Facebook Metrics Lab

Oluwamayowa Adesoye

THK Capstone 2023

TABLE OF CONTENTS

Problem Statement

In the realm of business analytics, there exists a compelling challenge centered around the performance metrics of a distinguished cosmetic brand's Facebook page. The dataset under consideration is multivariate and encompasses integer-based features, offering a comprehensive view of posts published during the year 2014 on the aforementioned brand's Facebook page. With 500 instances and a focus on regression tasks, the dataset, a subset of the original study by Moro et al. (2016), presents a valuable opportunity for analysis and insights. However, the omission of certain features due to confidentiality concerns raises the need for a rigorous exploration of the available data.

This project aims to leverage this dataset to derive meaningful regression models that can uncover patterns, trends, and influential factors affecting the Facebook page performance of the renowned cosmetics brand, ultimately contributing to a deeper understanding of social media dynamics in the business context.

Data Wrangling

The dataset, sourced from UC Irvine, comprises 500 rows and 19 columns, encapsulating various attributes related to the Facebook performance metrics of a renowned cosmetics brand. The columns encompass key dimensions such as 'Page total likes,' 'Type,' 'Category,' 'Post Month,' 'Post Weekday,' 'Post Hour,' 'Paid,' and several metrics related to lifetime post reach, engagement, and interactions. The initial dataset exhibited a structured format with a shape of (500, 19). Although largely clean, a discerning observation revealed the presence of a few missing values. In pursuit of data quality, I conducted a thorough exploration to understand the origins of these missing values. Subsequently, to ensure the robustness of the dataset, I made an informed decision to remove the instances with missing values. This meticulous approach to data wrangling lays the groundwork for reliable analyses, allowing for meaningful insights into the performance dynamics of the cosmetics brand's Facebook page during the specified timeframe.

Exploratory Data Analysis

Exploratory data analysis is the first and most important phase in any data analysis. In EDA, I conducted: Data sourcing, Data cleaning, Univariate analysis and Bi-variate/Multivariate analysis.In data cleaning, I checklisted for row corrections and columns corrections, checklisted for missing values corrections, checklisted for standardizing values corrections, checklisted for invalid values corrections. When checklisting for rows and collections, I checked for incorrect rows and columns because if too much information is missing from them, I could make the choice of deleting any incorrect rows or columns.

When checking for standardizing values corrections, I took a look at the outliers through the univariate graphs and confirmed it as well on the bivariate graphs. Univariate analysis is the simplest form of quantitative data analysis. It's used to describe, summarize, and find patterns in the data from a single variable. Histograms is a univariate graph which is used in my analysis. Histogramming counts how many samples fall in a certain interval. I choose to keep the outliers in my data.

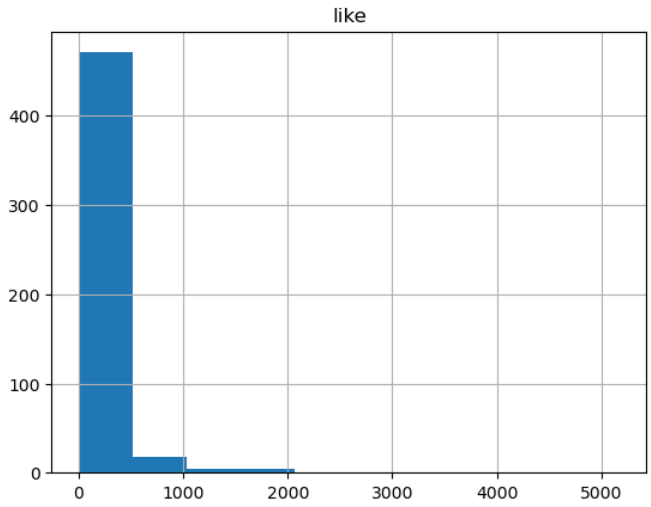

Figure 1: Histogram of 'like' data

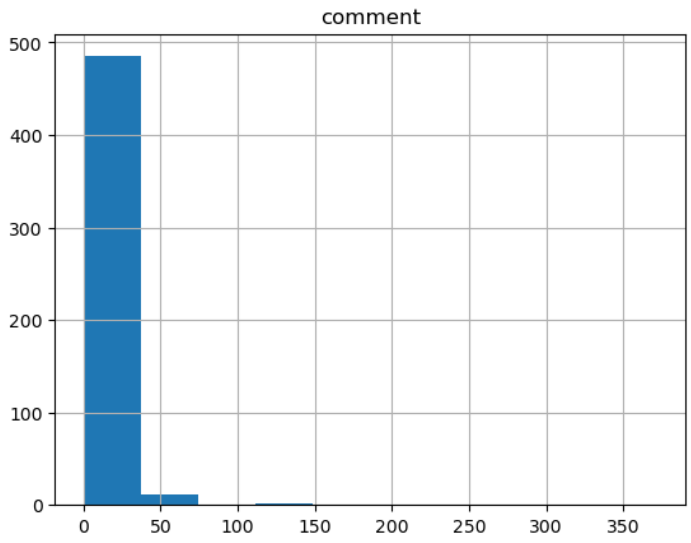

Figure 2: Histogram of 'comment' data

Box plots were used to source out properly look at the outliers in the data set. Some had outliers and some did not have outliers.

