I decide to use a new dataset different from the one used in hw3.

About the Data

1. Name / Title: Trending videos on Youtube
2. Link to Data: https://www.kaggle.com/datasets/anushabellam/trending-videos-on-youtube
3. Source / Origin:
   - Author or Creator: Mendeley.com
   - Publication Date: 2022
   - Publisher: anusha bellam
   - Version or Data Accessed: 1
4. License: CC0: Public Domain
5. Can You Use this Data Set for Your Intended Use Case? Yes

Format and Samples

Overview

Format: csv Size: 73 KB Number of Records: 115

Sample of Data

TODO show a few lines of data from the actual file. ⚠ Use "regular" Python to do this in this code block. Assuming that jupyter-lab was started in your root directory: with open('../data/raw/example-data.csv', 'r')

In [ ]:

```
import pandas as pd
import sys
file_path = sys.path[0] +'/../data/raw/Trending videos on youtube dataset.csv'

with open(file_path, 'r') as f:
    columns = f.readline().split(',')
    lines = f.readlines()
    first3lines = [tuple(line.split(',')) for line in lines][:3]

print('see first three lines of the dataset:')
for line in first3lines:
    print(line, '\n')
```

```
see first three lines of the dataset:
('0', 'UCU1_l0ZJyTK_7HZZ3Ruw8Dg', 'MAPS', 'pTnk3ziVVRM', '2014-01-10T01:24:57.000Z', 'Psychedelic Horizons Beyond
Psychotherapy Workshop - Part 3/4', '"Watch the full workshop at http://psychedelicscience.org In addition to thei
r psychotherapeutic uses', ' experimental evidence shows that psychedelics deepen daily life and enrich most acade
mic fields. As popular use still is 99% of all experiences', ' it is important to be aware of the effects of risin
g religious use', ' microdosing', ' problem--solving', ' self-help groups', ' effects on education and popular cul
ture. Entheogenic uses promote spiritual growth including heightened senses of significance', ' meaningfulness', '
and sacredness. Personality growth promotes altruism and open mindedness. Future leads propose advances through mi
crodosing', ' developing a Multistate Theory —even exploring previously unknown mindbody states that may contain u
nusual', ' rare and even undiscovered human abilities. James Fadiman', ' PhD completed his dissertation at Stanfor
d on the effectiveness of LSD-assisted therapy just as all research was shut down. During the subsequent 40 year l
ull', ' he has held a variety of teaching', ' consulting', ' training', ' counseling', ' and editorial positions.
He has taught in psychology departments', ' design engineering', ' and for the past three decades at the Institute
of Transpersonal Psychology that he co-founded. He has published textbooks', ' professional books', ' a self-help
book', ' a novel', ' and a series of videos (""Drugs:The Children are Choosing"") for National Public Television.
His books have been published in eight languages. He was featured in a National Geographic documentary and had thr
ee solo shows of his nature photography. He sits on two non-profit boards and has been the president of several sm
all natural resource companies. He has been involved in researching psychedelics for spiritual', ' therapeutic', "
and creative uses when it was legal and subsequently and recently published The Psychedelic Explorer's Guide: Saf
e", ' Therapeutic', ' and Sacred Journeys and released a series of videos (with Kokyo Henkel)', " Buddhism and Psy
chedelics. He is doing surveys of psychedelic use and has pioneered researching micro-dosing of a number of substa
nces for a host of conditions. More at http://jamesfadiman.com Dr. Roberts's major publication in the humanities a
nd social sciences is The Psychedelic Future of the Mind: Enhancing Cognition", ' Boosting Intelligence', ' and Ra
ising Values (2013)', ' in religion Spiritual Growth with Entheogens: Psychoactive Sacramentals and Human Transfor
mation (2012)', ' in medicine Psychedelic Medicine: New Evidence for Hallucinogenic Substances as Treatments (200
7)', ' and in mind development and education Psychedelic Horizons: Snow White', ' Immune System', ' Multistate Min
d', ' Enlarging Education (2006). He is a founding member of MAPS', ' co-founder of the Council on Spiritual Pract
ices', ' creator of Rising Researcher sessions', ' and originated the celebration of Bicycle Day. He has taught Ps
ychedelic Studies at Northern Illinois University since 1981. More info at: niu.academia.edu/ThomasRoberts"', '2
9', 'Nonprofits & Activism', 'PT1H19M40S', '4780', 'hd', 'False', '1512', '8.0', '0.0', '1.0\n')

('1', 'UCLuO2lUqHrPIIpx0hFenV2g', 'Tink Tink Club', 'cuJjSeHZIrg', '2015-06-18T16:56:04.000Z', 'Episode 35 - Dr. J
ames Fadiman', 'Dr. James Fadiman is the father of modern psychedelic therapy and the author of the Psychedelic Ex
plorers Guide. He is currently doing research on the effects of microdosing.', '22', 'People & Blogs', 'PT1H12M34
S', '4354', 'sd', 'False', '881', '14.0', '0.0', '3.0\n')

('2', 'UCihqrkaOgVMfLNo2W1hSliA', 'Podcast Bunk', 'IuyuZfWtGgg', '2016-05-01T05:33:13.000Z', '#325 Microdosing fro
m The Adam and Dr Drew Show on podbay', '"Adam and Dr. Drew are solo today and they open the show talking to a cal
ler about the new phenomenon of \'microdosing\' LSD and Drew gives his thoughts on how it will unfold. As the show
continues', ' they talk to other callers including one who has been addicted to pain killers for a long time tryin
g to ween himself off', ' as well as someone who wants to figure out how to help children who didn\'t grow up with
the best family support structure. AdamAndDrDrewShow.com"', '22', 'People & Blogs', 'PT50M40S', '3040', 'sd', 'Fal
se', '67', '0.0', '1.0', '3.0\n')
```

