INSY670-076

Individual Assignment

Is a Picture Worth a Thousand Words?

Name: Arham Anwar
McG ID: 261137773

This assignment aims to guide an aspiring influencer on what types of Instagram pictures they should and shouldn't post. That includes identifying the types of images that have a high chance of increasing engagement and those that have a high chance of low engagement.

To establish some definitions and expectations for the project, we've made two assumptions to maximize insights. First, 'engagement' in this project refers to comment activity, the ratio of comments to likes, and the ratio of likes to followers, in the mentioned order. Second, the client is assumed not to be the type that drives engagement due to external accolades such as mastery in singing, sports, or other endeavors. In other words, the client is assumed to be a regular person aspiring to increase engagement solely based on the content they post on Instagram, without relying on external mastery in the arts or other fields.

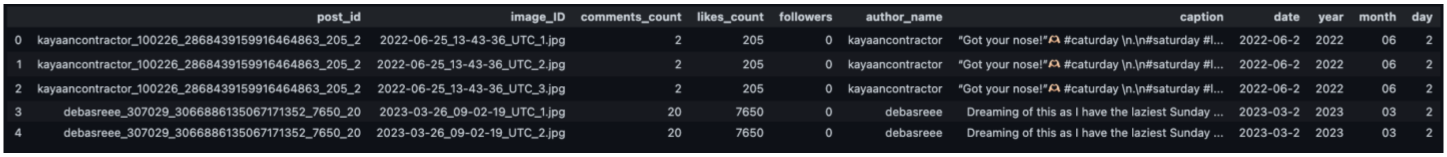
Task A: Social Media Content Performance Breakdown

Step 1: Automated Extraction of JSON files

Extracted JSON files containing detailed information about each post, including comments, likes, and engagement, to streamline data analysis through an automated script ("Step-1-Extract.py"). All the files are extracted in their home directory to keep files organized in the 'author:post:image' hierarchy

Step 2: Tabularizing Authors, Posts, Images hierarchy & engagement meta data

In this step, we had processed Instagram post data stored in the 'Data' directory. For each post, we had iterated through its folder, extracting relevant information from JSON files such as comments count, likes count, author name, and caption. We had also checked for a separate file containing follower count data. We had then compiled this data along with image file names into a list. This list had been converted into a Pandas DataFrame for easier manipulation. Additionally, we had extracted the date from the image file name and reordered the DataFrame columns. Finally, we had displayed the DataFrame and saved it as a CSV file named 'tabular_data.csv'. This step had facilitated data organization and had prepared the data for further analysis. A snapshot of the transformed data frame is shown in Figure 1.



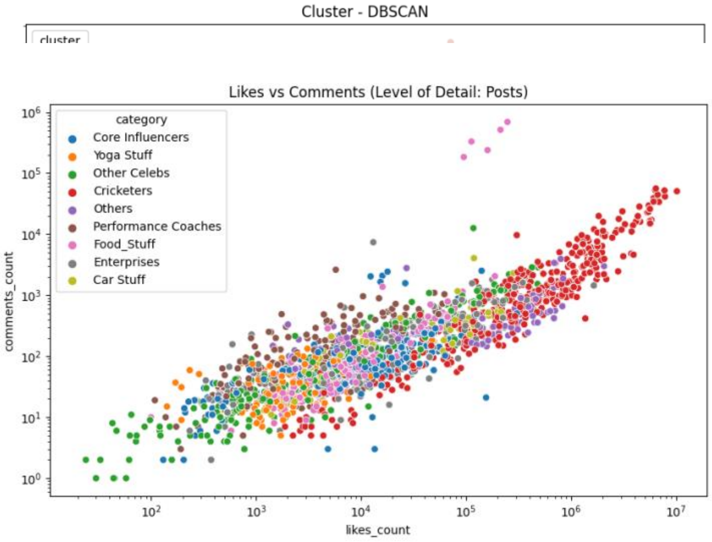
	post_id	image_ID	comments_count	likes_count	followers	author_name	caption	date	year	month	day
0	kayaancontractor_100226_2868439159916464863_205_2	2022-06-25_13-43-36_UTC_1.jpg	2	205	0	kayaancontractor	"Got your nose!"🫶 #caturday \n.\n#saturday #l...	2022-06-2	2022	06	2
1	kayaancontractor_100226_2868439159916464863_205_2	2022-06-25_13-43-36_UTC_2.jpg	2	205	0	kayaancontractor	"Got your nose!"🫶 #caturday \n.\n#saturday #l...	2022-06-2	2022	06	2
2	kayaancontractor_100226_2868439159916464863_205_2	2022-06-25_13-43-36_UTC_3.jpg	2	205	0	kayaancontractor	"Got your nose!"🫶 #caturday \n.\n#saturday #l...	2022-06-2	2022	06	2
3	debasreee_307029_3066886135067171352_7650_20	2023-03-26_09-02-19_UTC_1.jpg	20	7650	0	debasreee	Dreaming of this as I have the laziest Sunday ...	2023-03-2	2023	03	2
4	debasreee_307029_3066886135067171352_7650_20	2023-03-26_09-02-19_UTC_2.jpg	20	7650	0	debasreee	Dreaming of this as I have the laziest Sunday ...	2023-03-2	2023	03	2

Figure 1: Snippet of top 5 rows of tabularized dataframe

Step 3: Investigation of Posts

- A quick aggregation summary revealed there were 69 unique authors, 1968 unique post IDs and 5511 images. A Pivot table at the level of detail of authors was made summarizing number of posts, comments (minimum, maximum, average), number of likes and other basic attributes. It reveals that many of the influencers with top engagement are cricket players of the Indian national team as shown in Figure 2 snippet below. Note how, for an average client, being a cricket player from the get-go is very unlikely thus these top influencers are not a good sample to study for post recommendations. Therefore, we will ignore the cricketers from the study.

- Next, posts before 2021 were filtered out to ensure that the analysis focuses on recent data, which is more relevant for understanding current trends and behaviors on social media platforms. Removing irrelevant rows improves the accuracy and reliability of subsequent analyses & recommendations.
- Since Google Vision has an associated cost, using vision label API maybe a costly affair. So we created likes-vs-comments plots by author and by post to cluster authors. We first did clustering by DBScan and then by KMedoids. These cluster reveal interesting information about the Influencers as shown in figure 2 and figure 3. Finally using prior knowledge and some Instagram study, the influencers were categorized as shown in figure 4 and figure 5. Clustering enables the identification of groups of influencers with similar engagement profiles, facilitating targeted strategies for influencer marketing campaigns



