# SOME PERFORMANCE ASPECTS CONSIDERATIONS OF DATA ANALYSES On Instagram Insights

### PROJECT REPORT

**Submitted by**

**Harsh Vardhan Deepak Mittal  
Nikunj Goel   
Nisha Tyagi  
Vansh Kohli**

**BACHELOR OF COMPUTER APPLICATION**

 **SCHOOL OF COMPUTER SCIENCE ENGINEERING AND TECHNOLOGY**

BENNETT UNIVERSITY

Greater Noida-201310, Uttar Pradesh, India

01-December 2023/ 2Year

**TABLE OF CONTENTS**

1. Introduction…………………………………………………………………………….1
2. Objective overview …………………………………………………………………….1
3. [Data Exploration ……………………………………………………………….……...2](#_bookmark0)

3.1 Summarize Dataset Characteristics…………………………………………….......04

3.2 Explore and Visualizing Key Statistics………………………………...………….05.

1. [Cleaning And preprocessing …………………………………………...](#_bookmark0)........................04

4.1 Handling Missing Value……………………………………………………………04

4.2 Handling Duplicates……………………………………………………………….04

1. [Features Engineering …………………………………………...](#_bookmark0)...................................05

5.1 Create new meaningful features if necessary……………………………………....05

5.2 Handling categorical……………………………………………………………….05

5.3 Scaling Numerical Features………………………………………………………...05

5.4 Date and Time Handling……………………………………………………………05

1. Explanatory Data Analysis (EDA) […………………...](#_bookmark0)..................................................06

6.1 Data Distribution Analysis…………………………………………………………07

6.2 Aspects Relationship Analysis………………………….………………………….10

6.3Calculation of correlation coefficient……………………………………………...11

7. Statistical ID Analysis ………………...…………………….........................................12

8.Hypothesis Formulation ……………………………………..........................................13

9. Machine Learning (Prediction)……………………...………………………………….14

10. Conclusion…………………………………………………………………………….16

Insta Insights...!!

Introduction: -

In the era of digital communication and social media, understanding user engagement and content performance is essential for effective strategies and decision-making. This analysis focuses on

Instagram data, a popular social media platform, to extract insights into user interactions and trends. By leveraging exploratory data analysis (EDA), statistical testing, and machine learning techniques, we aim to uncover patterns, correlations, and predictive models that can inform content creators and marketers.

Objective Overview:

The primary goal is to gain actionable insights into Instagram user engagement and-

Content Category Effectiveness:

Follows and Impressions Correlation:

Temporal Trends in User Interactions:

Machine Learning Models for Predictions:

Data-Driven Content Strategy:

Holistic Approach:

Continuous Improvement:

Key Performance Indicators (KPIs)

Cross-Platform Consideration:

Actionable Recommendations:

Review and Adaptation:

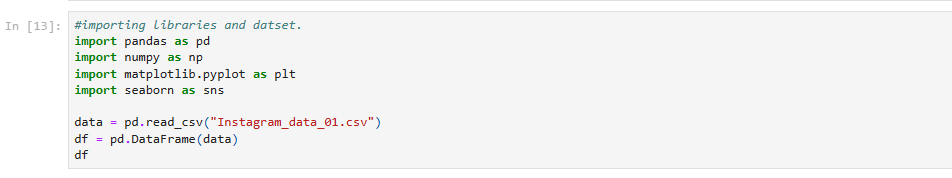
Schedule regular reviews to assess the effectiveness of implemented strategies.

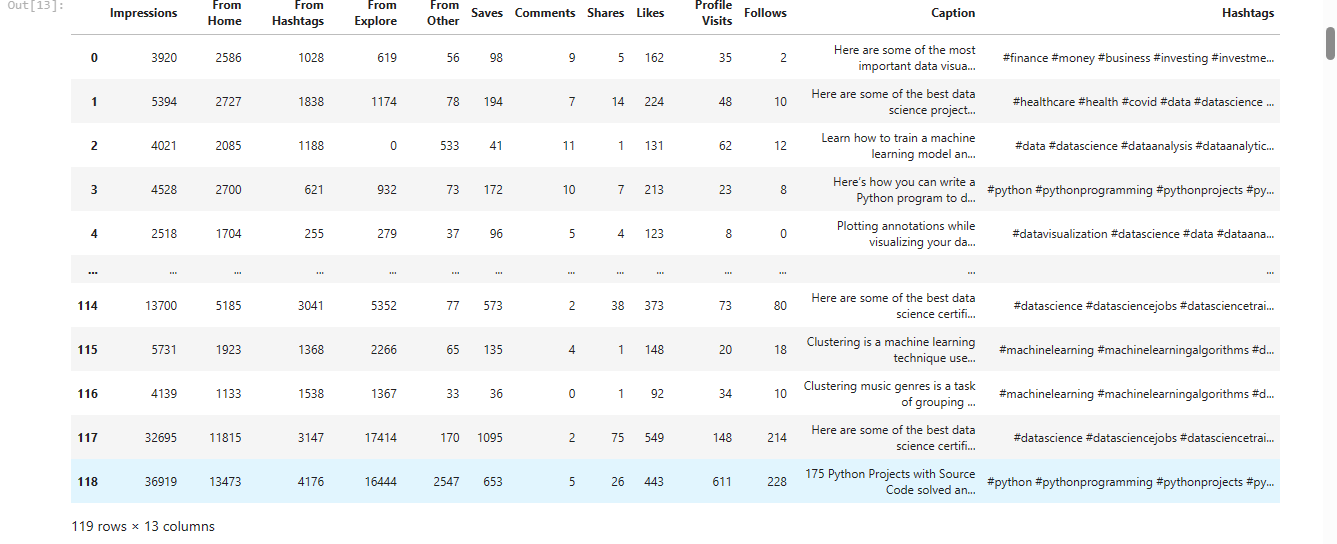
Be prepared to adapt and refine the approach based on ongoing analysis and changing platform dynamics.

