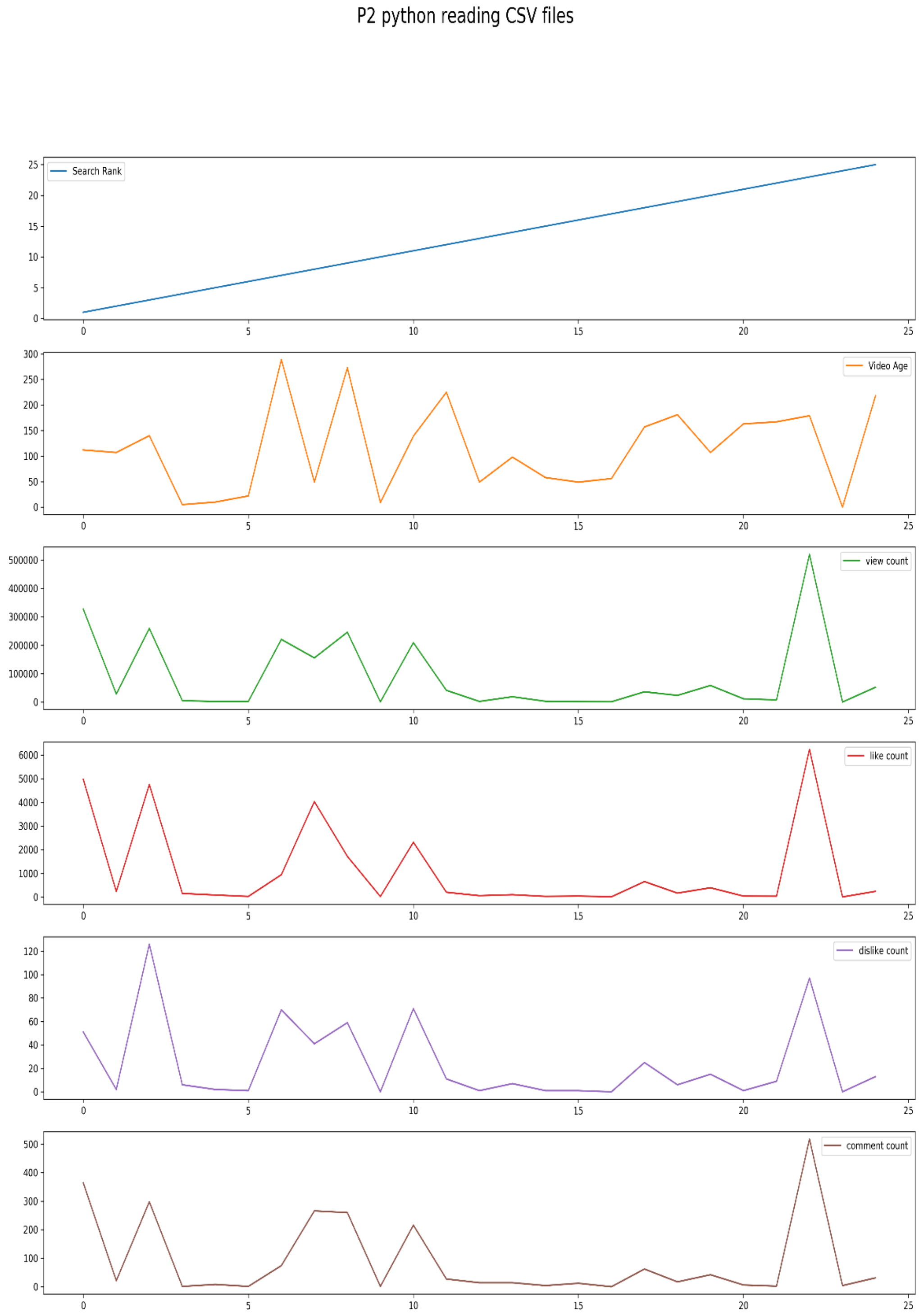
To be completed for Capstone 2 Milestone B

Appendix A – Python Query Statistics EDA

