- Product recommendations based on the key findings from EDA

  (Refer to "Correlation data analysis for product recommendation" in the next page)
  - Encourage users to create long duration video : 2~4 minutes
    - "Video watch duration" and "Number of share" are the key parameters that indicate the success of the products.
    - "Video watch duration" is positively correlated with "video duration".
    - "Video watch duration" is positively correlated with "video duration" and video watch duration
  - Recommend specific music and games to users based on the user's video profile in the product. Most impatient users would take the recommendation
    - Based on the correlation analysis, video id is correlated with game id, music id. Using this information, it is recommended to promote certain music and games to users
  - Minor recommendation: Product is mostly popular in India, Russia, China and a few European countries. There is room for expanding markets in western countries. Recommend to increase the marketing effort in northern America and more European countries.

- Analytic details (Google Colab with Pandas dataframe)
  - Step1: Basic information check based on pandas dataframe
    - Create 4 dataframes based on 4 csv files and run the first level data sanity check

0. Watched_Videos

	count	mean	std	min	25%	50%	75%	max
video_watch_duration	50000	3.41023	1.83826	0.1	2.05	3.3	4.88	7

	count	unique	top	freq
coordinates	50000	50000	(Decimal('14.439490'), Decimal('105.291233'))	1
os	50000	2	Android	25037
videoId	50000	5000	241ed4cb-e1bc-4892-9d29-304a9b5f8848	24
userId	50000	1000	39acf63e-c53c-43b9-a250-7a391070861a	72
appVersion	50000	10	2.1.2	5134
video_watch_dateTime	50000	49649	2021-12-11 05:13:52	3
feedId	50000	10	42568c8a-efbe-4293-81a3-513e33094ade	5116

1. videos

	count	mean	std	min	25%	50%	75%	max
video_duration	5000	4.53651	1.4027	2.1	3.33	4.555	5.72	7
num_shares	5000	9.8488	5.99553	0	5	10	15	20

	count	unique	top	freq
id	5000	5000	6e884071-8c0e-4782-aa40-e0fd78243360	1
attribute_id	5000	11	25e7fab9-abfb-4a74-a8cb-c3adbc6e3531	483
music_id	5000	10	c3721b09-6896-465a-8e13-4d73efd06874	550
animation_id	5000	15	d167d0af-a662-48ad-bb39-5aa306fc8141	359
creator	5000	992	ddfeab0f-274a-4cab-86c2-873cbd8e485e	12
video_created_at	5000	5000	2021-08-04 10:42:39	1
game_id	5000	5	8bd4c7be-fe87-45ad-990d-d131cd3d9c77	1021

2. Users

	count	unique	top	freq
id	1000	1000	37e369a8-4477-4d5d-acd7-4e3b8059586e	1
user_created_at	1000	1000	2021-03-23 00:24:53	1
class	1000	5	Gamer	213
coordinates	1000	626	('-6.25', '38.66667')	6

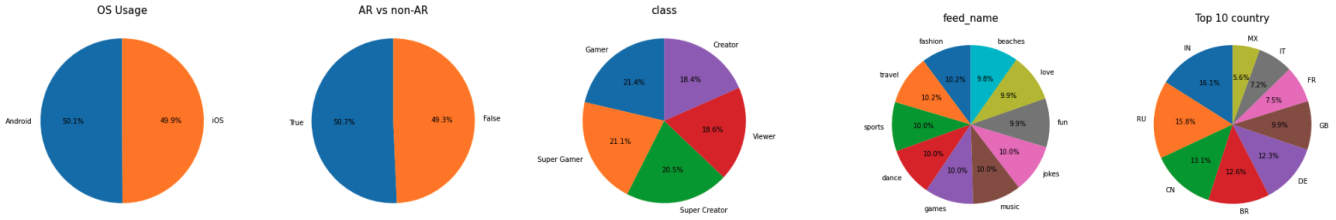
	count	unique	top	freq
id	1000	1000	37e369a8-4477-4d5d-acd7-4e3b8059586e	1
user_created_at	1000	1000	2021-03-23 00:24:53	1
class	1000	5	Gamer	213
coordinates	1000	626	('-6.25', '38.66667')	6

3. Feeds

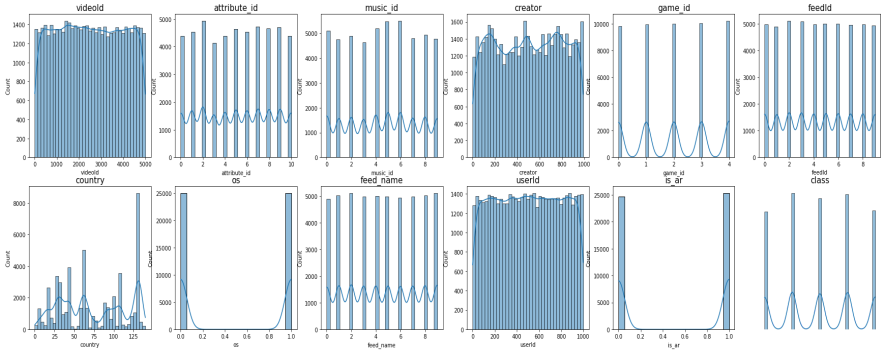
	count	unique	top	freq
id	10	10	27a87df4-da75-451f-b8f6-af9dd0a4ee00	1
feed_name	10	10	music	1

	count	unique	top	freq
id	10	10	27a87df4-da75-451f-b8f6-af9dd0a4ee00	1
feed_name	10	10	music	1