1

Data Exploration: -

1. Summarize Dataset Characteristics:

Use the Pandas library to load the Instagram dataset. Display the basic information about the dataset, including the number of records, variables, and data types.

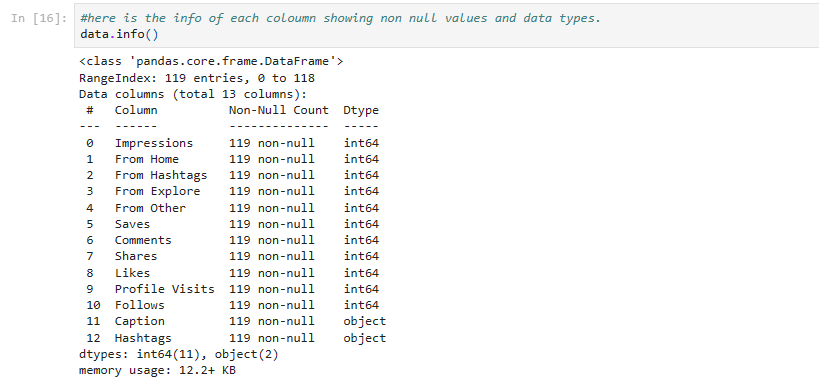


2. Explore and Visualize Key Statistics:

Utilize various visualization techniques to showcase key statistics for relevant numerical variables such as likes, comments, and shares and data info of the data set.

After converting our dataset into data frame, we used the info function to obtain the basic information about our data frame.

2



3

Cleaning and Preprocessing:

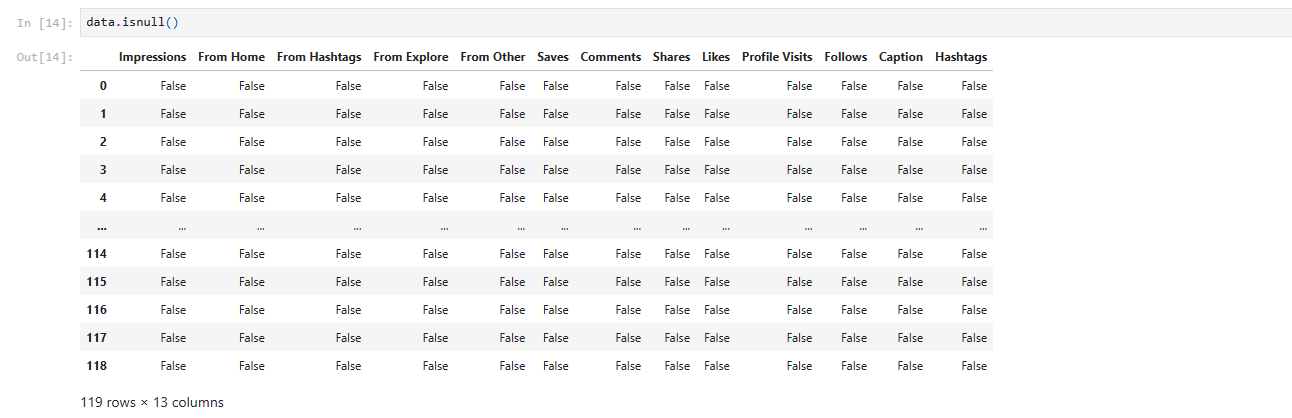
Data cleaning and preprocessing are crucial steps in the data analysis.

process, especially for sales analysis. Here are common steps taken to clean and preprocess sales data:

1.Handling Missing Values:

Check for missing values in the dataset.

Impute missing values using appropriate methods.

****

2. Handling Duplicates:

Identify and remove any duplicate records in the dataset.



4

Feature Engineering:

Create new meaningful features if necessary.

---Example-Engagement Rate

****

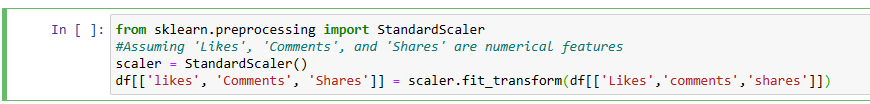
-Handling Categorical Variables:

Encode categorical variables if needed.