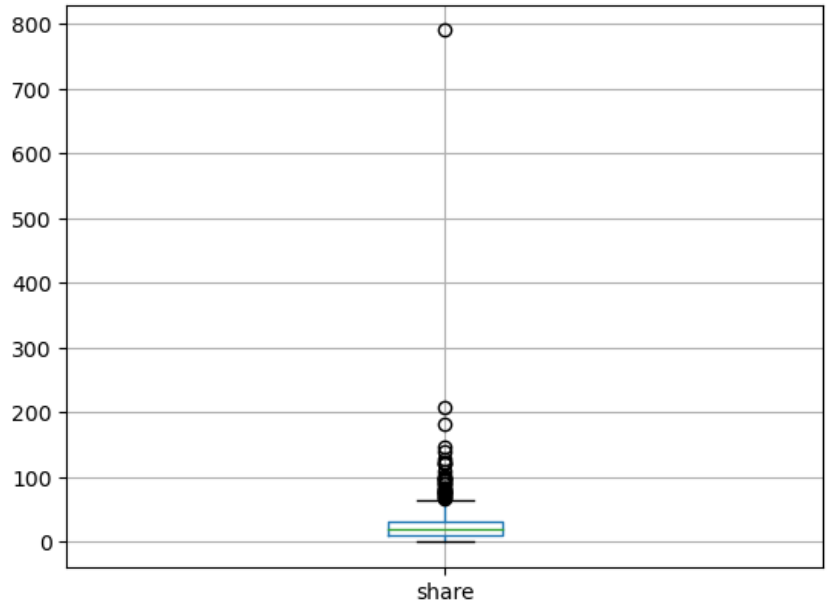

Figure 3*:* Outlier in the "Share" data.

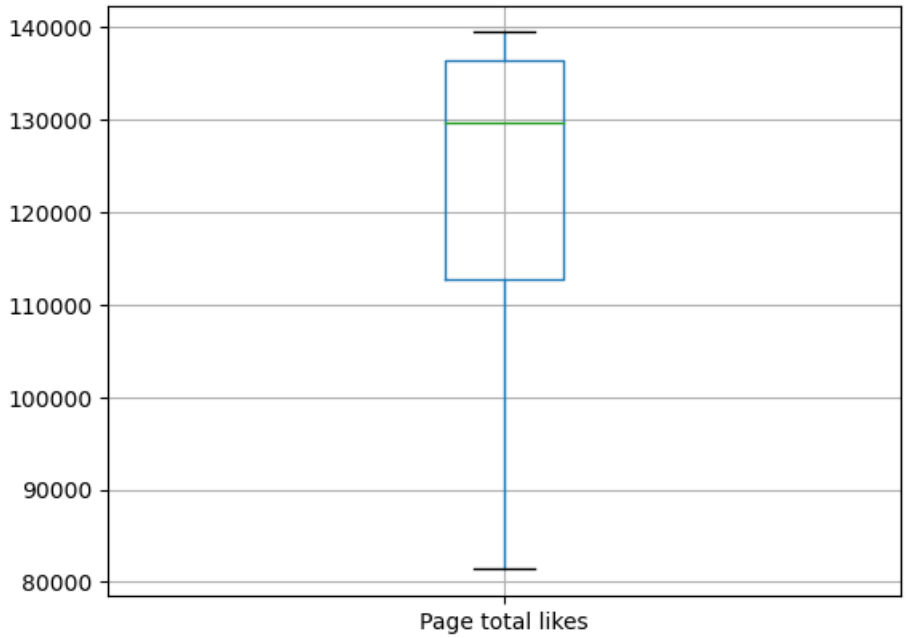

Figure 4: No outliers in the 'Page total likes" data.

When analyzing bivariate cases, I used scatter plots to describe relationships between pairs of variables, to assess linearity, find linearizing transformations and to also detect Outliers. I did a comparison between "like" and "shares." I want to see if an increase in shares leads to an increase in likes. I want to see if a correlation lies between them.

