Analysis of Social Media Influence on Consumer Behaviour

A Project work

Submitted in the partial fulfilment for the award of the degree of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE WITH SPECIALIZATION IN ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Submitted by:

20BCS6712	AMISHA KHANNA
20BCS6724	RIVI VIG
20BCS6735	CHARU GARG
20BCS6749	DEVI PRASAD SAMANTARAY

Under the Supervision of:

Mr. Pramod Vishwakarma

CHANDIGARH UNIVERSITY

Discover. Learn. Empower.

CHANDIGARH UNIVERSITY, GHARUAN, MOHALI - 140413, PUNJAB

April, 2024

CHANDIGARH UNIVERSITY

Discover. Learn. Empower.

DECLARATION

'Amisha Khanna', 'Rivi Vig' , 'Charu Garg' and 'Devi Prasad', students of 'Bachelor of Engineering in IBM SPECIALISEDARTIFICIAL INTELLIGENCE AND MACHINE LEARNING', session: 2023-24, Department of Computer Science and Engineering, Apex Institute of technology, Chandigarh University, Punjab, here by declare that the work presented in this Project Work entitled '**Analysis of Social Media Influence on Consumer Behaviour'** is the outcome of our own bonafide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics and is done under the Supervision of 'Mr. Pramod Vishwakarma'**.**

Date: 30/4/2024

20BCS6712	AMISHA KHANNA
20BCS6724	RIVI VIG
20BCS6735	CHARU GARG
20BCS6749	DEVI PRASAD SAMANTARAY

CHANDIGARH UNIVERSITY

Discover. Learn. Empower.

ACKNOWLEDGEMENT

We extend our heartfelt gratitude to our esteemed supervisor, Mr Pramod Vishwakarma, whose unwavering guidance and insightful feedback played a pivotal role in shaping this research endeavor. We also express our sincere thanks to the dedicated panelists who meticulously evaluated our work, offering invaluable perspectives and constructive critique.

We are deeply appreciative of Chandigarh University for providing us with a nurturing academic environment and resources essential for our research. The support and encouragement from the Department of AIT CSE have been instrumental in our academic journey.

Our friends and parents deserve our profound appreciation for their unwavering support, understanding, and encouragement throughout this endeavor. Their belief in us has been our greatest motivation, and we are profoundly grateful for their presence in our lives.

TABLE OF CONTENTS

List of Figures

Abstract

In an era dominated by digital connectivity, this research probes into the complex web of social media influence on customer behaviour. Motivated by the exponential rise of online platforms shaping consumer decisions, the study aims to unravel the underlying dynamics and implications for businesses. Our objectives include decrypting the key factors driving customer behaviour on these platforms and evaluating their impact on purchasing decisions. Employing a mixed-methods approach, combining qualitative and quantitative analyses, we aim to scrutinize user engagement, sentiment analysis, and behavioural patterns across diverse social media channels. This research endeavours to equip businesses with actionable insights, fostering strategic adaptations to meet evolving consumer expectations. In conclusion, this study contributes to the evolving landscape of digital marketing by clarifying the pivotal role that social media plays in shaping the contemporary consumer journey.

Keywords: Social media influence, Customer behaviour, Consumer decisions, Mixed-methods approach, Comprehensive insights, Actionable insights, Digital marketing

CHAPTER-1

INTRODUCTION

1.1 Problem Definition

In today's dynamic marketplace, where social media platforms serve as hubs of consumer interaction and influence, businesses are grappling with the challenge of deciphering the complex web of factors that shape consumer decision-making online. The rise of social media has transformed the traditional consumer journey, blurring the lines between brand awareness, consideration, and purchase. However, amidst this digital revolution, many businesses find themselves adrift, lacking the necessary insights to navigate effectively. This research endeavors to bridge this gap by delving into the nuanced mechanisms through which social media exerts its influence on consumer behavior. By identifying the key drivers and dynamics at play, businesses can gain a deeper understanding of their target audiences' digital behaviors, preferences, and motivations. Armed with these insights, companies can devise more tailored and impactful strategies to engage with consumers, foster brand loyalty, and drive conversions in the ever-evolving landscape of social commerce.