-Scaling Numerical Features:

Standardize or normalize numerical features.



-Date and Time Handling (if i applicable):

Convert date and time columns to datetime format.

A screenshot of a computer

Description automatically generated

5

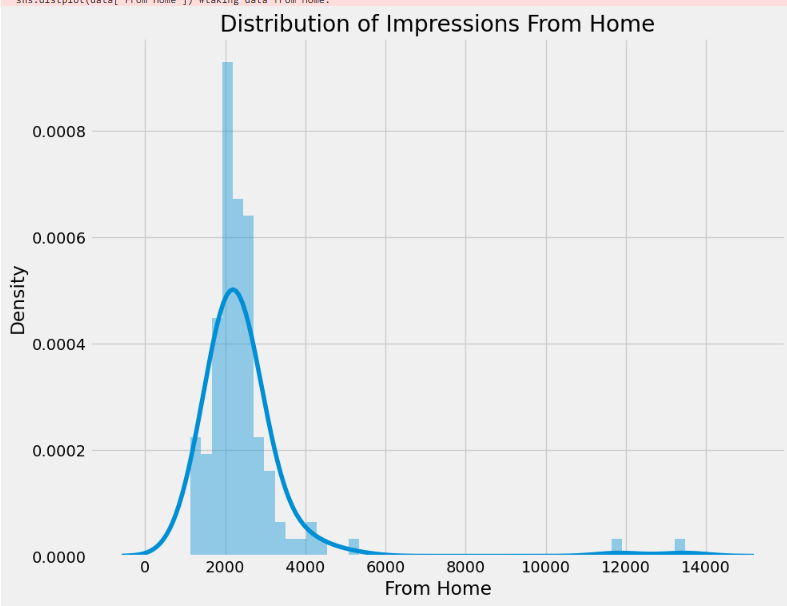
**Exploratory Data Analysis (Visualization):**

This section contains the visualization we applied on our dataset to compare explore and understand various aspects and dependencies of the variables and functions involved.

-Data Distribution Analysis

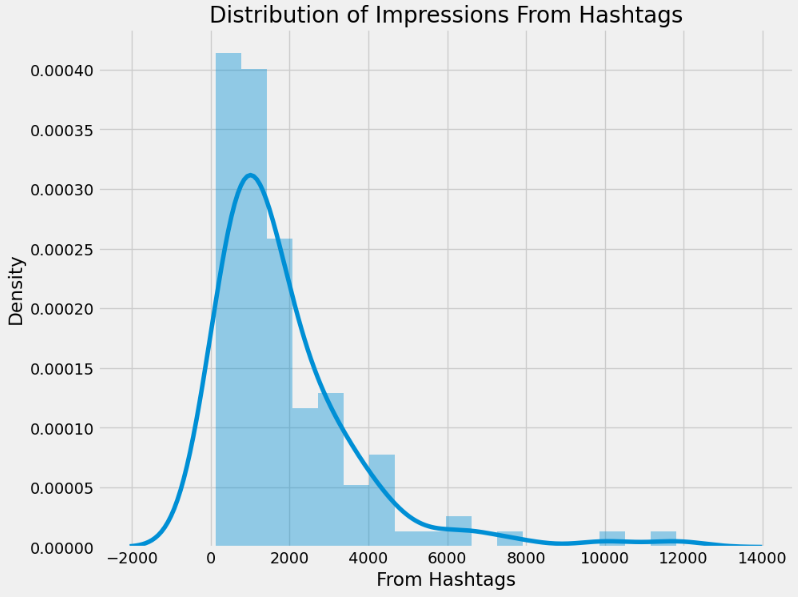
in the below section the data is analyzed, and we produced histogram graphs for impressions produced from various sources.



. 

The above graphs depict the impressions created from home page which helps us to evaluate the impact of home page on overall impressions.

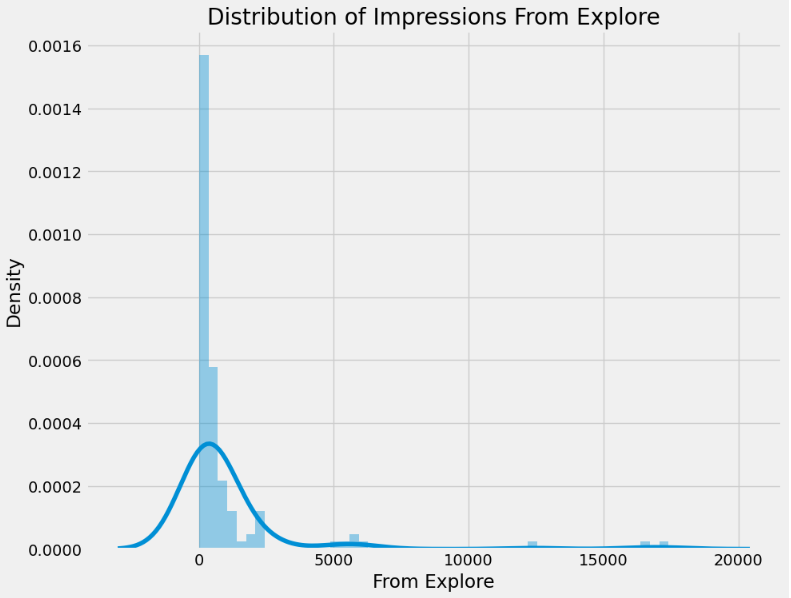
6



The above graphs depict the impressions created from hashtags which helps us to evaluate the impact of hashtags on overall impressions.