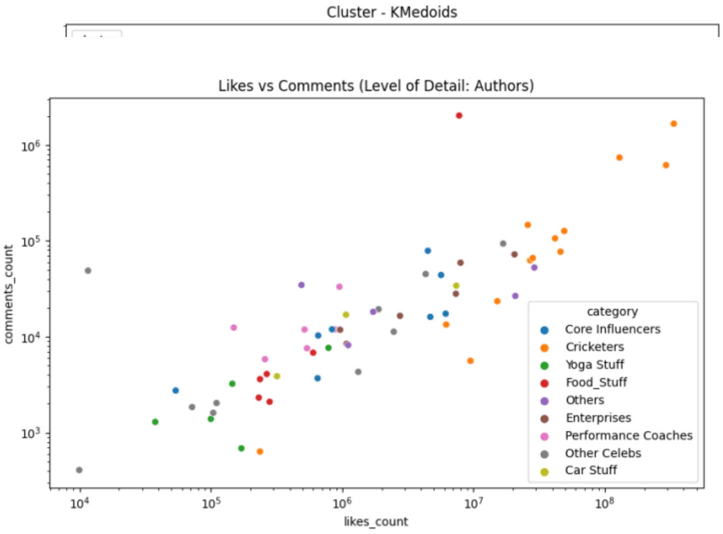


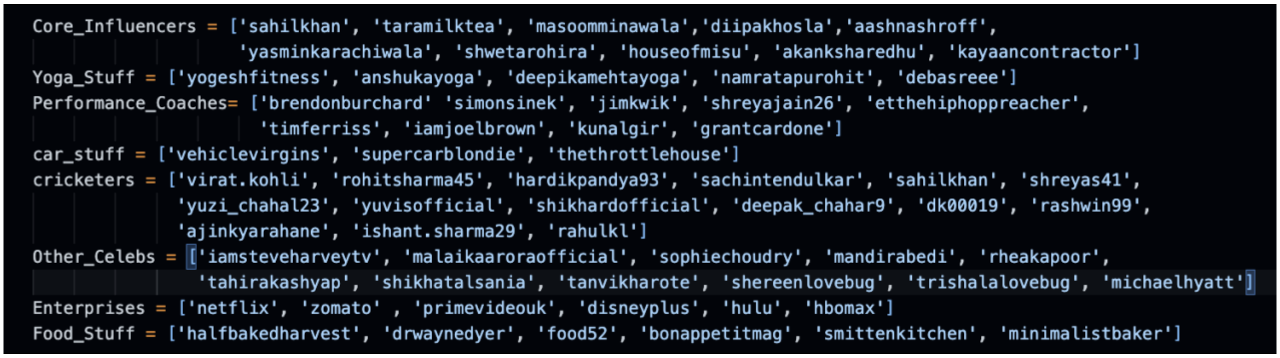
Figure 2: DBScan on Log of #Likes & #Comments

Figure 3: KMedoids Log of #Likes & #Comments

Figure 4: Author Categories Distribution By Posts

Figure 5: Author Categories Distribution

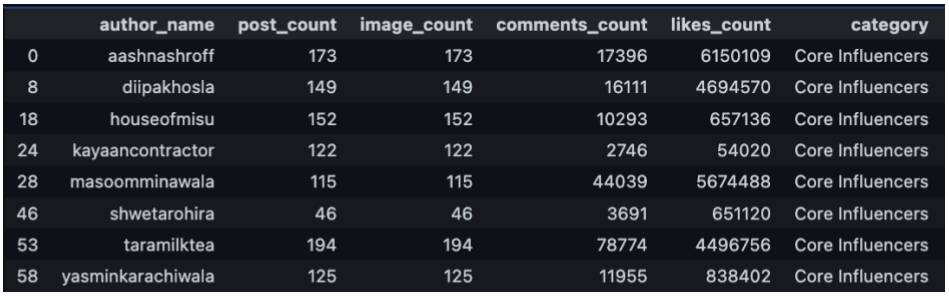
- Author Categorization: Categorizing authors based on their domain and engagement helps in understanding the composition of influencers within the dataset. This categorization allows for targeted analysis and insights into specific groups of influencers, such as core influencers or performance coaches, or cricketers or other categories which may have distinct engagement patterns. The categories and influencers are shown in figure 6.



```
Core_Influencers = ['sahilkhan', 'taramilktea', 'masoomminawala','diipakhosla','aashnashroff',
                    'yasminkarachiwala', 'shwetarohira', 'houseofmisu', 'akanksharedhu', 'kayaancontractor']
Yoga_Stuff = ['yogeshfitness', 'anshukayoga', 'deepikamehtayoga', 'namratapurohit', 'debasreee']
Performance_Coaches= ['brendonburchard' 'simonsinek', 'jimkwik', 'shreyajain26', 'etthehiphoppreacher',
                      'timferriss', 'iamjoelbrown', 'kunalgir', 'grantcardone']
car_stuff = ['vehiclevirgins', 'supercarblondie', 'thethrottlehouse']
cricketers = ['virat.kohli', 'rohitsharma45', 'hardikpandya93', 'sachintendulkar', 'sahilkhan', 'shreyas41',
              'yuzi_chahal23', 'yuvisofficial', 'shikhardofficial', 'deepak_chahar9', 'dk00019', 'rashwin99',
              'ajinkyarahane', 'ishant.sharma29', 'rahulkl']
Other_Celebs = ['iamsteveharveytv', 'malaikaaroraofficial', 'sophiechoudry', 'mandirabedi', 'rheakapoor',
                'tahirakashyap', 'shikhatalsania', 'tanvikharote', 'shereenlovebug', 'trishalalovebug', 'michaelhyatt']
Enterprises = ['netflix', 'zomato' , 'primevideouk', 'disneyplus', 'hulu', 'hbomax']
Food_Stuff = ['halfbakedharvest', 'drwaynedyer', 'food52', 'bonappetitmag', 'smittenkitchen', 'minimalistbaker']
```

Figure 6 : Categorization of Influencers based on clustering and manual screening

- Final Selection: Filtering out authors belonging to the 'Core Influencers' cluster allows for a focused analysis on influencers who have achieved significant traction through Instagram tactics only! This selection ensures that insights derived from further analysis are relevant and actionable for understanding successful influencer strategies and not driven by external successes such as accolades in sports, music, film, cooking etc. This selected batch of influencers are feeded to Google Vision via API calls for labelling of images. 'Core influencers' are listed below, and they lie in middle most region of the like-comment plots (log scaled), meaning they have less engagement than cricket star and other celebrities but more engagement than 'yoga' and 'guru' influencers. Selected batch is below in figure 7



	author_name	post_count	image_count	comments_count	likes_count	category
0	aashnashroff	173	173	17396	6150109	Core Influencers
8	diipakhosla	149	149	16111	4694570	Core Influencers
18	houseofmisu	152	152	10293	657136	Core Influencers
24	kayaancontractor	122	122	2746	54020	Core Influencers
28	masoomminawala	115	115	44039	5674488	Core Influencers
46	shwetarohira	46	46	3691	651120	Core Influencers
53	taramilktea	194	194	78774	4496756	Core Influencers
58	yasminkarachiwala	125	125	11955	838402	Core Influencers

Figure 7: Core Influencers (Focus of Study)

Step 4: Google Vision Image Labelling

- This code snippet had read a CSV file containing data of shortlisted influencers' posts, had filtered it to include only posts from specific authors (core influencers), and then had attempted to tag the images in these posts using Google Vision API. However, the tagging functionality had been commented out in the code to avoid accidental reruns and associated token costs.