1.2 Problem Overview

In the contemporary digital landscape, social media platforms like Instagram, Facebook, Twitter, and TikTok have become integral parts of consumers' lives, profoundly influencing their purchasing decisions. For instance, consider the phenomenon of influencer marketing, where individuals with large followings on social media endorse products or services. These influencers often have a

significant impact on their followers' purchasing behavior. A study by Mediakix found that 89% of marketers believe that ROI from influencer marketing is comparable to or better than other marketing channels. This highlights the power of social media personalities in shaping consumer perceptions and preferences.

Moreover, the rise of user-generated content (UGC) has further blurred the lines between marketing messages and authentic peer recommendations. Platforms like Yelp for restaurants or TripAdvisor for travel destinations thrive on user reviews and ratings, which heavily influence potential customers' decisions. Research by BrightLocal revealed that 88% of consumers trust online reviews as much as personal recommendations, emphasizing the profound impact of UGC on consumer behavior.

Additionally, the phenomenon of viral marketing showcases how social media can rapidly amplify brand messages and shape consumer perceptions. Take the example of the ALS Ice Bucket Challenge, which went viral on platforms like Facebook and Twitter. This grassroots campaign not only raised awareness about amyotrophic lateral sclerosis (ALS) but also generated significant donations to the cause. The viral nature of the challenge demonstrates the unparalleled reach and influence of social media in mobilizing communities and driving collective action.

Furthermore, the advent of social commerce has revolutionized the way consumers shop online. Platforms like Instagram Shopping and Pinterest Buyable Pins enable users to discover and purchase products seamlessly within the app, blurring the lines between content and commerce. According to eMarketer, social commerce sales are projected to reach $36.62 billion in the US alone by 2023, underscoring the transformative impact of social media on e-commerce.

In essence, these examples illustrate the multifaceted impact of social media on consumer behavior, from influencer endorsements and user-generated content to viral marketing and social commerce. Businesses must decipher these complex dynamics to craft effective strategies that resonate with their target audience in the ever-evolving digital landscape.

1.3 Timeline

Week 1-2: Project Initiation and Literature Review

Week 1: Project Kickoff

- Define project objectives, deliverables, and team roles.
- Set up project management tools and communication channels.
- Initiate the literature review process to gather existing research on social media influence on consumer behavior.

Week 2: Literature Review and Research Design

- Conduct an extensive literature review to identify gaps, trends, and key findings.
- Develop a research framework and methodology based on the literature review findings.
- Refine project objectives and finalize the research approach.

Week 3-5: Data Collection and Compilation

Week 3: Data Sources Identification

- Identify sources of data, including social media platforms, research databases, and industry reports.
- Establish criteria for selecting relevant data sources and datasets.

Week 4: Data Gathering

- Collect data from various social media platforms, including user-generated content, engagement metrics, and consumer feedback.
- Compile a diverse and comprehensive dataset for analysis.

Week 5: Data Preprocessing

- Clean and preprocess the collected data, including data deduplication, normalization, and formatting.
- Conduct initial exploratory data analysis to identify patterns and trends.

Week 6-8: Data Analysis and Model Selection

Week 6: Exploratory Data Analysis

- Analyze the preprocessed data to uncover insights into consumer behavior patterns and trends on social media.
- Identify key factors influencing consumer decisions and preferences.

Week 7: Model Exploration

- Explore and evaluate different machine learning models suitable for analyzing social media data, such as sentiment analysis algorithms and recommendation systems.
- Select the most appropriate model based on performance metrics and requirements.

Week 8: Model Adaptation and Optimization

- Adapt the selected machine learning model to the project's specific requirements and data characteristics.

- Optimize the model parameters and hyperparameters to improve performance and accuracy.

Week 9-11: Data Interpretation and Strategic Recommendations

Week 9: Interpretation of Results

- Interpret the findings from the data analysis and model evaluation to extract actionable insights.
- Identify trends, patterns, and correlations between social media activity and consumer behavior.

Week 10: Strategic Recommendations

- Develop strategic recommendations for businesses based on the research findings and insights.
- Provide guidance on how businesses can leverage social media to engage with consumers effectively and drive desired outcomes.

Week 11: Report Preparation

- Prepare a comprehensive report documenting the research methodology, findings, and strategic recommendations.
- Finalize the report format, structure, and visuals for maximum clarity and impact.

