Case #1 - Understanding Consumer Trends with images

Business case

In today's world, consumer preferences shift often as they are influenced by the growth in the use of social media. There has never been another moment in history in which information has been shared with so much ease.

Companies struggled with understanding their customers and developing products which are meaningful to their consumers. Understanding what customers talk about and what are the latest trends is by far one of the most important aspects of any business strategies.

The case focuses on the beauty industry, as its trends have a tendency to be very dynamic due to them being driven heavily by social media. Due to the fact that communications have become more and more visual, text is not enough to understand the world around us. This is why the case focuses heavily on extracting.

The main question of the case is to understand which trends the company should focus on in order to successfully tailor products to the correct customers.

Instructions

The case was created to show off your creativity and analytical thinking in a real world situation.

The analysis should focus on the following elements:

1. Discovering insights at a global level
2. Discovering insights at hairstyle level
3. Explore the relationships in between the hashtags used in the posts
4. Additionally, what other types of analysis you would do?

Detail the steps you have taken in your analysis and be as creative as you want because there are no right or wrong answers.

Description of the data

The file, ***Dashmote_data_set_case.json*** contains a sample 10000 images which have been run through visual recognition engine and clustered based on the hairstyle they showcase.

The variable, cluster represents the hairstyle cluster that the image has been assigned to by the visual recognition algorithm.

Each row contains the variable **url** which can serve as an identifier and the number of **likes** together with the **comments** per image. The **user_id** is the unique id of the Instagram account from which the post comes and the variable **id** is the unique identifier associated with the post itself.

Additionally a classifier **influencer_flag** was added to each of the images which have more than 500 likes, flagging them as influencer posts.

Not all variables need to be used and you will prioritize them based on the need of the type of analysis you will be using.

Schema of the Data

cluster	STRING	NULLABLE
id	STRING	NULLABLE
influencer_flag	STRING	NULLABLE
user_id	STRING	NULLABLE
influencer_num	INTEGER	NULLABLE
url	STRING	NULLABLE
date_month	STRING	NULLABLE
hashtags	STRING	REPEATED
date_unix	INTEGER	NULLABLE
comments	INTEGER	NULLABLE
date_week	STRING	NULLABLE
likes	INTEGER	NULLABLE
date_formated	STRING	NULLABLE
inf_true	STRING	NULLABLE