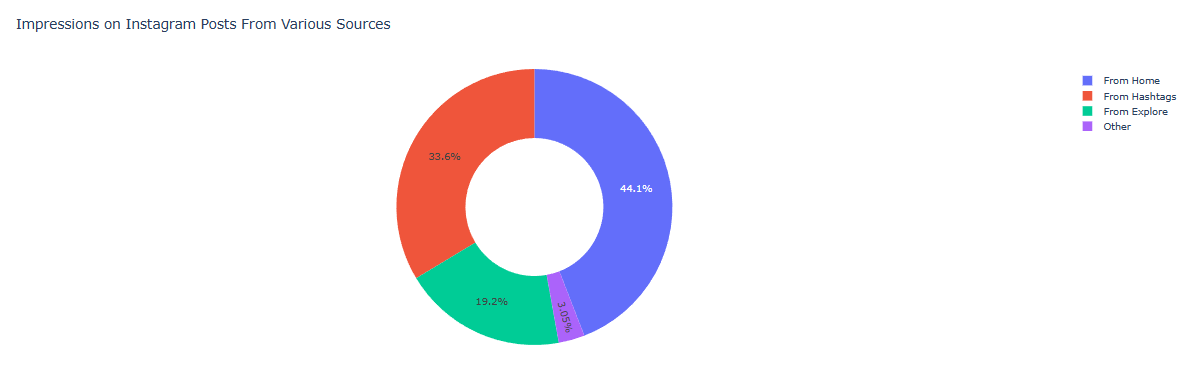
7



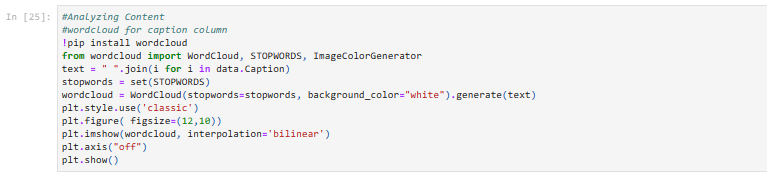
The above graphs depict the impressions created from explore which helps us to evaluate the impact of explore page on overall impressions.

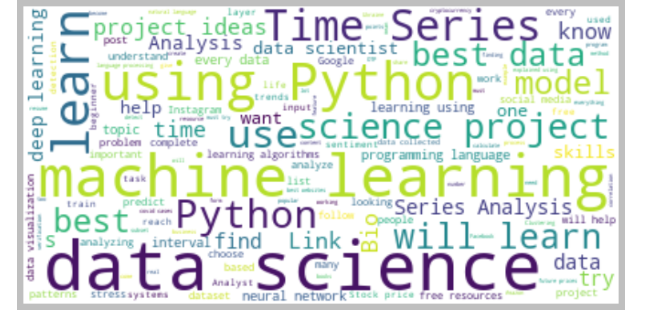
The below pie chart shows the distribution of impressions among the various sources on the Instagram which tells you where to focus on the id to gain views and increase followers.





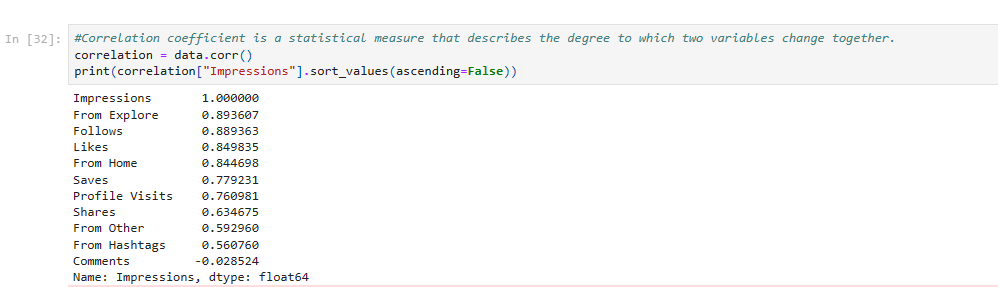
8





This visualization is called word cloud it highlights the focus point in a particular data set ,the more frequently used words appear larger according others are in size reference to the frequency.