This timeline outlines the project's progression from initiation to completion, with a focus on key activities and milestones during each phase. It allows for efficient project management and ensures that the research is conducted rigorously and delivers actionable insights to businesses.

1.4 Hardware Specification

To effectively analyze the influence of social media on consumer behavior, it's crucial to have hardware that can handle the demands of processing large datasets, running complex algorithms, and performing potentially intensive computations. Here's a comprehensive outline of the hardware requirements:

1. CPU (Central Processing Unit): A powerful multi-core processor is essential for efficiently handling data processing tasks. Look for CPUs with multiple cores and high clock speeds to accelerate computations. Modern CPUs from Intel, such as the Core i7 or Core i9 series, offer excellent performance for data-intensive tasks.

2. GPU (Graphics Processing Unit): Many machine learning algorithms benefit greatly from parallel processing, making a dedicated GPU a valuable asset. NVIDIA GPUs are widely preferred in the machine learning community for their exceptional performance and support for frameworks like TensorFlow and PyTorch. Options like the NVIDIA GeForce RTX or NVIDIA Quadro series provide robust computational capabilities.

3. RAM (Random Access Memory: Having ample RAM is crucial for efficiently handling large datasets. Aim for at least 16GB of RAM to ensure smooth operation, but if your datasets are particularly large, consider upgrading to 32GB or more for optimal performance.

4. Storage: Solid State Drives (SSDs) are highly recommended for faster data access and overall system responsiveness. Opt for SSDs with sufficient storage capacity to accommodate large datasets and software applications. Additionally, consider investing in a fast NVMe SSD for even quicker data

read/write speeds.

5. Monitor: A high-resolution monitor with good color accuracy is beneficial for data visualization and analysis. Choose a monitor with at least Full HD (1920 x 1080) resolution, but ideally, opt for a higher resolution such as Quad HD (2560 x 1440) or Ultra HD (3840 x 2160) for enhanced clarity and detail.

6. Peripherals: Don't overlook the importance of comfortable peripherals for prolonged work sessions. Invest in a quality keyboard and mouse that provide ergonomic support to minimize strain. Additionally, consider peripherals such as a high-quality headset for audio analysis or a graphics tablet for annotation tasks.

By ensuring that your hardware meets these requirements, you can create an optimal computing environment for conducting in-depth analysis of social media influence on consumer behavior, enabling efficient data processing, modeling, and visualization.

1.5 Software Specification

In the analysis of social media influence on consumer behavior, employing the right programming language, methodologies, and tools is crucial for extracting meaningful insights. Here's an extended overview of the key steps involved in the project:

1. Programming Language: Python is the preferred programming language for machine learning projects due to its extensive libraries for data analysis and machine learning. Libraries such as NumPy,

pandas, scikit-learn, TensorFlow, and PyTorch provide robust functionalities for data manipulation, modeling, and analysis.

2. Data Collection and Cleaning: Utilize social media APIs, such as the Twitter API or Facebook Graph API, to collect relevant data pertaining to consumer behavior. Preprocess the collected data to handle missing values, remove duplicates, and perform text normalization techniques like tokenization, stemming, and lemmatization to ensure data cleanliness and consistency.

3. Exploratory Data Analysis (EDA): Employ libraries like pandas, NumPy, and Matplotlib/Seaborn to perform exploratory data analysis. Visualize distributions, correlations, and trends in social media data and consumer behavior to gain insights into key patterns and relationships.

4. Feature Engineering: Extract relevant features from social media data, including text features, engagement metrics, and user demographics. Transform raw features into a format suitable for machine learning algorithms by encoding categorical variables, scaling numerical features, and performing dimensionality reduction techniques if necessary.

5. Machine Learning Models: Select appropriate machine learning algorithms based on the nature of the analysis. For sentiment analysis tasks, consider algorithms like Naive Bayes, Support Vector Machines, Recurrent Neural Networks (RNNs), or advanced transformer models like BERT. For predictive modeling, options include regression techniques (linear, logistic), decision trees, random forests, gradient boosting machines (e.g., XGBoost, LightGBM), or neural networks.