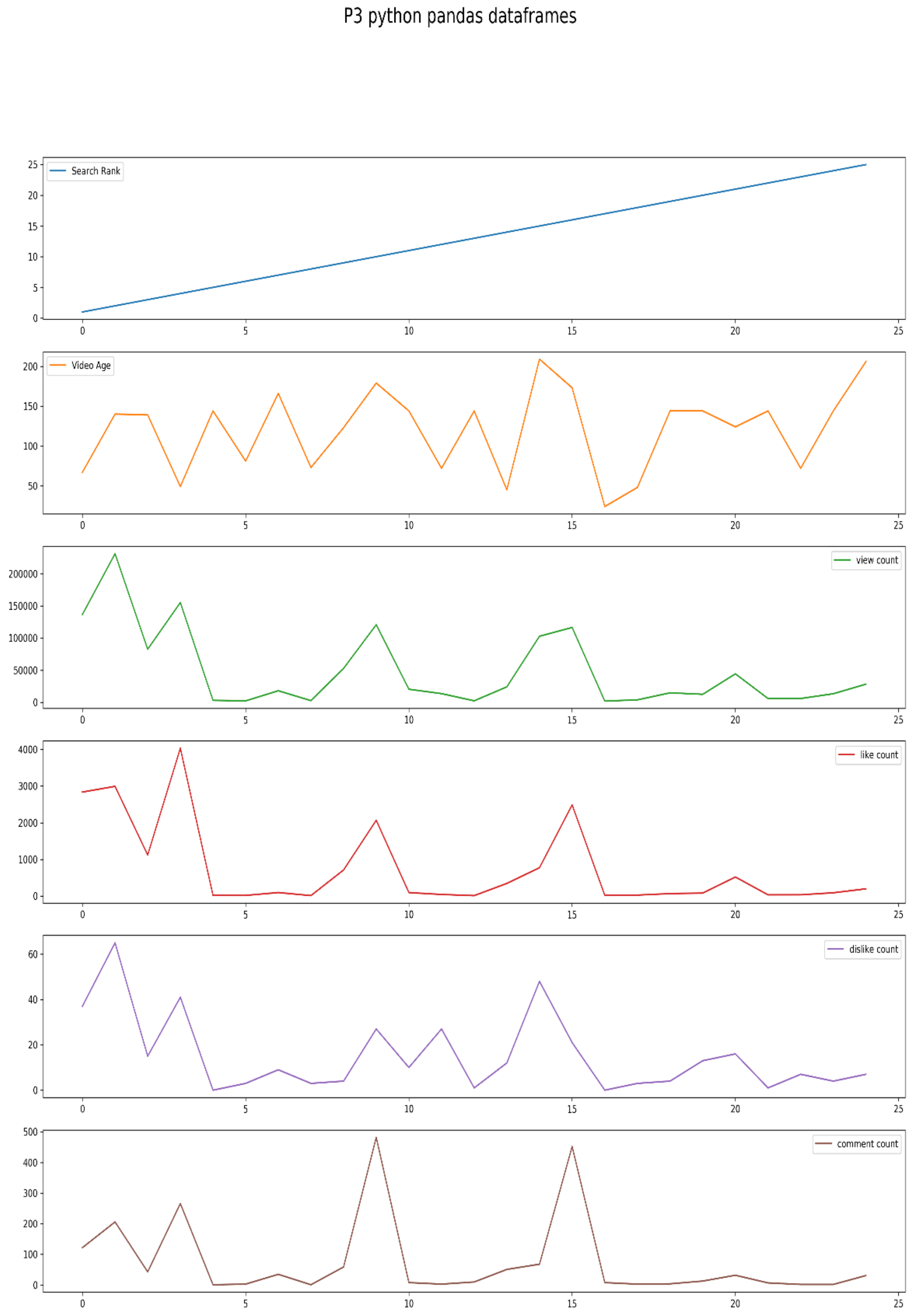
Python Query 1: “Python tutorial”