Calculation of correlation coefficient



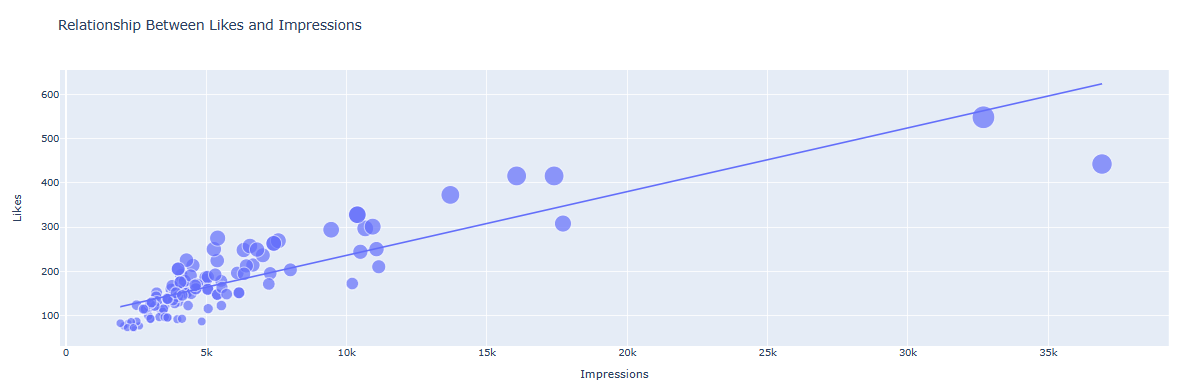
The correlation coefficient helps in determining the dependency of variables on each other here we calculated the coefficient with impressions and displayed them in a descending order the more the value is toward it represents positive relation and more toward -1 is negative here in this data set we can see that the impressions from explore are most effective and from comments are least effective.

9

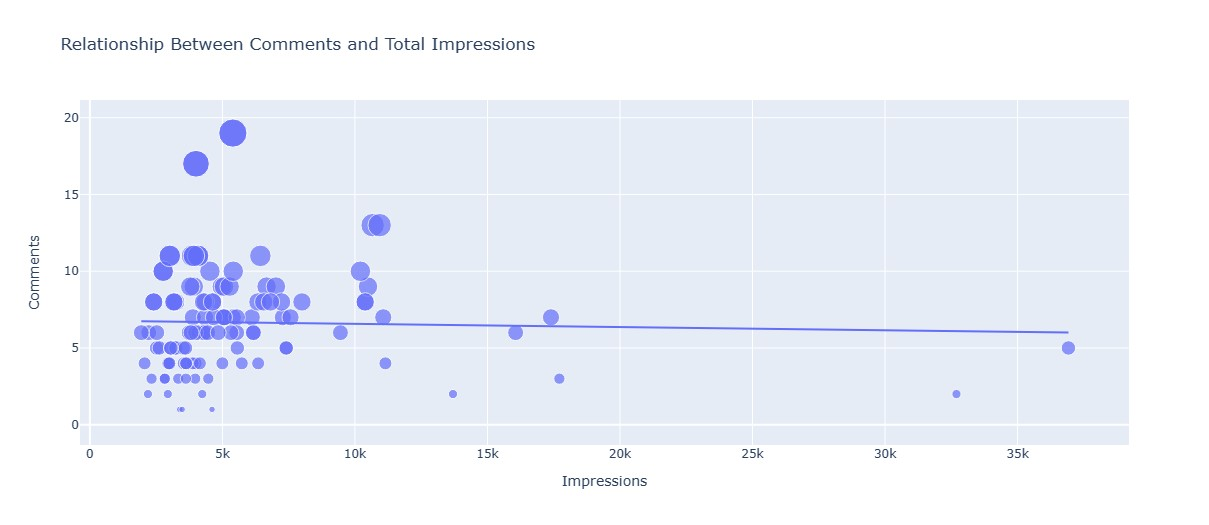
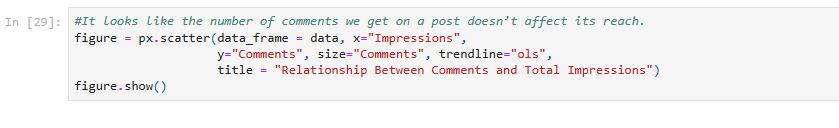
Aspects Relationship Analysis:

Analyze relationships between more than two variables.