- The structure of the file 'Shortlisted_Influencer_Posts_With_Labels.csv' would have likely included columns representing various attributes of each post, such as 'author_name', 'post_id', 'image_ID', and potentially additional columns for the detected labels from Google Vision. Each row in the file had corresponded to a specific post by a core influencer, with associated image IDs and any detected labels from image analysis. This file had served as a consolidated dataset for further analysis, potentially revealing insights into the content themes and characteristics associated with the selected influencers' posts.

- In the provided code, the data had already been read and filtered before the LDA topic modeling process began. Once the DataFrame was prepared, a preprocessing function had been meticulously crafted to transform the data into a bag-of-words representation, a crucial step in the LDA methodology.

Step 5: Topic Modelling & Optimal Number of Topics

- After this meticulous preprocessing, the process had ventured into the realm of determining the optimal number of topics. Here, a loop had been meticulously crafted to iterate through different numbers of topics, ranging from 2 to 20, with a discerning step of 1. Each iteration had meticulously trained an LDA model and evaluated its coherence, a measure of how interpretable and cohesive the topics were. Please see figure 8 for #topic-coherence plot.

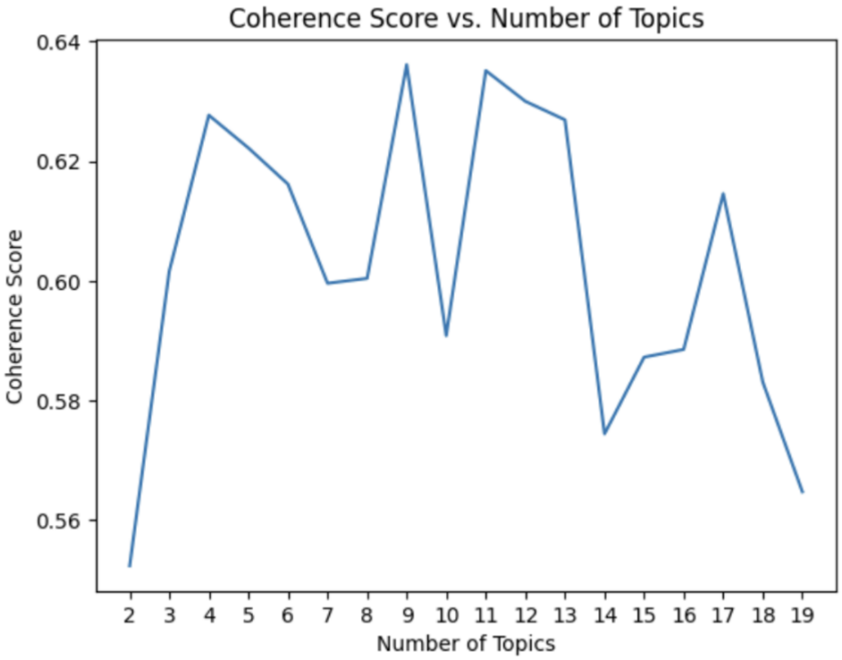


Figure 8: Coherence, #Topics plot

- The optimal number of topics had been judiciously selected based on the coherence score. This selection had been made after a thorough evaluation of the coherence scores across different numbers of topics, ensuring that the chosen number yielded the most coherent and insightful topics possible. #topics 9 had the highest coherent score. On the other side, #topics = 4 also had a high coherence score. Both were interpreted, because 9 can possibly have some duplication and maybe difficult to keep track of, therefore we checked if it makes sense to have 9 or 4, but #topics = 9 had very clear and meaningful topics.

- Following the selection of the optimal number of topics, the LDA model had been meticulously trained once again, this time with the chosen number of topics. The resulting model had been thoughtfully examined to extract the top words for each topic, providing valuable insights into the underlying themes within the dataset.

Topic 1: Gastronomic Delights:

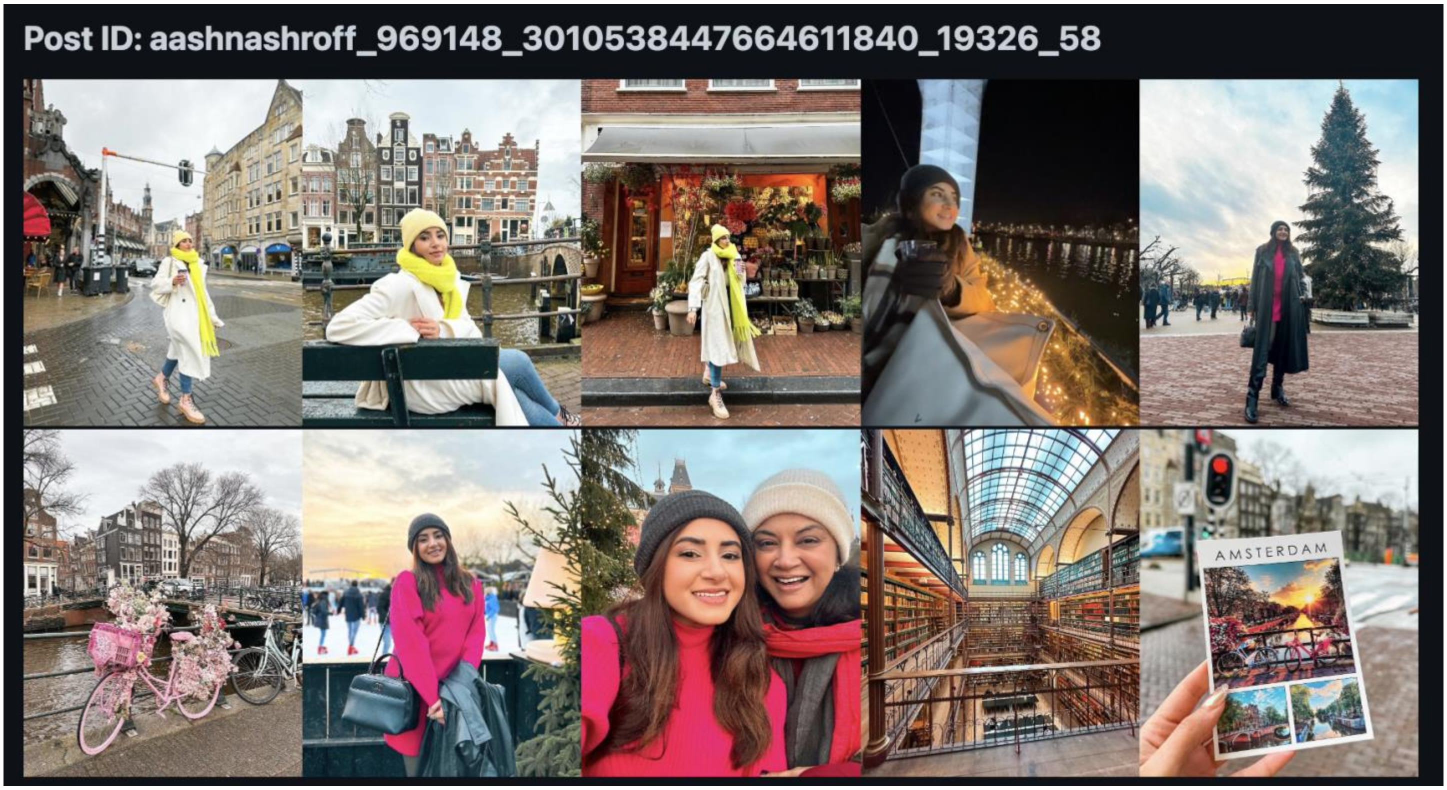
a. *Keywords: Food, Tableware, Ingredient, Cuisine, Recipe, Plate, Cake, Drinkware, Fruit, Sweetness, Natural foods, Cake decorating, Carnivore, Baked goods, Presentation, Culinary, Dessert, Refreshment, Meal, Culinary art, Joy, Food photography, Delicious, Cooking*