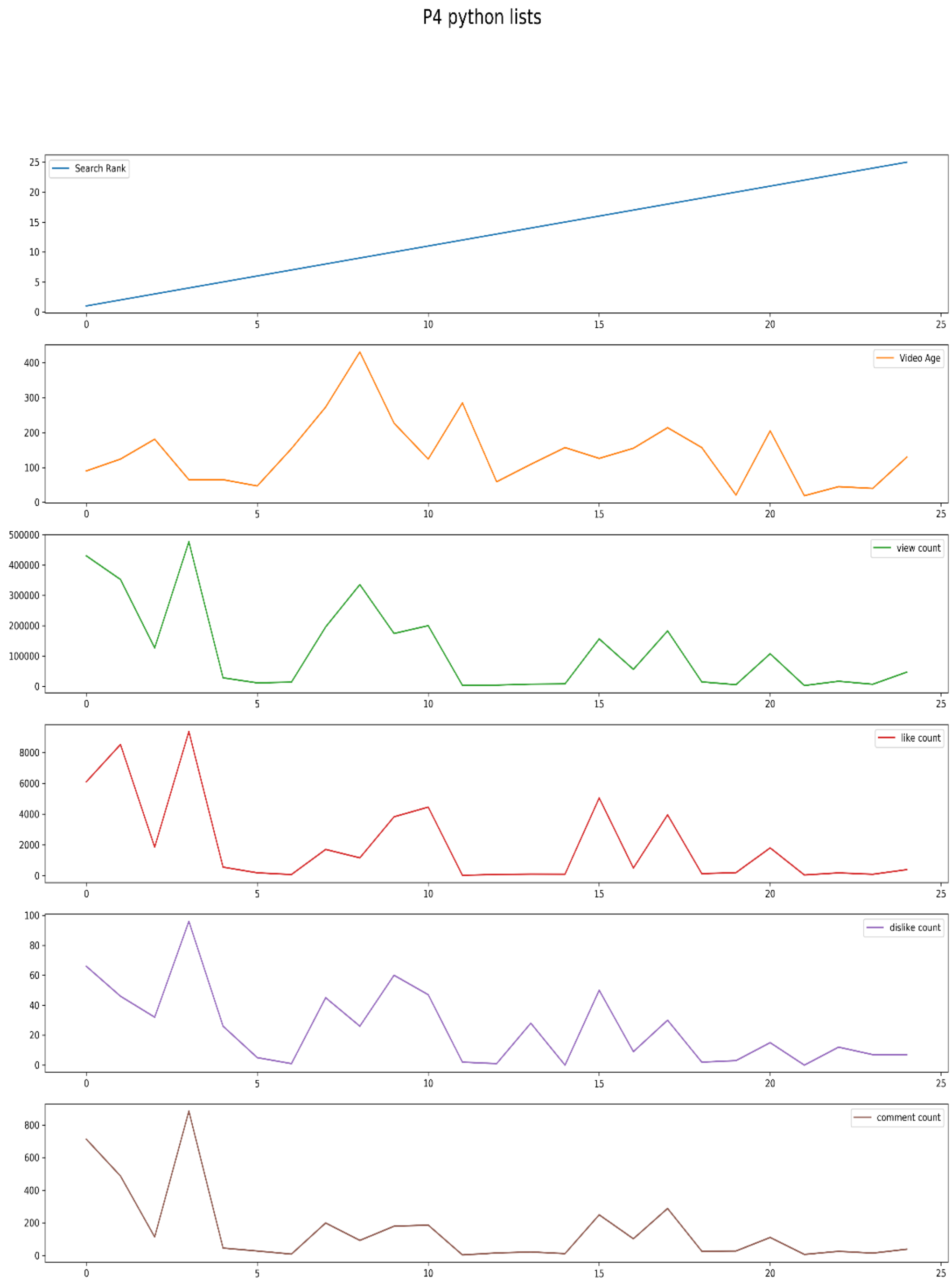
Python Query 2: “Python reading CSV files”



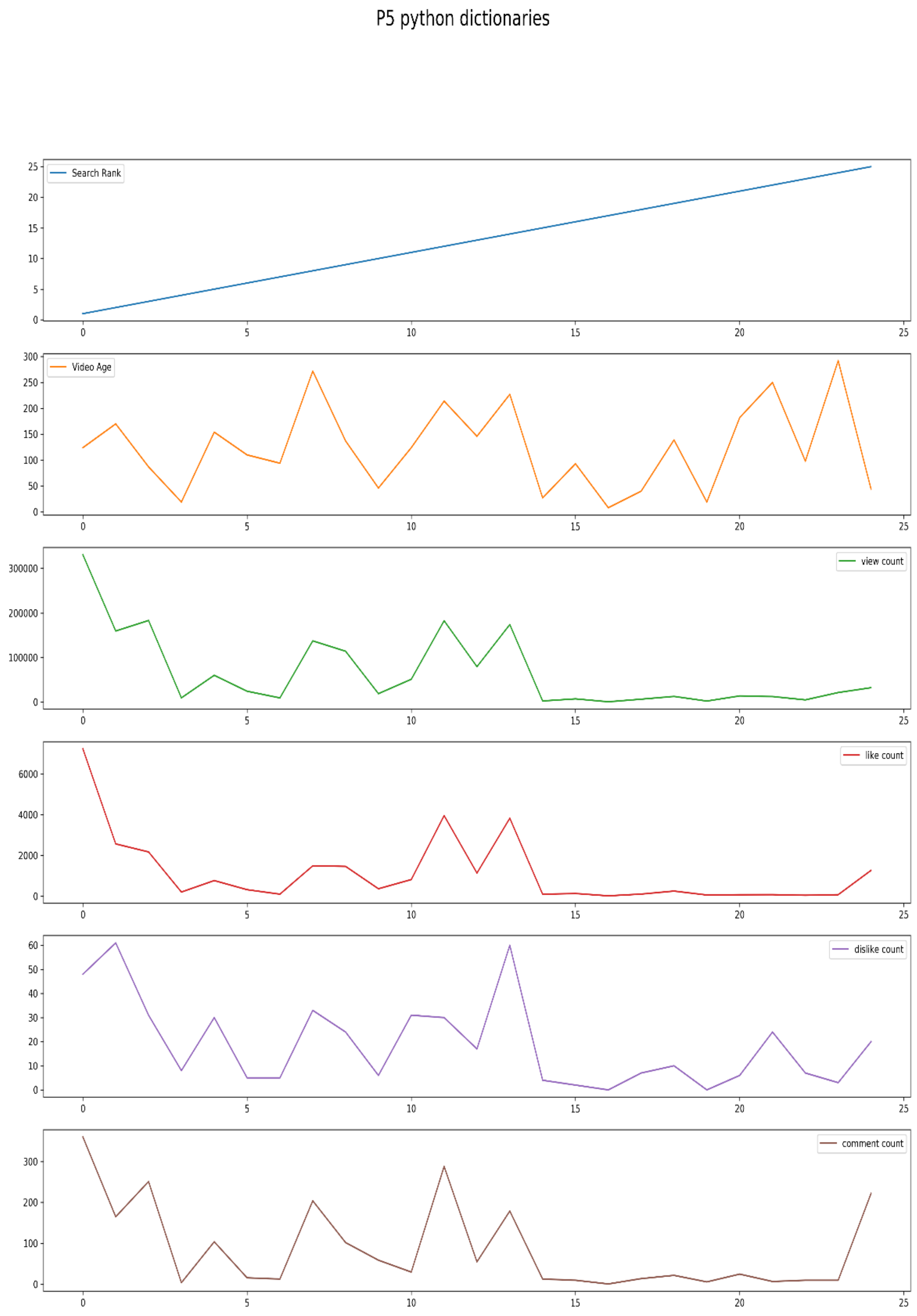