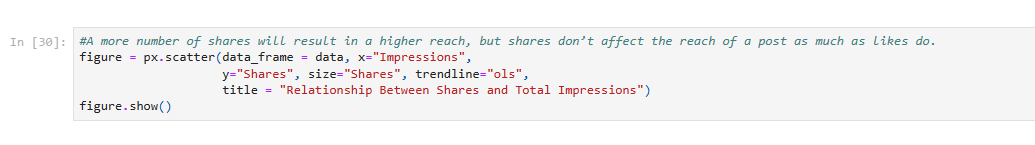




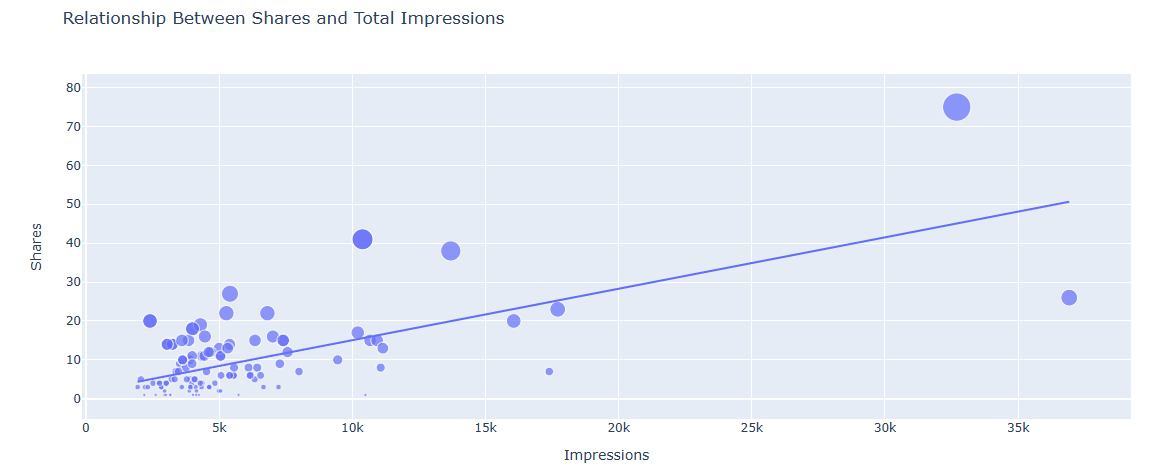
This graph produced a upward slope which represents positive dependency of the impressions due to likes, which concludes that more like results into more overall impressions



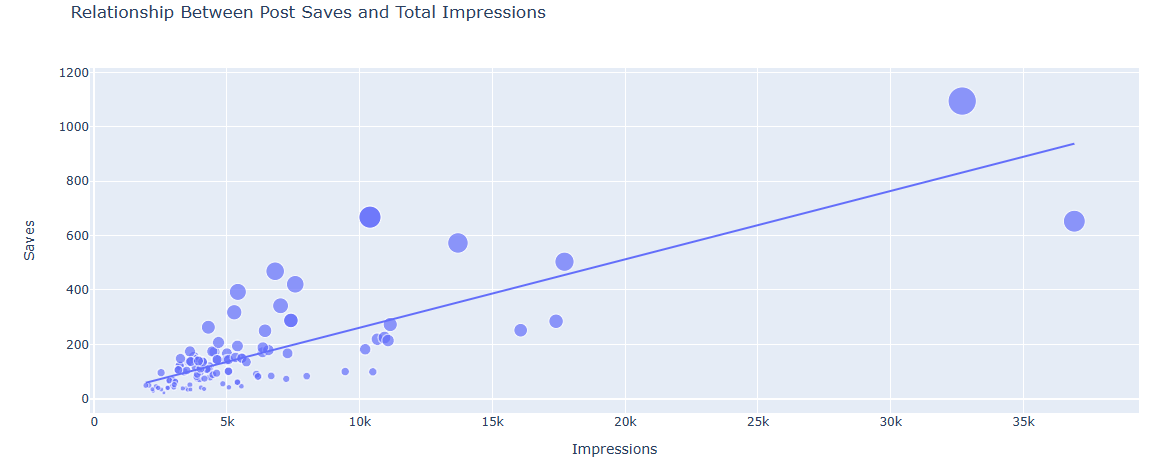
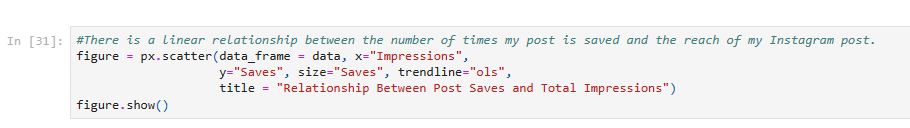
This graph produced almost no slope which represents positive dependency of the impressions due to comments, which concludes that comment results into less overall impressions.



10



This graph produced an upward slope which represents a positive dependency of the impressions due to shares, which concludes that more shares results into more overall impressions.

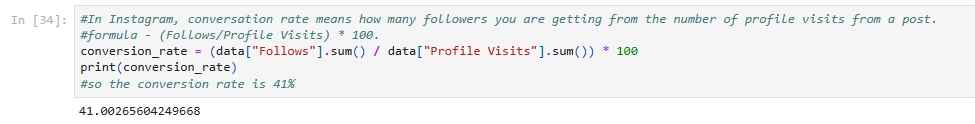


This graph produced a high upward slope which represents positive dependency of the impressions due to saves, which concludes that more like results into highly increased overall impressions.