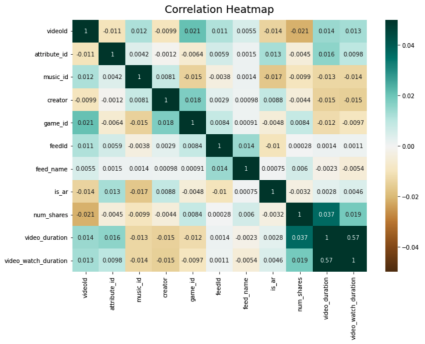
  - Step 2: Merge data based on the inner joint.
    - All data frames have common columns that enables merging all data frames into one universal frame
  - Step 3: Run data visualization and correlation study
    - Simple uni-variable data analysis
      - Pie chart

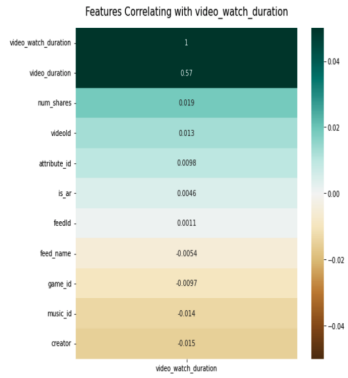


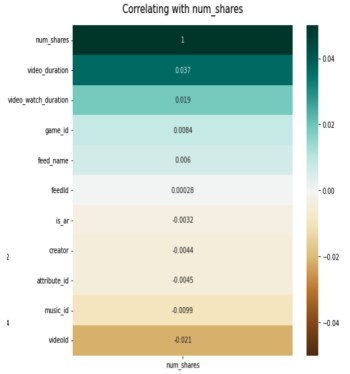
- Histogram



  - Correlation data analysis for product recommendation
    - 1st plot: Correlation among all the variables. Categorical variables are coded to unique integer based on pandas “dataframe.cat.codes”
    - 2nd plot: Correlation of Video watch duration as dependant variable
    - 3rd plot: Correlation of number of shares as dependant variable







- Dataframe analysis based on groupby ([“game_id”, “music_id”])

game_id	music_id	count	unique	top	freq
2609b6b5-3a08-451e-bce5-75fe506559d2	19e4db13-d854-4913-a8ca-3612ee1a6071	988	96	8aa4c9a8-ec26-411f-bc79-06ed62a4eb3b	17
	1f80192e-3bff-41b8-b272-7062b9d7b6de	997	99	23add3ab-f66d-49a7-a368-5303e113173c	18
	3de11516-8627-47c3-98ab-265b429cac91	933	93	5a716136-a4c9-4188-94d6-ca35a016fde6	17
	4ee0de5b-e27f-4d73-a3c6-52cf26e36c54	992	100	4b83d671-2164-4828-a855-b8b193c4da38	18
	8ff50feb-3b5d-4dd0-84b0-8409753bd166	822	84	56c9a3ec-5110-4d80-8191-27f8bc45b4ad	19
	b5b679a5-e769-4d95-82a6-f75e439b1ef9	1142	111	067e4a05-aa5f-4cca-86ac-b96a59dcf0b5	19
	c3721b09-6896-465a-8e13-4d73efd06874	1090	108	0e768791-47ec-45ce-92e8-95461199d3cb	19
	d53234f4-1efb-4299-97c1-1165818cba5	860	82	32ba0f44-5fba-40bd-9243-4d58f3555620	18
	da2fc9b0-32d5-4a1b-8b40-2706eb133bf0	1055	104	6eafe985-e3d8-41b2-b80a-9c32dc7734cd	17
	e2bf5d12-edc2-47cb-b23f-1f4f7c1a2292	942	93	58a64556-54ad-49f0-b475-885e1b9fd22b	18
4d54f6ab-cd8f-43a5-9be0-c497c2c32d24	19e4db13-d854-4913-a8ca-3612ee1a6071	931	99	d6f8a61c-5236-4040-a831-61449e2b840b	17

The operation reveals that certain game id and music are more tightly coupled with a specific video ID. Product can recommend game and music based on the user data

- Next steps
  - Change the data sampling method to capture more realistic data distribution. Currently all variables show the uniform distribution and it affects the correlation results. Gather data based on the stratified sampling and re-run the EDA
  - Build game and music recommendation algorithm development based on the user data : Classification model based on machine learning algorithm development