Python Query 3: “Python pandas dataframes”



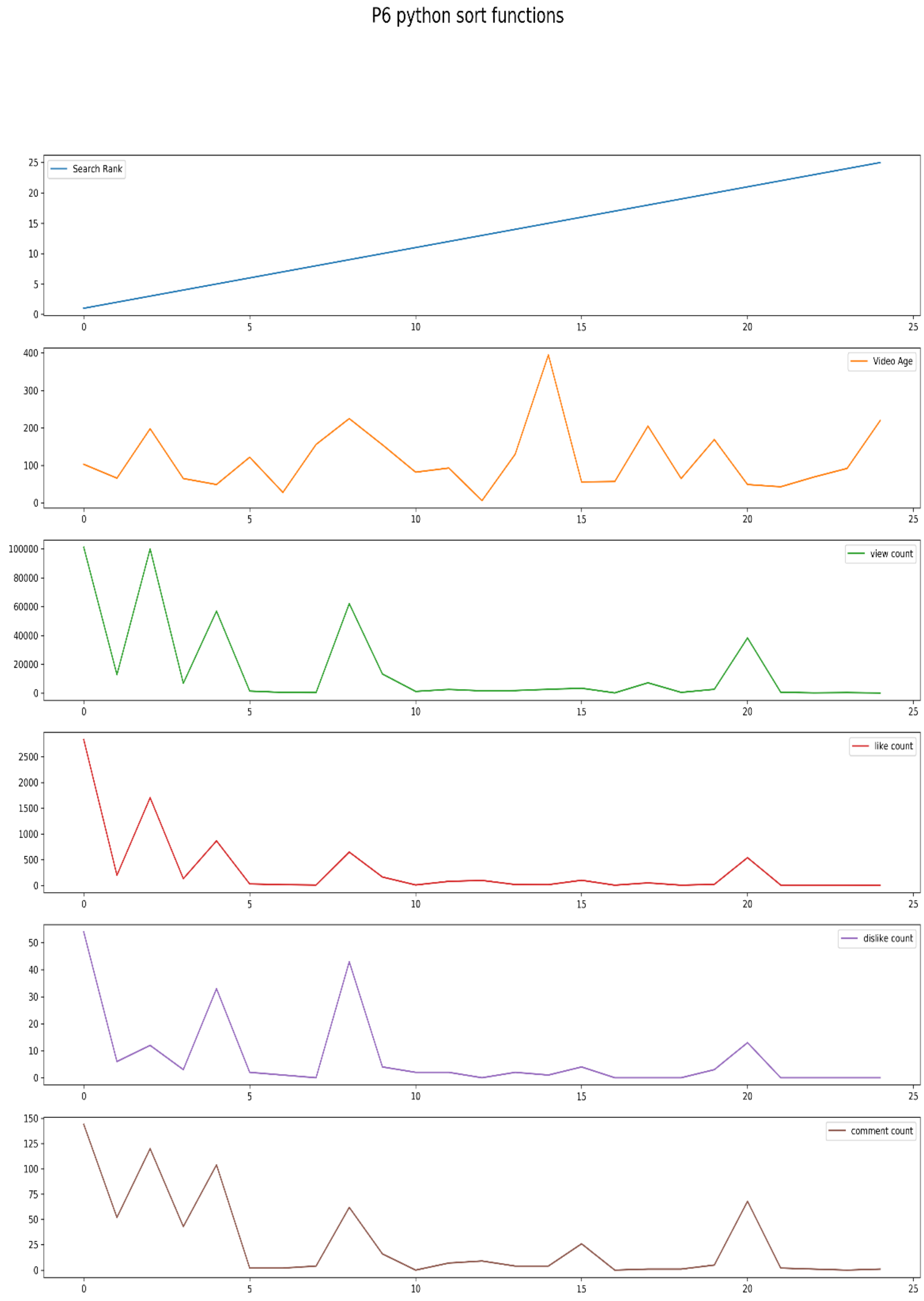
Python Query 4: “Python lists”