6. Model Evaluation: Split the data into training, validation, and testing sets to evaluate model

performance. Use metrics relevant to the task, such as accuracy, precision, recall, F1-score for classification tasks, or RMSE, MAE for regression tasks. Implement cross-validation techniques to ensure the robustness of the models and mitigate overfitting.

7. Deployment: Deploy the trained model as a service using frameworks like Flask for building web applications. Utilize cloud platforms such as AWS, Google Cloud Platform, or Microsoft Azure for hosting the application, ensuring scalability, and reliability.

8. Monitoring and Maintenance: Implement logging and monitoring mechanisms to track the performance of the deployed model. Regularly update the model with new data and retrain if necessary to maintain accuracy and relevance. Address any issues or bugs that arise in the deployed application promptly.

9. Documentation: Document the entire process, including data collection methods, preprocessing steps, model selection, evaluation metrics, and deployment procedures. Provide clear instructions for future maintenance and updates to ensure the reproducibility and scalability of the project.

10. Collaboration and Version Control: Utilize version control systems like Git to manage code changes and collaborate with team members efficiently. Document code using comments and docstrings for better understanding and maintainability, facilitating collaboration and knowledge sharing within the team.

By following these steps and leveraging the appropriate methodologies and tools, businesses can gain valuable insights into the influence of social media on consumer behavior, enabling them to make informed decisions and optimize their marketing strategies effectively.

CHAPTER-2

LITERATURE SURVEY

2.1 Timeline of the Problem as Investigated Throughout the World

In the current digital landscape, businesses are navigating a rapidly evolving marketplace where consumer behavior is increasingly shaped by social media interactions. Traditionally, businesses have relied on conventional market research methods such as surveys, focus groups, and demographic analysis to understand consumer preferences and purchasing decisions. While these methods have provided valuable insights, they often lack the depth and real-time data necessary to comprehensively capture the intricate nuances of social media influence on consumer behavior.

One of the primary limitations of traditional market research methods is their inability to capture real-time consumer sentiments and behaviors as they unfold on social media platforms. Surveys and focus groups, while informative, are often time-consuming and may not provide timely insights into shifting consumer trends or emerging patterns. Additionally, demographic analysis, while useful for segmenting consumer groups, may not fully capture the diverse and dynamic nature of social media interactions.

Existing social media analytics tools offer some insights into consumer behavior by tracking quantitative metrics such as likes, shares, and follower counts. However, these tools often fall short in providing a holistic understanding of consumer sentiments and motivations. They focus primarily on surface-level engagement metrics and may overlook the emotional resonance of content or the underlying drivers of consumer behavior.

Furthermore, existing social media analytics tools may lack advanced sentiment analysis capabilities, making it challenging for businesses to gauge the emotional impact of their marketing efforts accurately. Sentiment analysis is crucial for understanding how consumers perceive brand messaging and how it influences their purchasing decisions. Without robust sentiment analysis tools, businesses may miss out on valuable insights into consumer sentiment trends and fail to tailor their marketing strategies accordingly.

Moreover, behavioral tracking capabilities in existing social media analytics tools may be limited, hindering businesses' ability to understand the sequential patterns of consumer interactions and engagement on social media platforms. Behavioral tracking allows businesses to track user journeys, identify key touchpoints, and optimize their marketing strategies to drive conversion and retention.

In summary, while traditional market research methods and basic social media analytics tools have their merits, they may not provide the depth of insights needed to fully grasp the complexities of social media influence on consumer behavior. Businesses must invest in advanced analytics tools and methodologies that offer real-time data insights, sophisticated sentiment analysis, and comprehensive behavioral tracking capabilities to stay competitive in today's digital landscape.

2.2 Bibliographic Analysis

Year and Citation	Article/ Author	Tools/ Software	Technique	Source	Evaluation Parameter
2022	P. Grover, A.K. Kar and Y. Dwivedi et al (2022) 100116)	Scopus Database	TCCM (Theory, Context, Characteristics, and Methodology)	IJIMDI	Theory classification, Context classification, Characteristics classification
2023	N. K. Hoi Et al 2023	One-Way Analysis of Variance (ANOVA)	Quantitative survey, Questionnaire survey	European Journal of Business and Management Research	Cronbach's α, Internal consistency of the scale variables
2021	Sachin Gupta et al (2021)	Questionnaire survey in Delhi	TCCM (Theory, Context, Characteristics, and Methodology)	IJCRT	Consumer Behaviour Analysis, Social Media Influence Assessment