b. *Business Meaning: This topic focuses on showcasing culinary creations, recipes, and delightful food presentations. It signifies a strategy aimed at engaging followers with mouth-watering content, potentially appealing to food enthusiasts, and promoting culinary experiences or products.Sample Post:*

Topic 2: Travel Adventures:

a. *Keywords: Sky, Cloud, Water, Leisure, Nature, Smile, Building, City, Road, Tree, Sunglasses, Luggage, Landscape, Adventure, Exploration, Destination, Wanderlust, Excursion, Travelogue, Journey, Discovery, Outdoors, Serenity, Adventure travel*

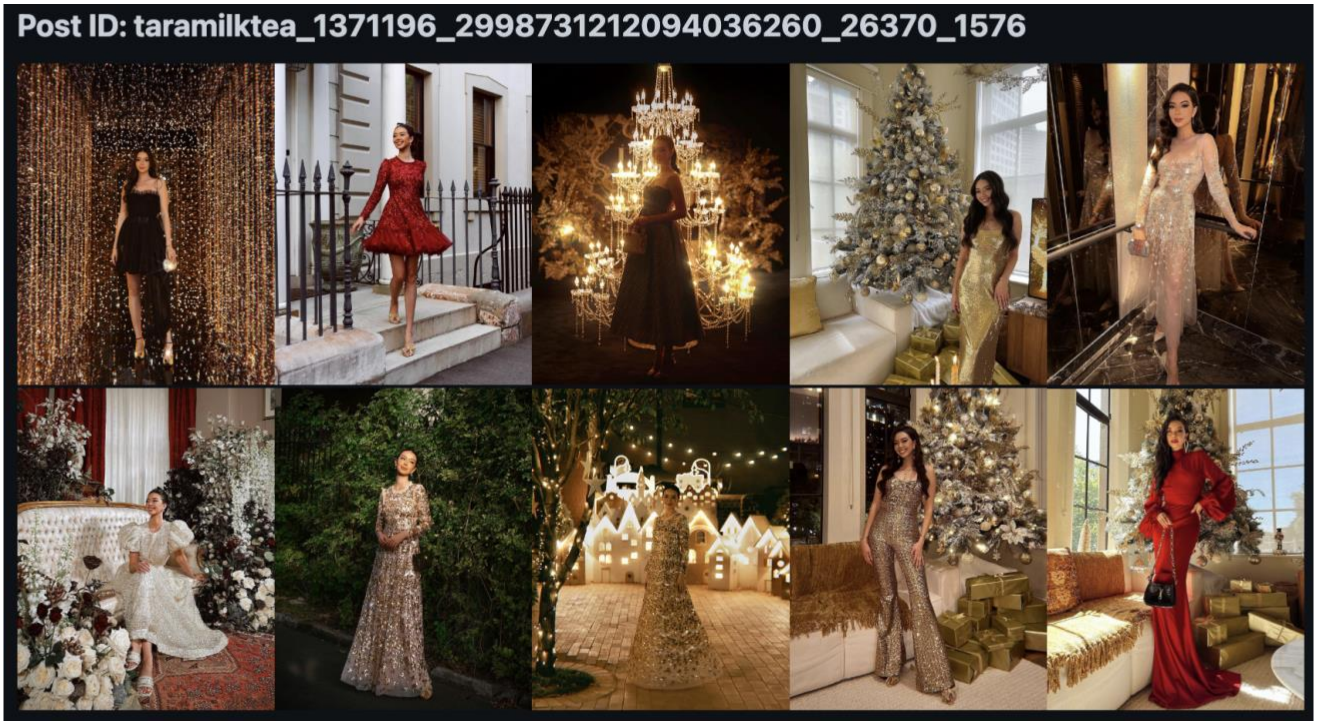
b. *Business Meaning: These posts capture moments from various travel destinations, promoting a sense of adventure and wanderlust. This strategy aims to showcase travel experiences, attract travel enthusiasts, and potentially collaborate with travel brands to promote destinations or travel-related products. Sample Post:*


Post ID: aashnashroff_969148_3010538447664611840_19326_58

Topic 3: Fashion Elegance:

a. *Keywords: Dress, Waist, Fashion, Sleeve, Smile, Shoulder, Flower, Hairstyle, Elegant, Formal wear, Gown, Temple, Textile, Pink, Chic, Trendy, Stylish, Glamorous, Classy, Runway, Fashionista, Couture, Style, Sophisticated*