11

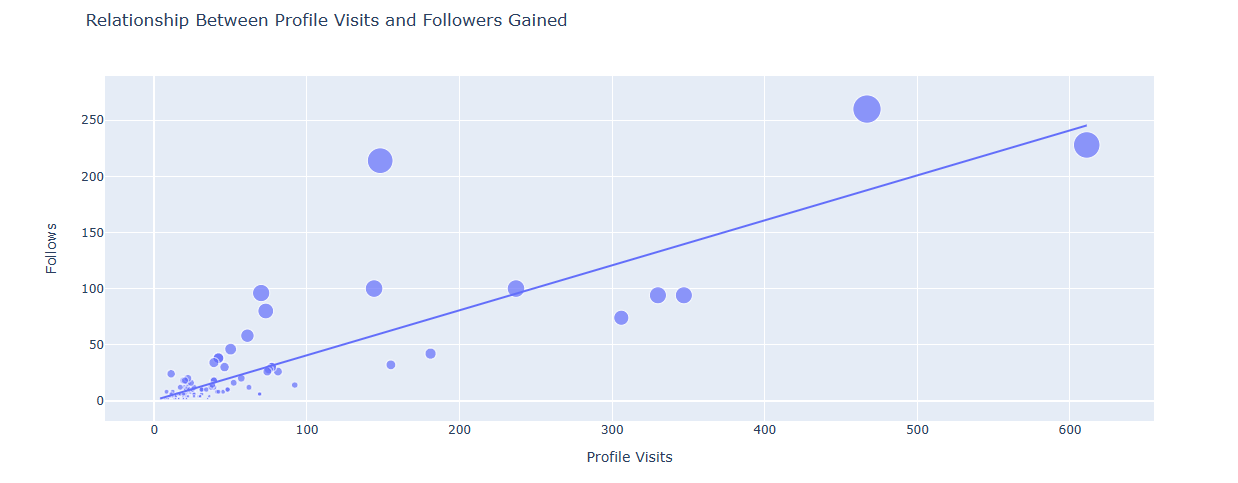
Statical analysis of the Id

We calculated the conversion rate which tells how many of the profile visits are converted into follows we then multiplied it to 100 to convert it into percentage.



The conversion rate in 41% here which depicts high follow rate.

Profile visits and follows relationship.



This graph produced a high upward slope which represents positive dependency of the profile visits with follows, which concludes almost half of the people which visit the profile are likely to follow.

12

Hypothesis Formulation: -

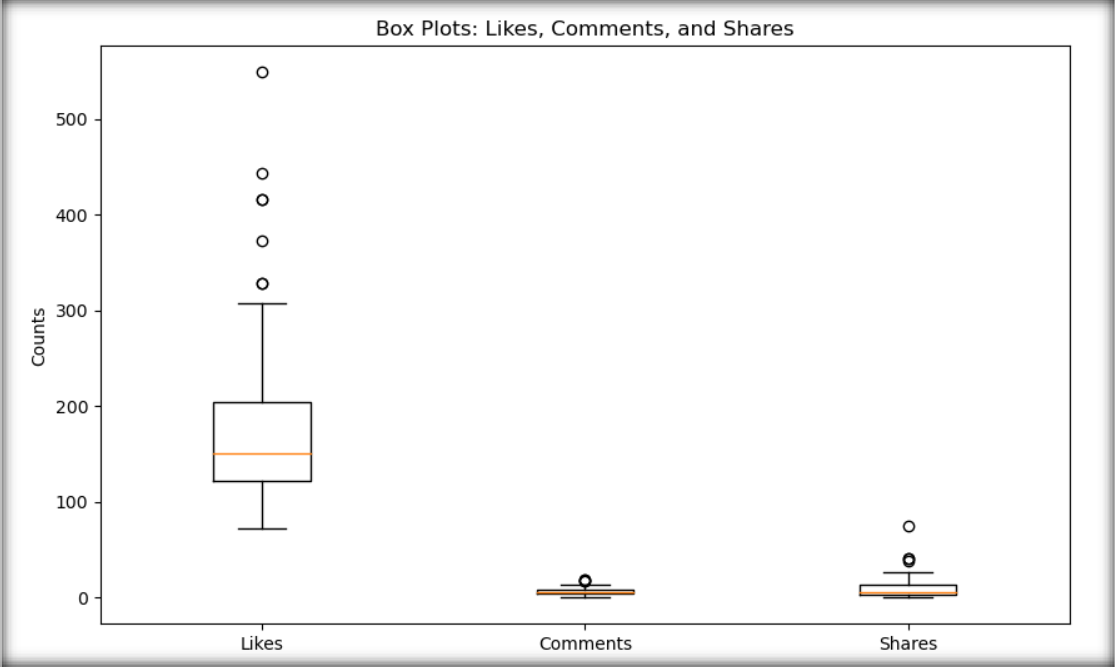
Hypothesis formulation involves stating clear and testable statements about the relationships or patterns you expect to find in your data. Based on the code snippets you provided for Instagram data analysis, let's formulate a few hypothetical

statements:

Hypothesis about Engagement Metrics:

 1.Null Hypothesis (H0): The average number of likes, comments, and shares does not significantly differ across different categories (From Home, From Hashtags, From Explore, etc.).