2020	Brown, A. et al Joan-Francesc Fondevila-Gascón et al (2020).	Social media analysis programs	Statistical tests, Student's t-test	MDPI (Multidisciplinary Digital Publishing Institute)	Categorical data analysis, Consumer Behaviour Analysis
2023	Prakash Singh el at (2023)	Smart PLS 4.0 software	Path analysis, Structural equation modelling (SEM)	Taylor and Francis (Information & Technology Management)	Common method bias (CMB) Cronbach's alpha Composite reliability
2022	Hasan Mahmud et al (2022)	Excel sheets for data extraction and synthesis	Thematic Analysis	Technological Forecasting and Social Change Vol 175	Quality assessment criteria (QAC)
2022	Sykora, M., Elayan et al. 2022	VADER sentiment analysis tool,	Predicting customer response sentiment	Journal of the Academy of Marketing Science,	Accuracy of sentiment prediction

		Support Vector Machines (SVM)		48(4), 630–648	

.

2.3 Proposed System

The proposed system aims to address the limitations of the existing approach by leveraging advanced data analytics techniques and integrating qualitative and quantitative methodologies. Key features of the proposed system include:

Advanced Sentiment Analysis:

- Utilize natural language processing (NLP) algorithms to conduct sentiment analysis on social media content, providing deeper insights into the emotional resonance of posts and user-generated content.

Behavioral Tracking:

- Implement advanced tracking mechanisms to monitor user behaviour across various social media platforms. This includes analysing browsing habits, interaction patterns, and decision-making processes to uncover nuanced insights into consumer behaviour.

Integrated Data Analytics Platform:

- Develop a centralized data analytics platform that integrates data from multiple social media channels. This platform provides businesses with a holistic view of consumer behaviour, enabling them to identify cross-channel trends and patterns.

Real-time Insights:

- Offer real-time analytics capabilities to provide businesses with up-to-date insights into consumer behaviour trends. This enables agile decision-making and allows businesses to adapt their marketing strategies in response to changing consumer preferences.

Predictive Analytics:

- Implement predictive analytics models to forecast future consumer behaviour trends based on historical data and current market dynamics. This empowers businesses to anticipate consumer needs and proactively tailor their marketing efforts accordingly.

CHAPTER-3

DESIGN PROCESS

3.1 Concept Generation

Problem Formulation:

The dynamic interplay between social media and consumer behavior presents a significant challenge for businesses in navigating the evolving digital landscape. At the heart of this challenge is the imperative need to address the intricate complexities surrounding the influence of social media on the decision-making processes of contemporary consumers.

Understanding the Core Challenge:

1. Fundamental Shift in Consumer Behavior:

The traditional consumer journey has undergone a paradigmatic transformation, with social media emerging as a primary touchpoint throughout the decision-making process. Individuals increasingly rely on social media platforms for information, recommendations, and peer interactions, shaping their choices in unprecedented ways.

2. Scarcity of Comprehensive Insights:

Despite the evident role of social media in shaping consumer decisions, there is a scarcity of comprehensive insights into the nuanced dynamics of social media influence. The specific mechanisms driving these choices remain elusive, leaving businesses grappling with a lack of clarity regarding factors such as user engagement, sentiment dynamics, and behavioral patterns across diverse social media channels.

3. Strategic Adaptation and Business Implications:

The necessity for businesses to strategically adapt to this evolving landscape underscores the urgency of addressing the identified problem. Without a profound understanding of the underlying forces at play, companies struggle to formulate effective strategies to connect with their target audience. This lack of strategic clarity has direct implications on market positioning, brand perception, and ultimately, the ability to meet the dynamic expectations of consumers navigating the digital marketplace.

Addressing the Challenge:

To overcome the challenges posed by the dynamic interplay between social media and consumer behavior, businesses must prioritize the following approaches:

1. In-Depth Research and Analysis:

Conduct comprehensive research and analysis to uncover the intricate dynamics of social media influence on consumer behavior. This involves exploring factors such as user engagement, sentiment analysis, and behavioral patterns across various social media platforms.