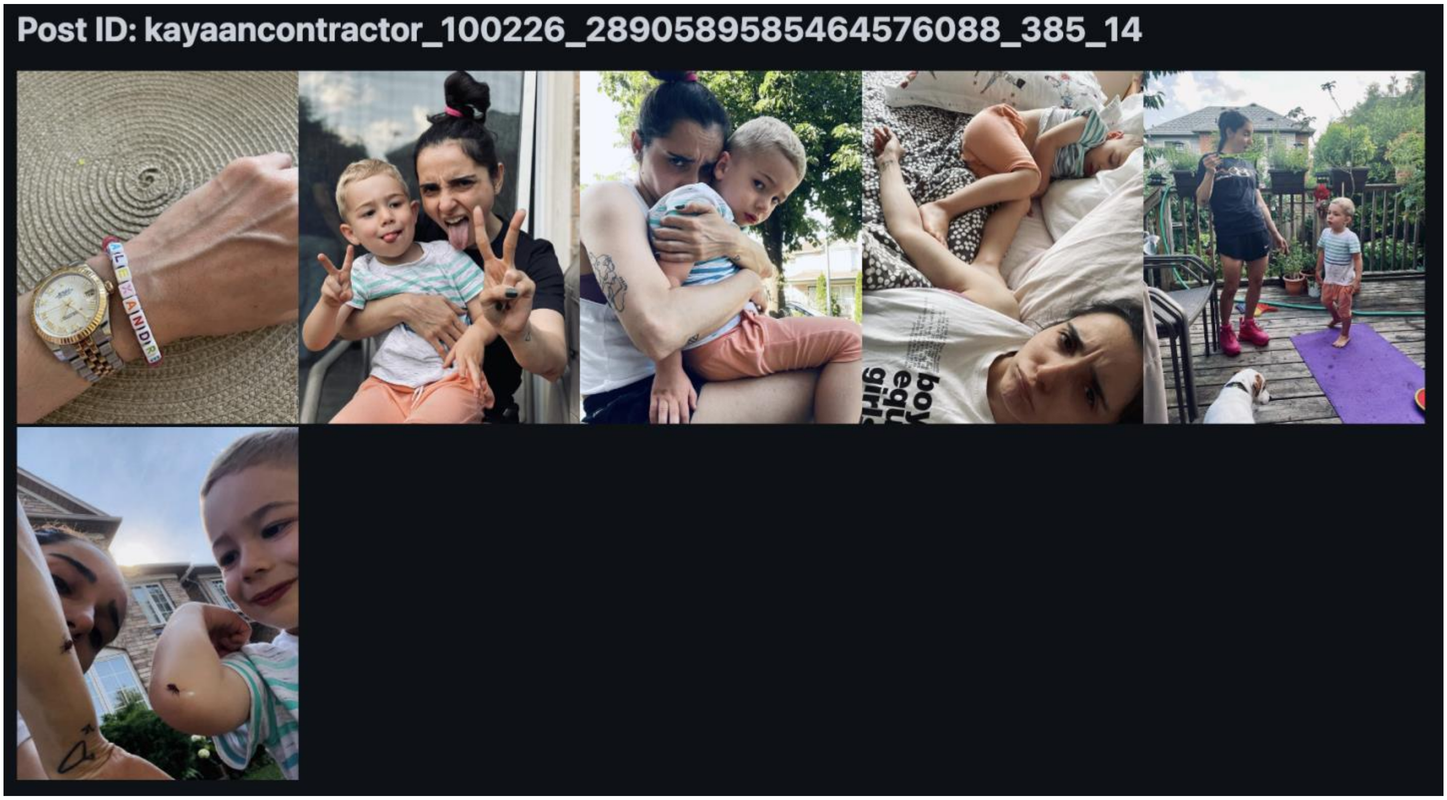
b. *Business Meaning: This topic focuses on elegant fashion trends, featuring stylish outfits and accessories. It represents a strategy targeting fashion-conscious individuals, influencing style choices, and promoting fashion brands or collections to enhance brand image and appeal. Sample:*


Post ID: taramilktea_1371196_2998731212094036260_26370_1576

Topic 4: Joyful Moments:

a. *Keywords: Happy, Smile, Gesture, Skin, Hair, Head, Lip, Face, Fun, Eyebrow, Sleeve, Expression, Eye, Comfort, Leisure, Relaxation, Excitement, Laughter, Contentment, Joy, Emotion, Happiness, Delight*

b. *Business Meaning: These posts capture moments of happiness and joy, emphasizing positivity and emotional connection with the audience. This strategy aims to create uplifting content, strengthen brand perception, and foster customer relationships by evoking positive emotions and associations with the brand. Sample Post:*

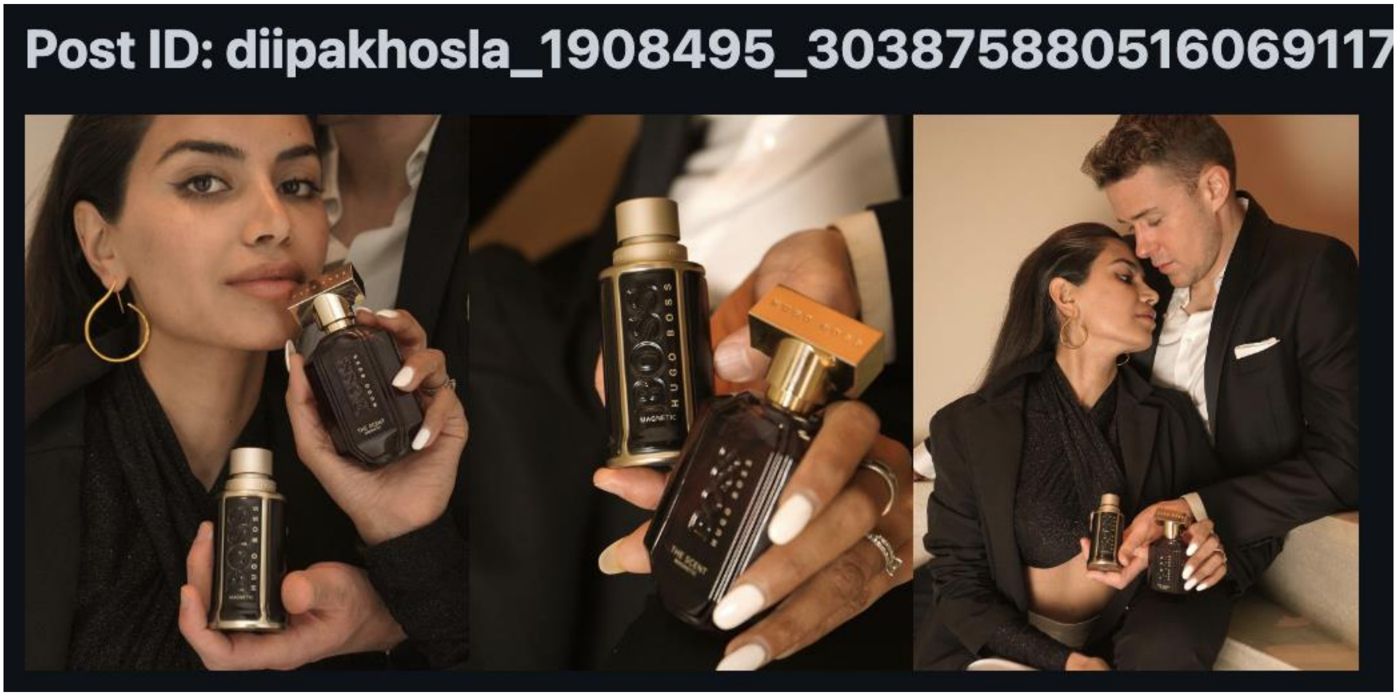


Post ID: kayaancontractor_100226_2890589585464576088_385_14

Topic 5: Entertainment and Leisure:

a. *Keywords: Entertainment, Beverage, Book, Gadget, Watch, Publication, Flash photography, Hand, Font, Finger, Cool, Human, Jewellery, Number, Art, Nail, Sunglasses, Beard, Beverage, Leisure, Fun, Enjoyment, Relaxation*

b. *Business Meaning: This topic encompasses entertainment-related content, including leisure activities and indulging in beverages. It signifies a strategy aimed at providing entertainment value, attracting a diverse audience interested in leisure pursuits, and promoting lifestyle products or experiences associated with relaxation and enjoyment. Sample Post*