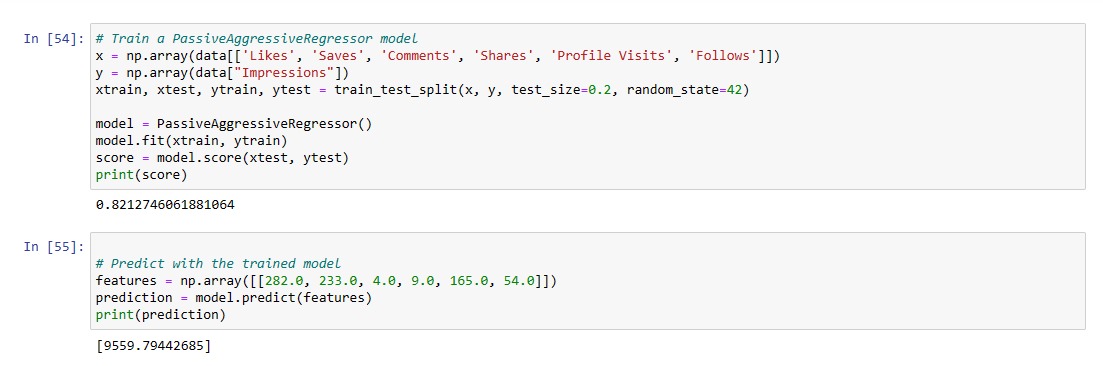
2.Alternative Hypothesis (H1): There is a significant difference in the average engagement metrics across different categories.



13

Prediction using our dataset.

We used the PassiveagressiveRegressor model for training and testing our data set.



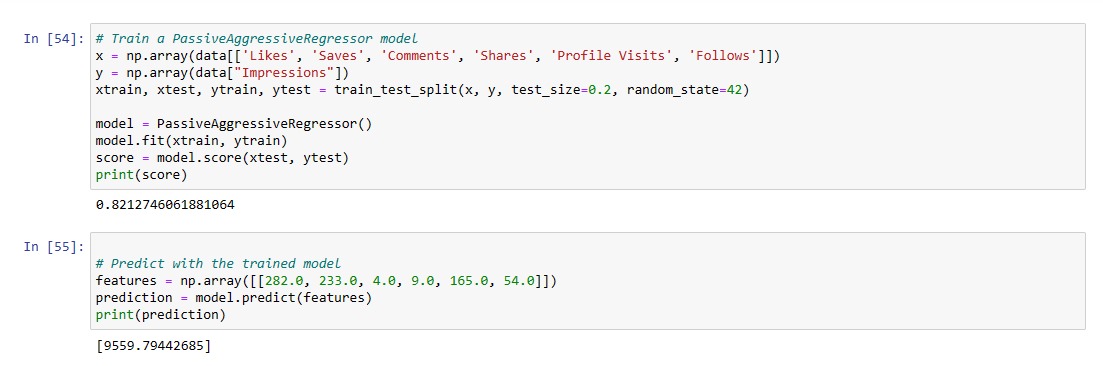
We used the model to split the data set into testing and training sets .20% is used for testing whereas the 80% is used for training.

xtest is the feature data for testing, and ytest is the corresponding actual target values.

The calculated R^2 score provides a measure of how well the model predicts the target variable on the test set. The score ranges from -1 to 1, where 1 indicates a perfect prediction.

Here, our score is 0.821274606 which depicts at fairly accurate prediction.

After the testing the model is ready to predict



model. predict is a method of the trained model (PassiveAggressiveRegressor) that takes a set of features as input and produces a predicted output.

In this case, it predicts the 'Impressions' based on the input features.

Print the Prediction.

Here, the model predicts 9559.79442685 impressions predicted for the Instagram Id.

14

Conclusion: -

In conclusion, the analysis of Instagram data provided valuable insights into user engagement and content performance. Through exploratory data analysis (EDA) and statistical testing, we gained a comprehensive understanding of the relationships between various metrics. Here is a summary of the key findings and steps taken:

1.Exploratory Data Analysis (EDA):

Conducted a thorough exploration of the dataset, summarizing its characteristics and visualizing key statistics.

Explored relationships between different metrics using scatter plots, histograms, and pair plots.

Analyzed the distribution of engagement metrics across categories and identified potential trends over time.

2.Statistical Testing:

Tested hypotheses related to engagement metrics,

correlation between follows and impressions, and time trends in likes.

Utilized statistical tests such as ANOVA, Pearson

correlation coefficient, and regression analysis to assess the significance of relationships.

3.Machine Learning Integration:

Prepared and preprocessed the data for machine learning.

Selected relevant features and split the data into training and testing sets.

Chose a machine learning algorithm (Random Forest Regressor in this example) and trained the model.

Evaluated the model's performance using mean squared error and explored hyperparameter tuning for optimization.

4. Insights and Recommendations:

Identified patterns and correlations in the data, providing valuable insights for content strategy and user engagement optimization.

The machine learning model can be used for predictive analysis, potentially forecasting future engagement metrics based on historical data.

5. Limitations and Future Work:

Acknowledged any limitations in the analysis, such as assumptions made during imputation or the scope of the dataset 15

Considered areas for future work, such as collecting additional data, refining models, or exploring more advanced machine learning techniques.

Overall, this analysis serves as a foundation for data-driven decision-making in optimizing Instagram content and user.

engagement. The combination of EDA, statistical testing, and machine learning provides a comprehensive approach to extract meaningful insights from the data. Continued refinement and exploration can contribute to an even deeper understanding of user behavior on the platform.

16