DS 201X – homework 3

More seaborn pairplot

Instruction:

Using the USvideos.csv from Kaggle's "Trending YouTube Video Statistics" to visualize the correlations of "Views" and "Likes" and "Channel ID" on Youtube.

(Ref: https://www.kaggle.com/datasnaek/youtube-new/data)
Import the data from the **USvideos.csv** to Pandas DataFrame

Explore the Data

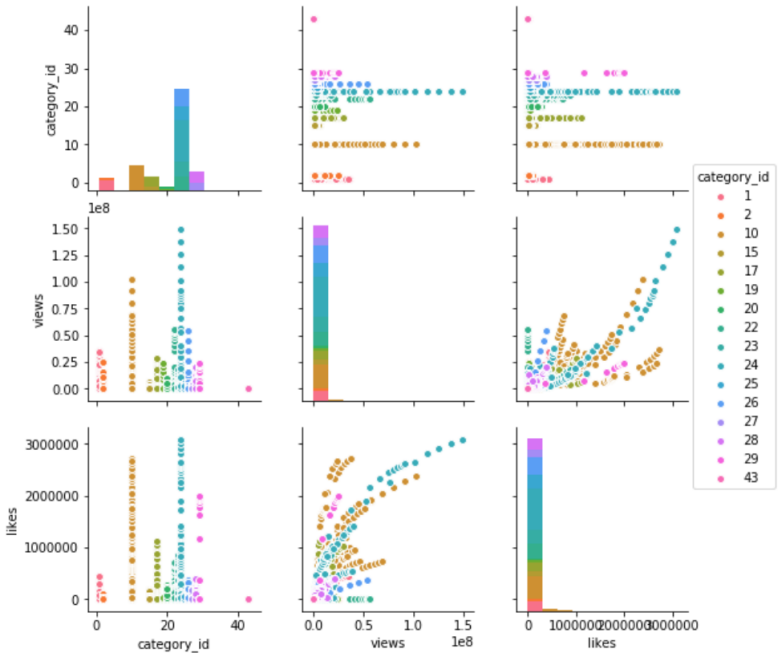
1. Explore your data by using .head(), .info()
2. Find and remove any duplicate rows in the data (if any).

Work on your dataset:

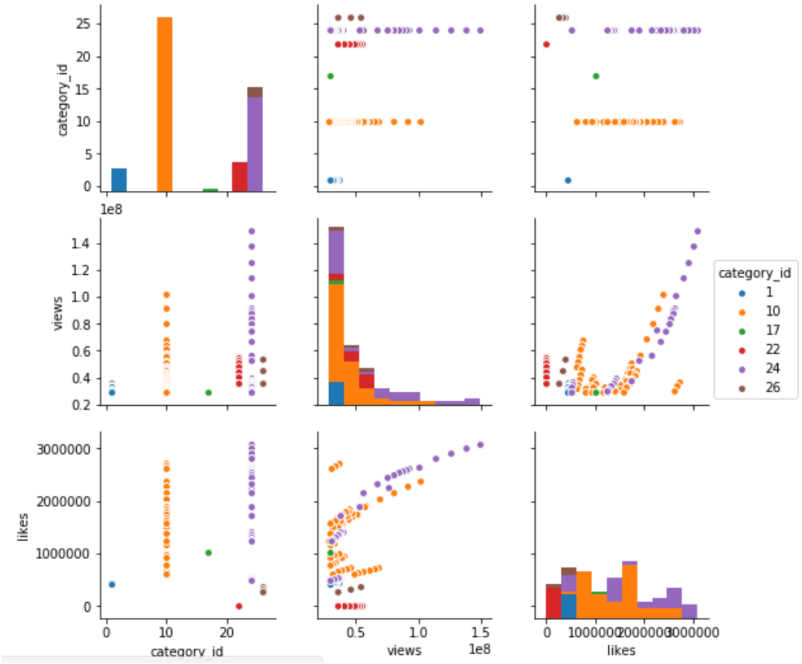
1. To find a correlations of "views" and "likes" and "category_id" on Youtube.
2. [Optional,] you can marge the category name from another dataset to the category_id for better understanding of the data
3. Try select top rows from the DataFrame by "views", using df.nlargest(100, "views")
4. Try select top rows from the DataFrame by "likes" and "views", using df.nlargest(10, ['likes','views'])
5. Since the top views videos on Youtube could behaved very differently from the rest of 20K videos in the list, Select the top 100 videos by it "views" then create a second DataFrame name "df_top100views".

Create following charts:

1. Create a seaborn pairplot chart that shows "views" and "likes" and "category_id" relationship.



2. Create a seaborn pairplot chart that shows “views” and “likes” and “category_id” relationship from only top views 100 videos, by df_top100views



3. Observe and describe the differences from those two plots.

Show your chart in class.

Also submit your Jupyter Notebook file (.ipynb file) on Canvas.