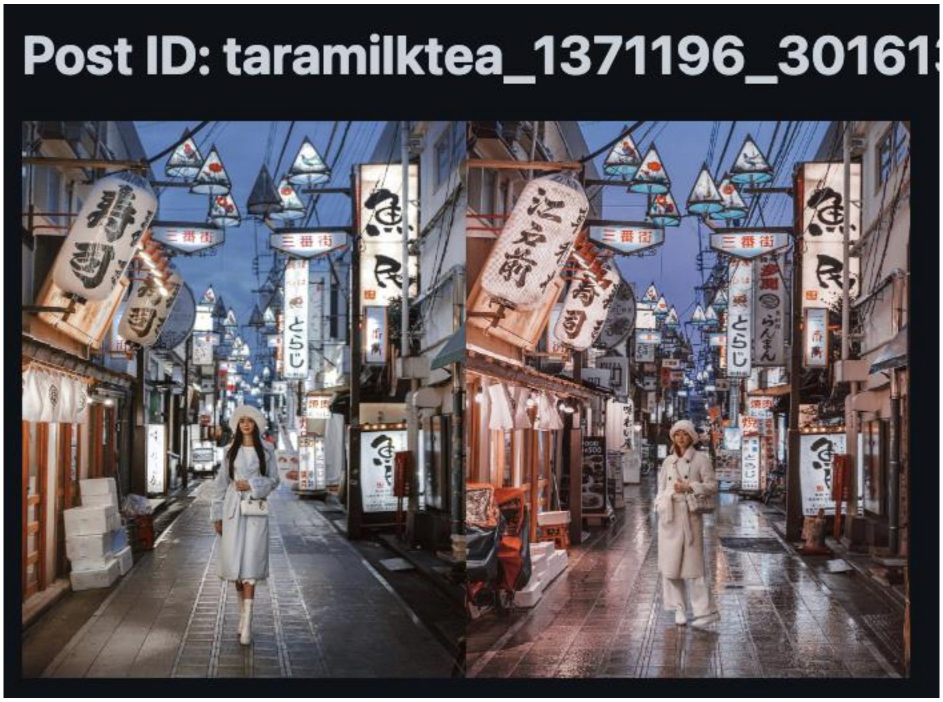
Post ID: diipakhosla_1908495_303875880516069117

Topic 6: Urban Lifestyle:

a. Keywords: Building, Road, Infrastructure, Window, City, Retail, Car, Vehicle, Sidewalk, Architecture, Automotive, Red, Daytime, Food, Crossing, Tire, Lighting, Design, Urban, Street, Modern, Pavement, Urban living

b. Business Meaning: These posts showcase urban landscapes and lifestyle elements associated with city living. This strategy targets urban dwellers, highlighting urban culture, architecture, and lifestyle products or experiences to appeal to consumers interested in city life and urban living.

c. Sample Post



Post ID: taramilktea_1371196_30161

Topic 7: Elegant Fashion Trends:

a. Keywords: Fashion, Sleeve, Shoulder, Gown, Dress, Formal wear, Sunglasses, Lip, Standing, Hairstyle, Smile, Body, Apparel, Chic, Elegance, Trend, Stylish, Classy, Bridal, Magazine, Textile, Clothing, Purple

b. Business Meaning: This topic highlights elegant fashion trends and stylish attire for formal occasions. It represents a strategy aimed at promoting fashion brands, influencing style choices, and engaging with fashion-forward audiences to enhance brand image and appeal in the fashion industry.

c. Sample Post

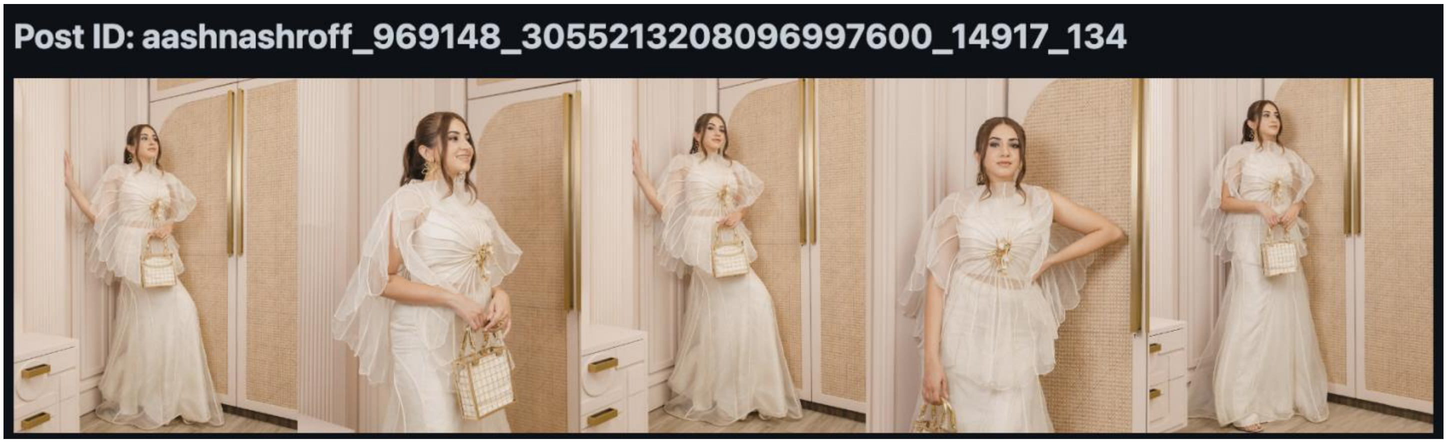
Post ID: kayaancontractor_100226_2892795825581264307_389_36

Topic 8: Relaxation and Leisure:

a. Keywords: Plant, Event, Tree, Sky, Building, Interior, Pool, Decoration, Green, Furniture, Water, Magenta, Comfort, Swimming, Window, Wood, Waist, Yellow, Blue, Happy, Leisure, Relaxation, Tranquility