2. Strategic Adaptation:

Develop strategic frameworks and adaptation strategies to align with the evolving digital landscape. This may include leveraging insights gained from research to refine marketing strategies, enhance brand positioning, and optimize engagement tactics on social media platforms.

3. Continuous Monitoring and Evaluation:

Implement robust monitoring and evaluation mechanisms to track changes in consumer behavior trends and sentiment dynamics on social media. This allows businesses to adapt and iterate their strategies in

real-time, ensuring relevance and effectiveness in engaging with their target audience.

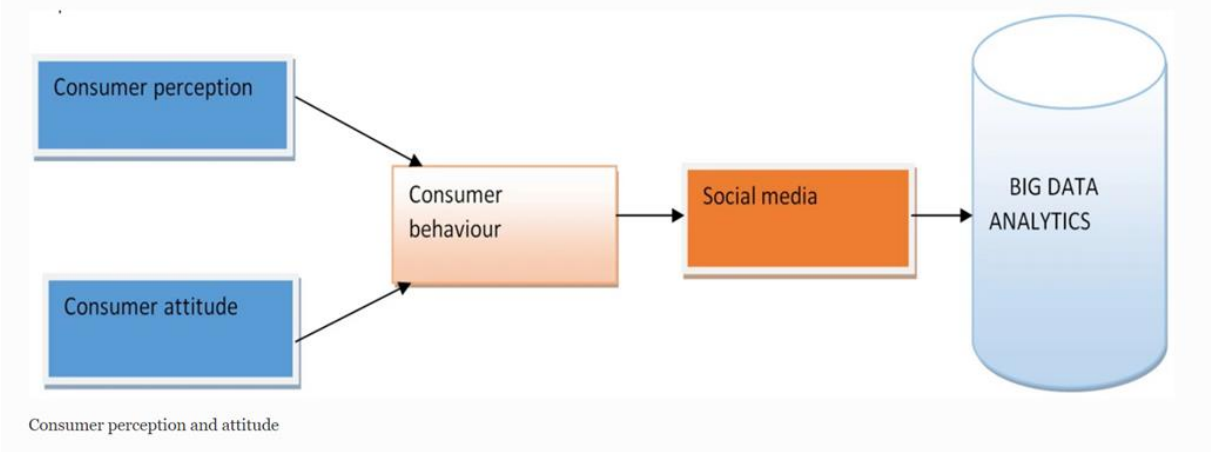


Fig1:Consumer perception and attitude pipeline

3.2 Evaluation & Selection of Specifications/features:

This study aims to untangle the intricate relationship between social media and consumer behaviour, providing businesses with actionable insights to navigate the evolving digital landscape. The core objectives include:

1. Decrypting Key Factors Driving Customer Behaviour:
    - Identify and analyse the key factors shaping consumer decisions on social media.
2. Evaluating Impact on Purchasing Decisions:
    - Assess the direct influence of social media on consumer purchasing decisions.
3. Understanding User Engagement Dynamics:
    - Scrutinize user engagement metrics across various social media channels.
    - Analyse patterns of user interactions and their implications on consumer choices.
4. Conducting Sentiment Analysis:
    - Employ sentiment analysis to gauge the emotional resonance of content on social media.

- Investigate how expressed sentiments influence consumer purchasing decisions.

5. Analysing Behavioural Patterns Across Platforms:
    - Examine behavioural patterns exhibited by consumers on different social media platforms.
    - Identify trends and variations in user behaviour, guiding businesses to tailor strategies accordingly.
6. Utilizing a Mixed-Methods Approach:
    - Combine qualitative and quantitative methods for a comprehensive exploration.
    - Utilize qualitative analyses for in-depth insights and quantitative data for statistical assessments.
7. Providing Actionable Insights for Businesses:
    - Translate research findings into practical recommendations for optimizing social media strategies.
    - Offer guidance on adapting marketing approaches based on identified social media influences.
8. Fostering Strategic Adaptations
    - Facilitate businesses in aligning strategies with the evolving landscape of consumer expectations.
    - Provide insights on adjusting marketing strategies to leverage social media dynamics.
9. Optimizing Online Presence:
    - Guide businesses in optimizing their online presence based on user preferences and behaviours.
    - Assist in creating content and engagement strategies that resonate with the target audience.

