**DSA210 PROJECT | Luxury Makeup Brand Analysis Final Report**

**ZEYNEP ÖZİLER**

**30986**

TABLE OF CONTENTS

[1. INTRODUCTION 3](#_Toc199451238)

[2. STEP BY STEP METHODOLOGY 4](#_Toc199451239)

[a. Data Collection 4](#_Toc199451240)

[b. AI Assisted Engineering 4](#_Toc199451241)

[c. Data Expansion(AI Assisted) 5](#_Toc199451242)

[d. Exploratory Data Analysis(EDA) 5](#_Toc199451243)

[e. Hypothesis Testing 6](#_Toc199451244)

[f. Machine Learning Modelling 7](#_Toc199451245)

[g. Tools & Documentation 7](#_Toc199451246)

[3. CONCLUSION 7](#_Toc199451247)

# **INTRODUCTION**

The aim of this project is to determine if engaging with TikTok impacts the perceived sales growth of luxury makeup brands. YSL Beauty and Lancôme were chosen as the subject brands of the project considering their popularity and online visibility. This project comprised an entire data science structure which was performed in Python, including but not limited to: data collecting, feature construction, data preprocessing, hypothesis testing, and machine learning applications.

# **STEP BY STEP METHODOLOGY**

## **a. Data Collection**

- 32 TikTok videos (16 per brand) were selected by hand from the official TikTok accounts of @yslbeauty and @lancome.official.

- The data was collected through the months of January to April 2025, having 4 videos per month per brand.

- Each row included the following columns for the data: Date, Brand, Views, Likes, Comments, Caption, Video Type, and Why Selected.

Data collection was done manually due to the platform’s scraping limitations and was curated carefully to ensure a diversity of contents and engagement levels.

## **b. AI Assisted Engineering**

To deepen the analysis, two additional columns were added using AI:

1- *Sentiment Analysis (column name: SENTIMENT)*

\* Caption analysis was made using a rule-based sentiment classifier which was created in Python.

\* A manually constructed dictionary of keywords (e.g., “love”, “great”, “hate”, “worst”) was used to highlight and detect the sentiment of each post as Positive, Neutral, or Negative.

2- *Growth Rate Estimation (GROWTH RATE %)*

\* Growth estimation was done using a formula that’s been created:

Engagement Score = (Likes + Comments) / Views

\* Engagement score was adjusted with a sentiment multiplier:

+15% for Positive

+5% for Neutral

–5% for Negative

\* The resulting value was the predicted growth rate for each post.

## **c. Data Expansion(AI Assisted)**

To come closer to the necessities to be able to applying machine learning algorithms, the dataset was synthetically extended to 64 rows.

\* This was done by AI-assisted row generation, sustaining statistical similarity with the original data distribution.

\* Generated rows have a similar structure to the patterns that are observed in already collected post statistics while preserving logical combinations of features like Views, Likes, Sentiment, and Growth Rate.

## **d. Exploratory Data Analysis(EDA)**

Using pandas, matplotlib, and seaborn, several visualizations were conducted:

- Average Growth Rate per Brand – barplot

- Sentiment Distribution – countplot

- Likes vs Growth Rate – scatterplot

- Video Type vs Growth – categorical analysis

These graphs are stored under the visuals/ folder and support the visualization of the related trends that are in the dataset

## **e. Hypothesis Testing**

To validate our observations with statistical proof, two major hypothesis tests were done by using the scipy.stats library in Python.

1- Independent Two-Sample T-Test:

- Null Hypothesis (H\_0): No significant difference between the mean estimated growth rates of YSL and Lancôme contents.

- Alternative Hypothesis (H\_1): There is a significant difference in mean growth rates between the two brands.

- Results: The p-value was 0.8978, indicating *no statistically significant difference* between the growth rates of the two brands.

2- Pearson Correlation Test:

- Null Hypothesis (H\_0): No linear correlation between likes and estimated growth.

- Alternative Hypothesis (H\_1): Existing correlation between likes and estimated growth.

- Results: The Pearson correlation coefficient (r) = 0.366, with a p-value = 0.043.

- Interpretation: *Moderate positive correlation* between likes and growth rate, which is statistically significant suggesting that videos that are liked more more likely to correspond a higher growth rate.

Related python codes for these tests are located in code/analysis\_with\_hypothesis.py.

## **f. Machine Learning Modelling**

After extending the dataset to 64 rows, ML applications were done to predict Growth Rate Category (High vs Low):

- Growth rate column was converted into categorical (binary)

- *Two models* were tested:

Decision Tree Classifier

Logistic Regression

- Model evaluation metrics included: Accuracy, Precision, Recall, F1 Score

All code is available in: code/final\_analysis\_model.py

## **g. Tools & Documentation**

- The project is published on GitHub in the following link: https://github.com/zeynepoziler/luxury\_makeup\_brands\_analysis

- Repository structure includes: data/ for all datasets, visuals/ for figures, code/ for all Python code files

- README.md explains the methodology and findings detailly through the project.

- requirements.txt lists all required Python libraries.

# **CONCLUSION**

This project is a demonstration on how social media content can be used alongside AI-assisted tools and machine learning techniques to derive insights in the luxury cosmetics marketing sector. Despite having multiple challenges in data collection process, the combination of manual and synthetic data collection allowed for reasonable statistical and predictive analysis.