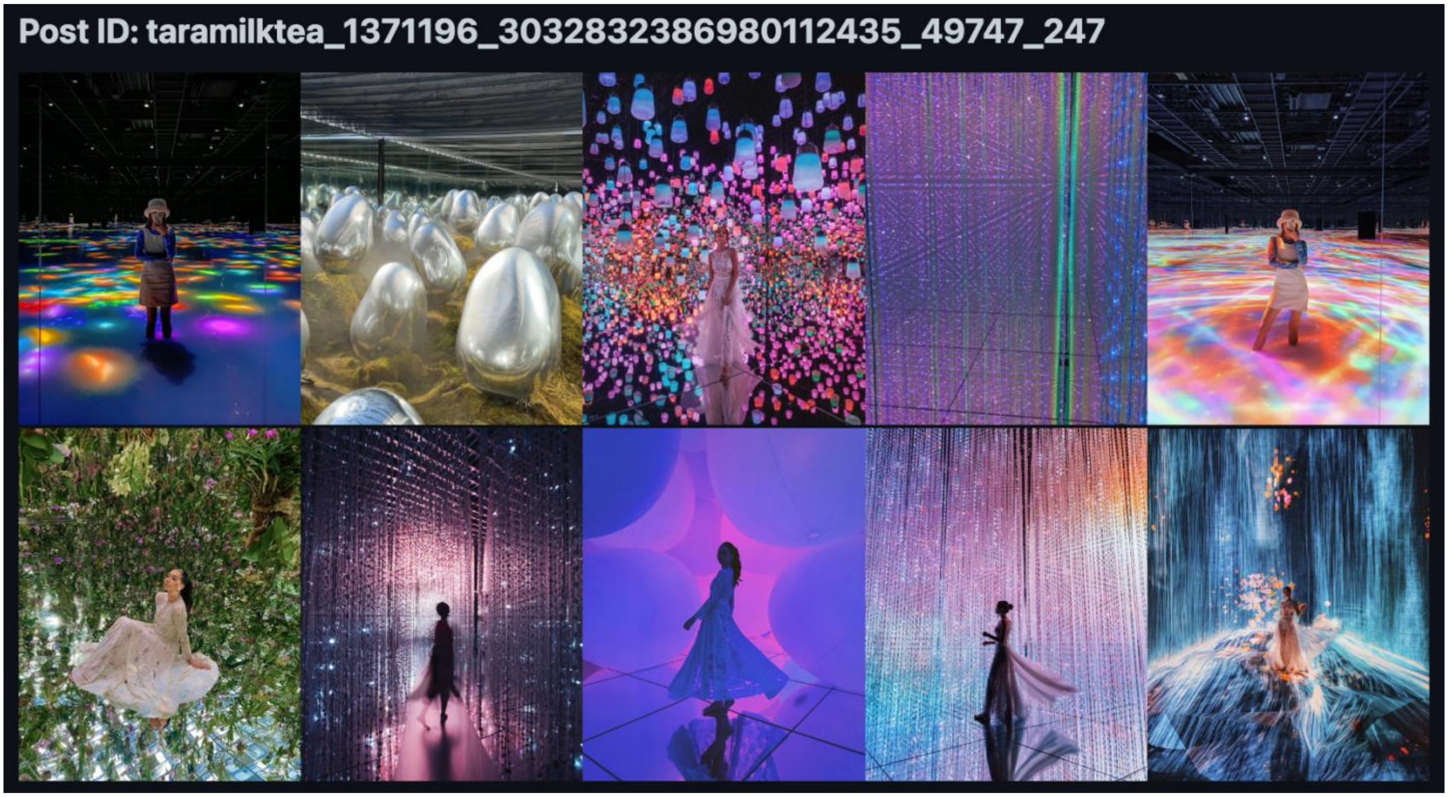
b. Business Meaning: These posts depict scenes of relaxation and leisure activities, promoting tranquility and well-being. This strategy aims to promote leisure products or experiences, wellness retreats, and lifestyle offerings associated with relaxation and rejuvenation, appealing to consumers seeking relaxation and leisure opportunities.

c. Sample Post



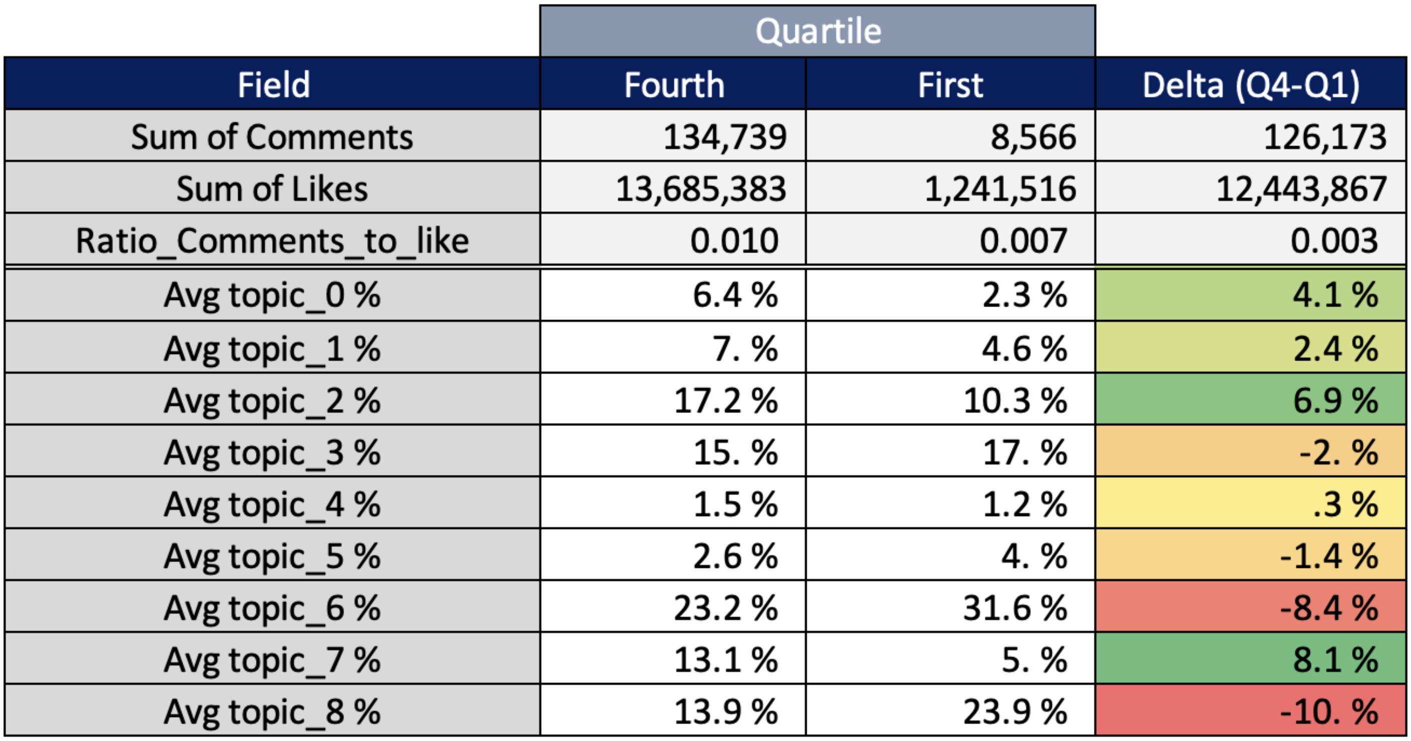
Topic 9: *Artistic Expression:*

- *a.* *Keywords: Photography, Joint, Eyewear, Knee, Leg, Gesture, Shoe, Fashion, Textile, Vision, Standing, Entertainment, Hairstyle, Accessory, Dress, Eyelash, Style, Fashionable, Artistic, Creativity, Expression, Aesthetic*
- ***b.*** *Business Meaning: This topic revolves around artistic expression and creativity, showcasing artistic fashion designs and visually appealing compositions. It signifies a strategy targeting artistic and creative individuals, fostering brand appreciation and engagement among aesthetically inclined audiences interested in fashion and artistic expression.*
- *c.* *Sample Post*



- **Next,** we added weight of each topic for each image, which was then consolidated to each post! This was done because a post is assumed to be having a combined influence. Images within a post don’t have independent influence for the common comment engagement.