10. Fostering a Deeper Connection with the Audience:

- Facilitate businesses in fostering a deeper and more meaningful connection with their target audience.
- Explore strategies to build brand loyalty through effective utilization of social media channels.

In summary, these research objectives collectively form a robust framework to unravel the intricacies of social media influence on consumer behaviour. By achieving these objectives, this study endeavours to equip businesses with insights essential for successful navigation of the ever-evolving digital landscape.

3.3 Methodology

This research employs a mixed-methods approach to comprehensively explore the influence of social media on consumer behaviour. The methodology encompasses both qualitative and quantitative techniques to provide a multifaceted understanding of the phenomenon.

Data Collection:

- Qualitative Data: Conduct in-depth interviews and focus group discussions with consumers to gain insights into their perceptions, experiences, and decision-making processes influenced by social media.
- Quantitative Data: Gather data from social media platforms using web scraping techniques or through API access. Collect metrics such as user engagement (likes, shares, comments), sentiment analysis scores, and demographic information.

Content Analysis:

- Qualitative Analysis: Utilize thematic analysis to identify recurring themes and patterns in qualitative data obtained from interviews and focus groups. Extract insights regarding consumer attitudes, preferences, and motivations related to social media influence.
- Quantitative Analysis: Conduct statistical analysis of quantitative data to identify correlations and trends. Analyze user engagement metrics, sentiment scores, and demographic distributions to uncover patterns indicative of social media influence on consumer behaviour.

Sentiment Analysis:

- Utilize natural language processing (NLP) techniques to conduct sentiment analysis on social media content. Classify user-generated content into positive, negative, or neutral sentiments to assess the emotional resonance of posts and their potential impact on consumer perceptions and purchasing decisions.

Behavioural Analysis:

- Qualitative Observation: Observe and document consumer behaviour on social media platforms through participant observation or online ethnography. Capture behavioural patterns, such as browsing habits, interaction with content, and decision-making processes.
- Quantitative Analysis: Analyse quantitative data to identify behavioural patterns across different social media platforms. Utilize techniques such as cluster analysis or factor analysis to categorize users based on their behaviour and preferences.

Ethical Considerations:

- Ensure ethical standards are upheld throughout the research process, particularly concerning data privacy and informed consent.
- Adhere to ethical guidelines for conducting research involving human subjects, including transparency in data collection procedures and maintaining participant confidentiality.

Limitations:

Acknowledge potential limitations of the study, such as sample biases in qualitative research and data inaccuracies in quantitative analysis.