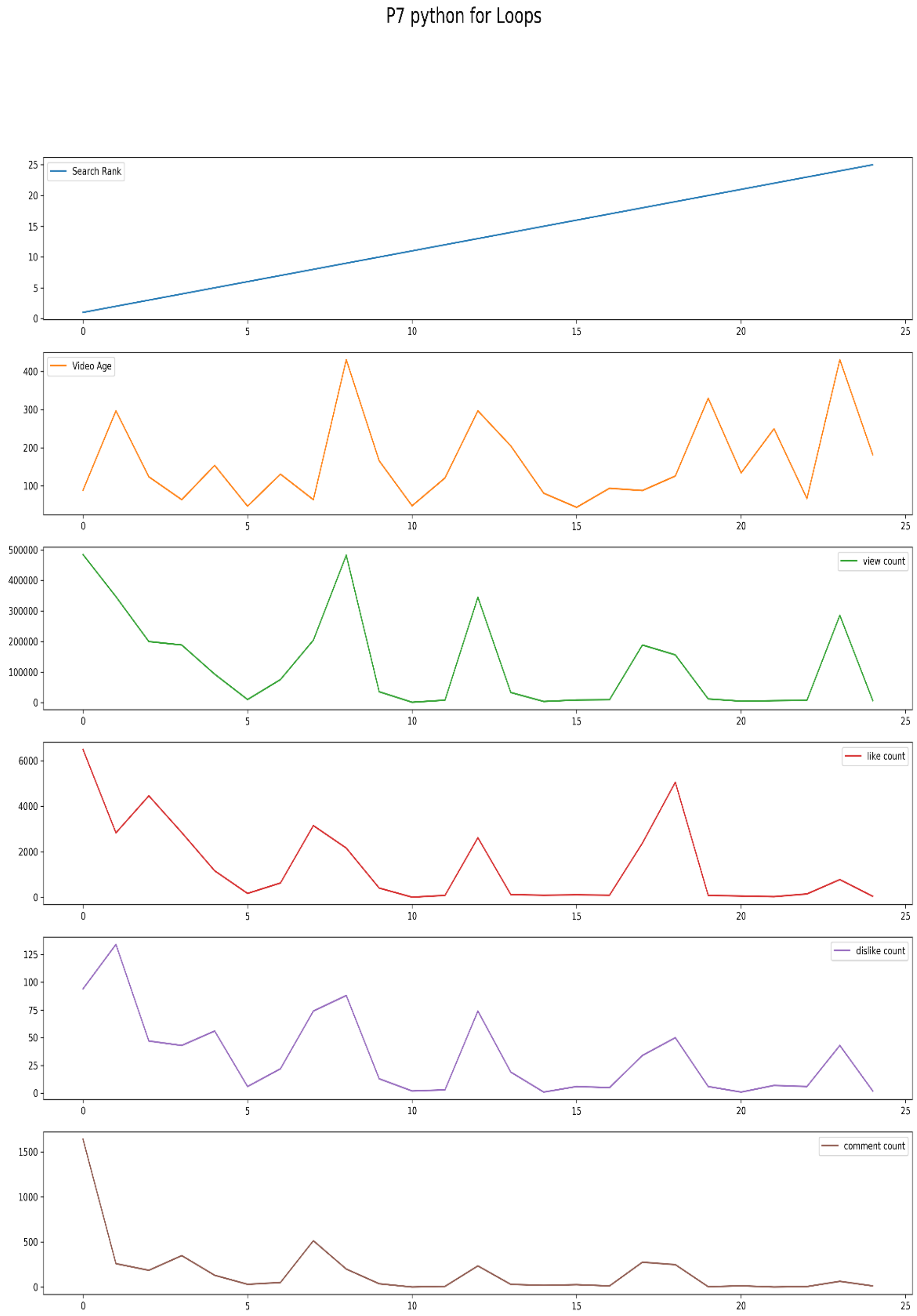
Python Query 5: “Python dictionaries”



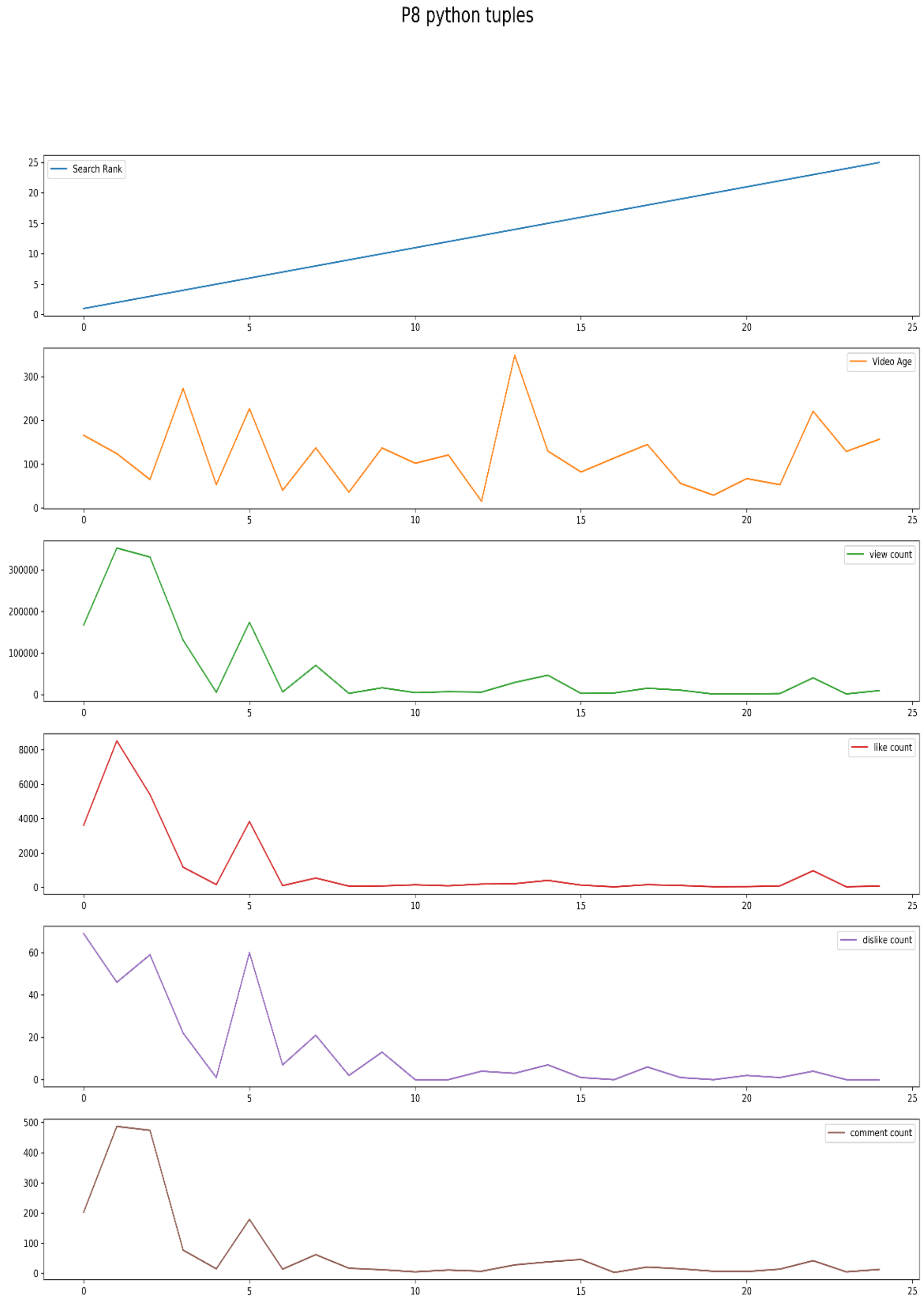
Python Query 6: “Python sort functions”



Python Query 7: “Python for loops”



Python Query 8: “Python tuples”


P8 python tuples
Search Rank
Video Age
view count
like count
dislike count
comment count