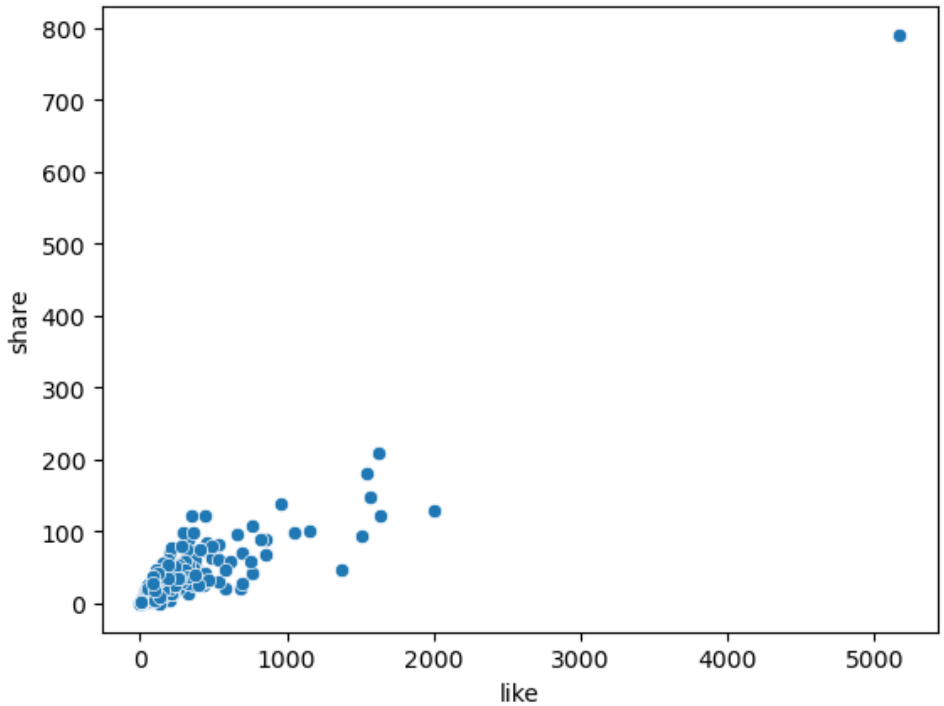

***Figure 5*:** Scatterplot between like and shares

I hypothesize that there is a correlation between likes and shares in that an increase in shares, increases the amount of likes it gets. My hypothesis was proven right by the multivariate graph used to compare my data. The null hypothesis was no correlation between shares and likes in a post.
There's a positive correlation between likes and total interactions. There's no correlation between likes and post hours. There's a positive correlation between like and share.

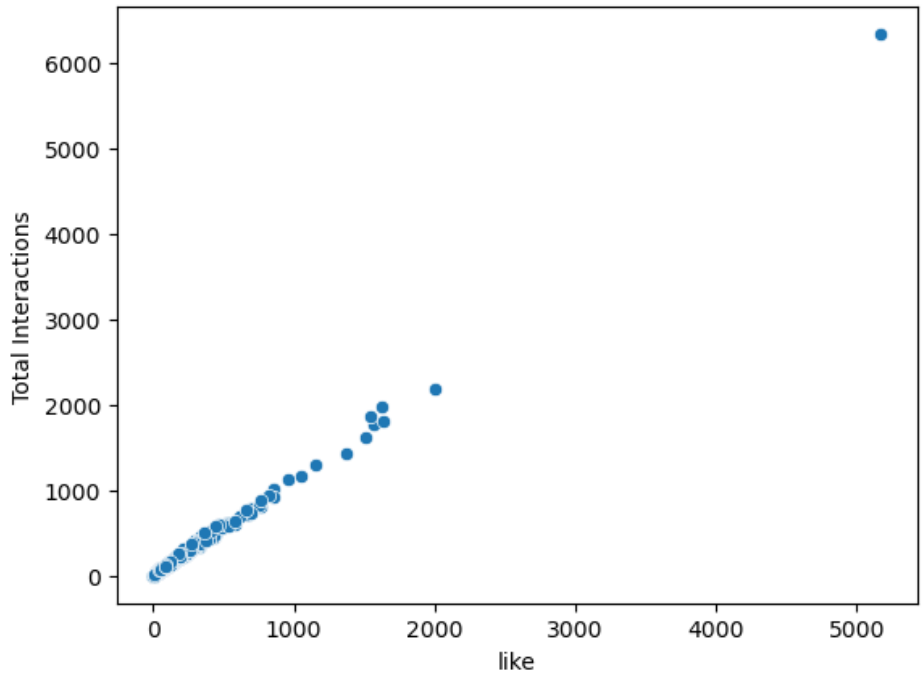

Figure 6: Scatterplot between like and total interactions

There's no correlation between page total likes and lifetime post impressions by people who have liked your page.