Fields or Column Headers

- column2', 'channelid'
- column3', 'channeltitle'
- column4', 'videoid'
- column5', 'publishedat'
- column6', 'videotitle'
- column7', 'videodescription'
- column8', 'videocategoryid'
- column9', 'videocategorylabel'
- column10', 'duration'
- column11', 'durationsec'
- column12', 'definition'
- column13', 'caption'
- column14', 'viewcount'
- column15', 'likecount'
- column16', 'dislikecount'
- column17', 'commentcount'

```
In [ ]: [(f'column{i+1}',columns[i].lower()) for i in range(len(columns))]
```

```
Out[ ]: [('column1', ''),
         ('column2', 'channelid'),
         ('column3', 'channeltitle'),
         ('column4', 'videoid'),
         ('column5', 'publishedat'),
         ('column6', 'videotitle'),
         ('column7', 'videodescription'),
         ('column8', 'videocategoryid'),
         ('column9', 'videocategorylabel'),
         ('column10', 'duration'),
         ('column11', 'durationsec'),
         ('column12', 'definition'),
         ('column13', 'caption'),
         ('column14', 'viewcount'),
         ('column15', 'likecount'),
         ('column16', 'dislikecount'),
         ('column17', 'commentcount\n')]
```

1. Retrieve the data, create a DataFrame

- Include the data for the graders by either:
  - downloading it and placing it into your data/raw directory
  - linking to it in a markdown cell
- Create a data frame from the data; use any method to do this (for example read_csv)

```
In [ ]: import matplotlib.pyplot as plt
        import numpy as np
```

```
In [ ]: video = pd.read_csv(file_path, index_col = [0])
        video
```

Out[ ]:

	channelId	channelTitle	videoId	publishedAt	videoTitle	videoDescriptio
0	UCU1_l0ZJyTK_7HZZ3Ruw8Dg	MAPS	pTnk3ziVVRM	2014-01-10T01:24:57.000Z	Psychedelic Horizons Beyond Psychotherapy Work...	Watch the full workshop a http://psychedelics.
1	UCLuO2lUqHrPIlpx0hFenV2g	Tink Tink Club	cuJjSeHZIrg	2015-06-18T16:56:04.000Z	Episode 35 - Dr. James Fadiman	Dr. James Fadiman is the father c modern psyc.
2	UCihqrkaOgVMfLNo2W1hSliA	Podcast Bunk	luyuZfWtGgg	2016-05-01T05:33:13.000Z	#325 Microdosing from The Adam and Dr Drew Sho...	Adam and Dr. Drew are solo today an they open.
3	UCgbWWPn3VYYzxjffZbfj9GQ	Alan Springwind	cng_ZhQf8iY	2016-01-25T04:48:22.000Z	Microdosing Away The Same Old Blues	Source https://www.spreaker.com/user/springwi.
4	UCFmLi6X1mojkFZOFngNR9tQ	Drug Education Agency	OpQIQEx7J5A	2014-08-15T10:53:58.000Z	Erschossener Kiffer / Drogen in Mikro-Dosierun...	Von erschossenen "Dealern", vo demonstrierend.
...	...	...	...	...	...	.
110	UCjnXuO6mfE0-czvcbOYnpbQ	Ayahuasca Magic	j9HFlDff2XY	2016-03-25T04:58:34.000Z	Time differentials: mystery wisdom and microdo...	The lectures I talk about are free (som of th.
111	UC0dsjSDsLZ5hbmoNjvsHJsQ	fluidresearch	lOGAkCtT4wk	2010-08-20T14:46:56.000Z	Fluid filling with exact precision dosing by F...	See how the industry leaders in Precisio micr.
112	UCmNZB3-zCyMuHCOTraws0dg	Helix Steel Victoria	sTNaGjt6564	2014-05-19T12:14:36.000Z	Accurate dosing of Helix Twisted Steel Micro R...	© Copyright --- Look at the steady rai of Hel.
113	UC281yIVp-U8Ijg4KUVRgDRA	Osiron X	fMlQoGlNGHo	2016-04-15T19:31:48.000Z	Psilocybin Mushroom Micro-Dosing	When too much time lapses betwee trips, the e.
114	UCH4ezCqRXHQhsETfl0dainA	FabFormIndustries	ODX4tzCWhhU	2015-08-19T18:25:20.000Z	Dosing Helix Micro Rebar	Coquitlam concrete goes into detail hov they d.

115 rows × 16 columns

1. Using the Data

- In a markdown cell, describe what you'd like to use the data for:
  - repeat the same analysis that you did previously in plain python, but with pandas instead
  - perhaps you would simply like to clean it up for further analysis later
- Write code to achieve what you've written out above. The code should contain at least 4 (repetition is allowed) of the following:
  - 1 filling in missing values
  - 1 type conversion
  - 1 transform a column
  - 1 create a new calculated column
  - 1 visualization
  - 1 calculate summary statistics
  - 1 calculate value counts
- In a markdown cell above your code, write out which of the above requirements you're implementing. As you write your code, document your process in an accompanying markdown cell.

Question 1:

When was the most video published in this dataset?

- Who (population): youtube viewers

- What (subject, discipline): youtube videos
- Where (location): online
- When (snapshot, longitudinal): not applicable
- How much data do you need to do the analysis/work: column 'publishedAt'

Question 2:

What is the statistics of view count of videos in this dataset (mean, median, mode, min, max)?

- Who (population): youtube viewers
- What (subject, discipline): youtube videos
- Where (location): online
- When (snapshot, longitudinal): not applicable
- How much data do you need to do the analysis/work: column 'viewCount' of the dataset

Question 3:

Is duration of the video correlated with view counts of videos?

- Who (population): youtube viewers
- What (subject, discipline): youtube videos
- Where (location): online
- When (snapshot, longitudinal): not applicable
- How much data do you need to do the analysis/work: column 'durationSec', 'viewCount' of the dataset

Question 3:

Is video Category associated with video like counts?

- Who (population): youtube viewers
- What (subject, discipline): youtube videos
- Where (location): online
- When (snapshot, longitudinal): not applicable
- How much data do you need to do the analysis/work: column 'likeCount' and 'videoCategoryId'of the dataset

---

I transformed the "publishAt" column ty datetime type to get the year that has the most number of published video in the dataset.

In [ ]:

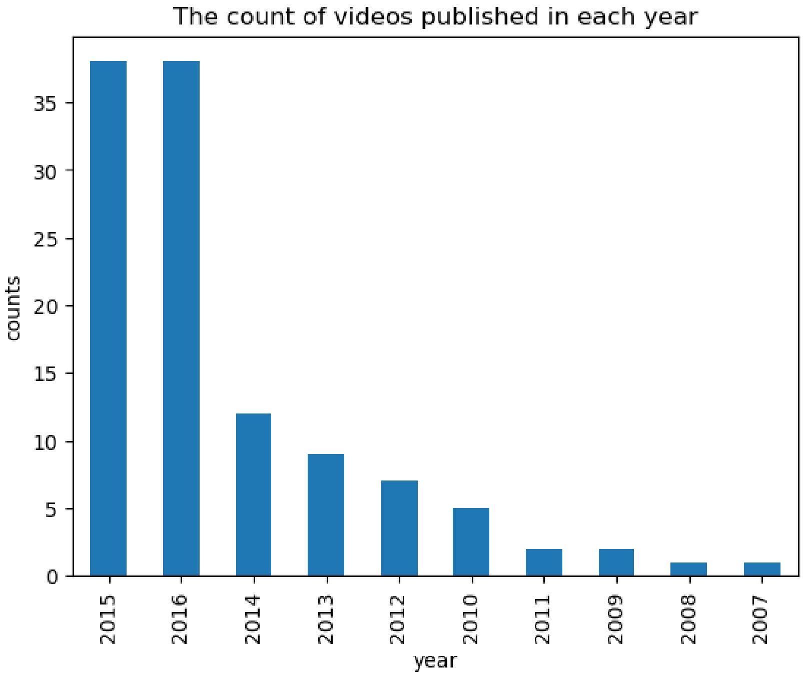
```
video.publishedAt = pd.to_datetime(video.publishedAt)
video.publishedAt.dt.year.value_counts()

print(f'Question1: year {video.publishedAt.dt.year.value_counts().index[0]} is the year that has the most number o
    {video.publishedAt.dt.year.value_counts().values[0]} ')
```

```
Question1: year 2015 is the year that has the most number of published video in the dataset:     38
```

I used the value_counts() on the year attribute of the "publishedAt" columns, and then visualize it by making a bar plot of the count of videos published in each year.

In [ ]:

```
video.publishedAt.dt.year.value_counts().plot.bar()
plt.title('The count of videos published in each year')
plt.xlabel('year')
plt.ylabel('counts')
plt.show()
```



I calculate summary statistics for the "viewCount" column of the dataset. I calculate its mean, median, min value, max value and standard deviation.

In [ ]:

```
mean = np.mean(video.viewCount)
median = np.median(video.viewCount)
min = video.viewCount.min()
max= video.viewCount.max()
std = np.std(video.viewCount)

print(f'Question2: summary statistics of view counts:\n \
    mean:{mean:.2f},\n \
    median:{median:.2f},\n \
    min:{min},\n \
    max:{max},\n \
    standard deviation:{std:.2f},\n')
```

```
Question2: summary statistics of view counts:
     mean:9164.05,
     median:356.00,
     min:2,
     max:526243,
     standard deviation:51784.62,
```

I fill out nan values in "likeCount", "dislikeCount", "commentCount" columns by replacing them with their means respectively.

In [ ]:

```
video.likeCount.fillna(video.likeCount.mean(), inplace = True)
video.dislikeCount.fillna(video.dislikeCount.mean(), inplace = True)
video.commentCount.fillna(video.commentCount.mean(), inplace = True)
```

I transform the "durationSec" column of the dataset from string to floats to see its correlation with view counts.

In [ ]:

```
video.durationSec = pd.to_numeric(video.durationSec, errors = 'coerce')
corr = video[['durationSec', 'viewCount']].corr()

print(f'Question3: The correlation between video duration and view counts is {corr.iloc[0, 1]:.4f}')
print('There does not seem to be a correlation between video duration and video view counts.')
```

```
Question3: The correlation between video duration and view counts is 0.0744
There does not seem to be a correlation between video duration and video view counts.
```

In [ ]:

```
corr = video[['videoCategoryId', 'likeCount']].corr()

print(f'Question4: the correlation between video category and like count is: {corr.iloc[0, 1]:.4f}')
print('There does not seem to be a correlation between video category and video like counts.')
```

```
Question4: the correlation between video category and like count is: 0.0720
There does not seem to be a correlation between video category and video like counts.
```

- 2 type conversion

    - I transformed the "publishAt" column ty datetime type to get the year that has the most number of published video in the dataset.
    - I transform the "durationSec" column of the dataset from string to floats to see its correlation with view counts.
- 1 calculate value counts, 1 visualization
    - I used the value_counts() on the year attribute of the "publishedAt" columns
    - I visualize it by making a bar plot of the count of videos published in each year.
- 1 calculate summary statistics
    - I calculate summary statistics for the "viewCount" column of the dataset. I calculate its mean, median, min value, max value and standard deviation.
- 1 filling in missing values
    - I fill out nan values in "likeCount", "dislikeCount", "commentCount" columns by replacing them with thir means respectively.