Main differences in the average topic weights of images across the two quartiles



	Quartile		
Field	Fourth	First	Delta (Q4-Q1)
Sum of Comments	134,739	8,566	126,173
Sum of Likes	13,685,383	1,241,516	12,443,867
Ratio_Comments_to_like	0.010	0.007	0.003
Avg topic_0 %	6.4 %	2.3 %	4.1 %
Avg topic_1 %	7. %	4.6 %	2.4 %
Avg topic_2 %	17.2 %	10.3 %	6.9 %
Avg topic_3 %	15. %	17. %	-2. %
Avg topic_4 %	1.5 %	1.2 %	.3 %
Avg topic_5 %	2.6 %	4. %	-1.4 %
Avg topic_6 %	23.2 %	31.6 %	-8.4 %
Avg topic_7 %	13.1 %	5. %	8.1 %
Avg topic_8 %	13.9 %	23.9 %	-10. %

- Engagement Disparity: The engagement disparity between the fourth quartile (Q4) and the first quartile (Q1) is stark. Specifically, Q4 accumulated 126,173 more comments than Q1, showcasing a notable difference in audience interaction and engagement levels between these quartiles.
- Topic Distribution Differences:
  - Topic 0 (Gastronomic Delights): Posts related to gastronomic delights in Q4 had an average percentage of 6.41%, while in Q1, it was 2.31%. This indicates a 4.10% higher engagement with food-related content in Q4, highlighting a strong audience interest in culinary creations and food-related posts.
  - Topic 1 (Travel Adventures): Travel-related content in Q4 garnered an average percentage of 7.00%, compared to 4.57% in Q1, representing a 2.43% higher engagement. This underscores a significant interest in travel experiences and adventure among the audience.
  - Topic 2 (Fashion Elegance): Fashion-related content in Q4 had an average percentage of 17.22%, while in Q1, it was 10.34%. This shows a 6.88% higher engagement with fashion trends and elegant attire in Q4.
  - Topic 3 (Joyful Moments): Although Q1 had a slightly higher average percentage (17.03%) compared to Q4 (15.04%) for joyful moments, the difference is minimal. This indicates relatively consistent engagement with content focusing on happiness and joy across both quartiles.
  - Topic 4 (Entertainment and Leisure): Q4 posts had a slightly higher average percentage of 2.65% for entertainment and leisure compared to Q1's 4.04%, indicating a 1.39% difference. This suggests that entertainment-related content also contributes to higher engagement levels, albeit slightly more in Q1.

- Topic 5 (Urban Lifestyle): Urban lifestyle content in Q4 saw an average percentage of 23.17%, while in Q1, it was 31.61%, showcasing an 8.44% difference. This implies a stronger interest in urban lifestyle and city-related content among Q1 audience.
- Topic 6 (Elegant Fashion Trends): Fashion-related content in Q1 had a significantly higher average percentage (23.90%) compared to Q4's 13.08%, reflecting a 10.82% difference. This suggests that stylish and elegant fashion trends resonate more with the audience in Q1.
- Topic 7 (Relaxation and Leisure): Posts related to relaxation and leisure activities in Q4 had an average percentage of 13.91%, while in Q1, it was 4.99%. This indicates an 8.09% difference, showcasing a relatively lower engagement with content related to relaxation and leisure activities in Q4.
- Topic 8 (Artistic Expression): Q1 posts showed a higher average percentage (9.99%) for artistic expression compared to Q4 (8.09%), representing a 1.90% difference. This suggests a higher engagement with artistic and creative content in Q1.

Task B: advice for the client to increase engagement on its Instagram page based on findings

Recommendations are based on performance analyses and topic modelling of influencers who monetize engagement primarily on the type of Instagram posts they are making rather than their external accolades. For this purpose. Based on this we have the following tailored recommendations to boost influence and captivate the audience:

1. Capture Gastronomic Delights and Travel Adventures:

- Gastronomic delights and travel adventures garner higher engagement in Q4. Showcase visually appealing culinary creations like artisanal desserts (e.g., a decadent chocolate lava cake) or exotic dishes (e.g., sushi rolls bursting with vibrant colors), coupled with travel adventures featuring breathtaking landscapes (e.g., a sunset over a tropical beach) and cultural experiences (e.g., exploring bustling markets in foreign cities).

2. Focus on Fashion Elegance and Urban Lifestyle:

- Fashion elegance sees a remarkable increase in engagement in Q4, while urban lifestyle content remains a strong favorite among the audience in Q1. Share stylish outfits against urban backdrops, offering fashion-forward tips that resonate with followers' sense of elegance and city living. For instance, showcase a chic ensemble featuring a tailored blazer and statement accessories against the backdrop of a vibrant city street.

3. Emphasize Joyful Moments and Artistic Expression:

- Although joyful moments maintain consistent engagement levels across both quartiles, Q1 shows a slight preference for artistic expression. Share heartwarming moments like a family picnic in the park or a spontaneous dance party with friends. Unleash creativity through artistic endeavors like photography (e.g., capturing the play of light and shadow in a serene landscape) or painting (e.g., creating a vibrant abstract artwork inspired by nature).

4. Integrate Entertainment and Leisure:

- Despite a slight dip in engagement in Q4, entertainment and leisure content captivates the audience's interest. Incorporate entertaining content like movie reviews (e.g., a review of the latest blockbuster film) and hobby-related posts (e.g., a DIY tutorial for crafting handmade candles). Offer behind-the-scenes glimpses into leisure activities such as a cozy night in with a favorite book or a fun-filled day exploring a local amusement park.

5. Balance Elegant Fashion Trends with Relaxation and Leisure:

- The audience shows higher engagement with elegant fashion trends in Q1, while relaxation and leisure activities see an increase in engagement in Q4. Strike a harmonious balance by sharing stylish fashion inspiration (e.g., showcasing a classic little black dress paired with statement jewelry) alongside moments of relaxation and self-care (e.g., a tranquil yoga session at sunrise or a soothing bubble bath with scented candles).

Aligning the content strategy with these insights will deepen the connection with the audience and foster sustained engagement and growth on the Instagram platform. Keep an eye on audience feedback and adapt the strategy accordingly to ensure continued success on the influencer journey.