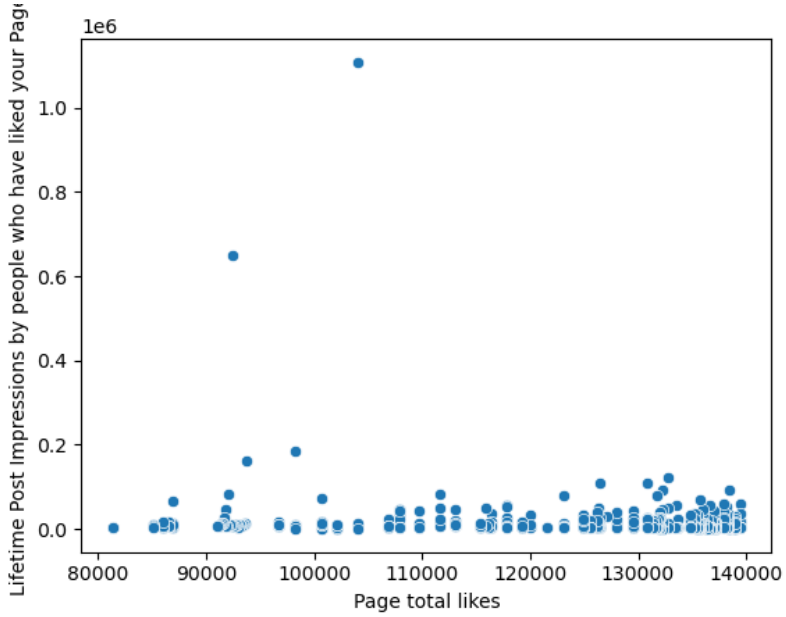

Figure 7: Scatterplot between page total likes and lifetime post impressions by people who have liked your page.

As stated a few graphs had outliers such as share total interactions, lifetime, post impressions by people who have liked your page, like, and comment. I am choosing to leave these outliers in our data, because there could be a reason why it's there, such as: a pregnancy announcement post, a birthday post, a graduation post, a marriage engagement post, and much more that will cause more of these to flare up.

Modeling

In our data modeling phase, we employed linear regression, a statistical method utilized to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. This method is particularly valuable for predicting a continuous dependent variable, also known as the target variable, based on several independent variables, referred to as predictive variables. In my analysis, the target variable identified was "like," and the objective was to assess its relationship with other independent variables. The procedural steps involved training the data, scaling it, and subsequently performing linear regression.

The regression line generated from this analysis can be employed to predict the value of Y for a given value of X. Notably, the actual versus predictive values plot revealed a positive slope, indicating a robust correlation between the model's predictions and the actual results. However, when examining the residual plots, a contrasting negative or declining slope was observed. This finding suggests that the model may not adequately capture the underlying linear relationship in

the data. Further investigation is warranted to enhance the model's performance and ensure a more accurate representation of the intricate dynamics among the variables.

Results and Discussion

In conclusion, this project embarked on a comprehensive exploration of the performance metrics of a distinguished cosmetic brand's Facebook page, using a dataset comprising 500 instances and 19 columns sourced from UC Irvine. The data wrangling phase ensured the dataset's cleanliness and reliability by addressing missing values, setting the stage for a robust analysis. The exploratory data analysis (EDA) phase involved thorough checks for data accuracy, standardization, and outlier considerations. Notably, the investigation into the correlation between "likes" and "shares" revealed a positive relationship, validating the initial hypothesis.

Moving to the data modeling phase, linear regression was employed to predict "likes" based on other independent variables. While the model demonstrated a positive correlation in the actual versus predicted values plot, the residual plots suggested potential limitations in capturing the underlying relationships. Further refinement and investigation are essential to enhance the model's accuracy and ensure a more nuanced representation of the intricate dynamics within the dataset. This project has laid a foundation for deeper insights into the factors influencing the Facebook page performance of the cosmetic brand, contributing to a more profound understanding of social media dynamics in the business context. Future iterations may involve fine-tuning the model and exploring additional features to enhance the predictive capabilities and practical applications of the analysis.

In the ongoing pursuit of refining the analysis of the cosmetic brand's Facebook page performance, several future steps could be undertaken to enhance the model's robustness and predictive capabilities. A crucial avenue for improvement involves a meticulous review of the model specification to ensure its correctness and inclusion of all relevant predictors. Exploring potential interactions or introducing nonlinear terms may further capture nuanced relationships within the data. Additionally, considering variable transformations, such as log transformations or power transformations, could be explored to better align the variables with the underlying relationships. Another avenue worth exploring is the evaluation of alternative models. If the observed relationship proves to be nonlinear, contemplating the use of nonlinear regression models or other machine learning algorithms capable of capturing complex patterns may offer valuable insights and contribute to a more accurate representation of the dynamic interplay among variables. These future steps aim to refine and optimize the modeling process, ultimately leading to a more nuanced understanding of the factors influencing the Facebook page performance of the cosmetic brand.