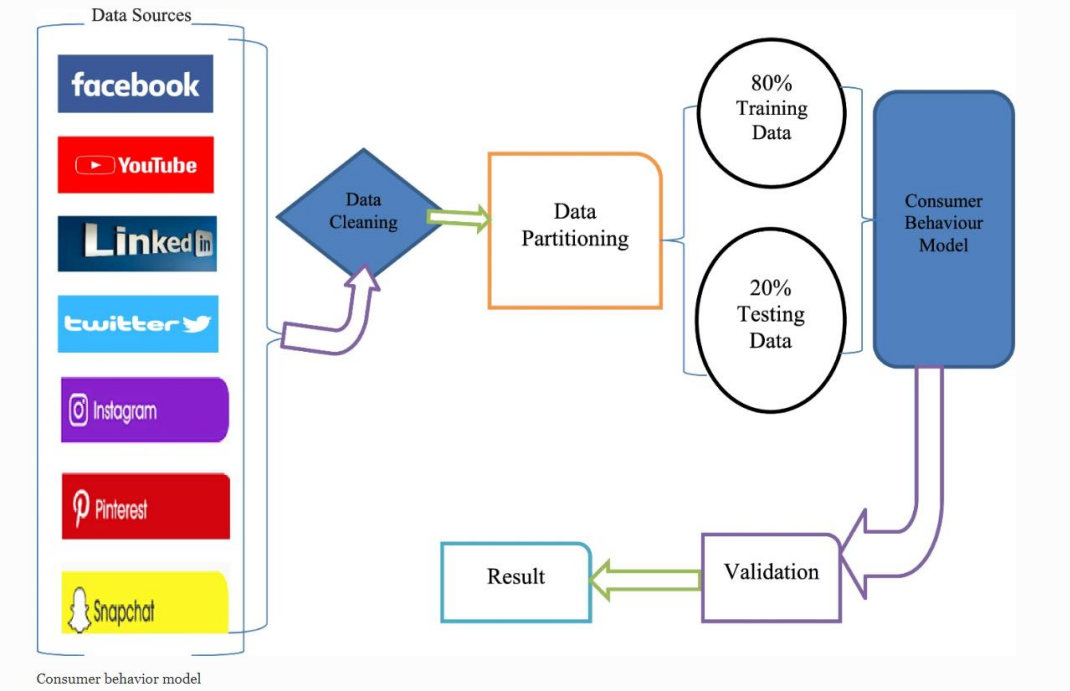


Fig2:Consumer behavior model

3.4 Implementation Plan

Social media platforms, particularly Twitter, have become pivotal in shaping consumer behavior and influencing purchasing decisions. Understanding the impact of social media on consumer behavior is crucial for businesses to develop effective marketing strategies. In this project, we aim to analyze the relationship between social media activity on Twitter and consumer behavior.

Data Collection:

We will collect Twitter data using the Twitter API, focusing on tweets related to specific brands, products, or industries. The data collection process will involve defining relevant keywords, hashtags, and filters to capture tweets that reflect consumer opinions, sentiments, and interactions.

Dataset Preparation:

The collected Twitter data will be preprocessed to extract relevant features such as text content, user engagement metrics (likes, retweets, replies), timestamps, and user information (followers count, verified status). Preprocessing steps will include text normalization, tokenization, removing stop words, and sentiment analysis.

Feature Engineering:

We will engineer additional features from the raw data to capture various aspects of social media influence, such as user influence metrics (follower count, retweet count), temporal features (time of day, day of week), and content-based features (sentiment polarity, topic modeling).

Model Selection:

Several machine learning models will be considered for analysis, including but not limited to:

1. Linear Regression: A simple yet powerful statistical method used for modeling the relationship between a dependent variable and one or more independent variables, providing insights into the linear associations within the data.

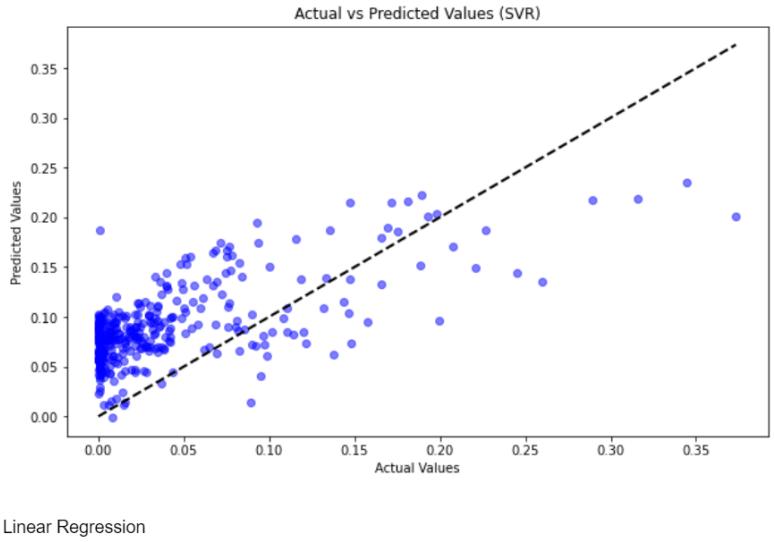
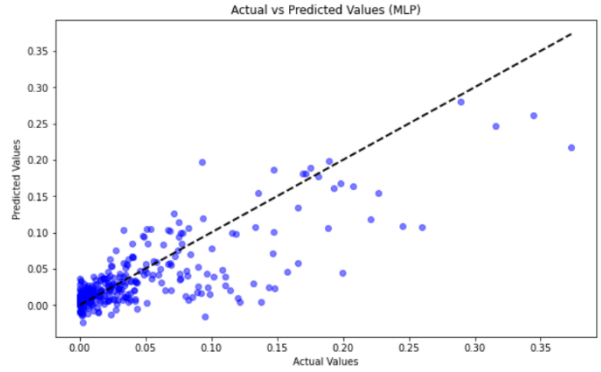


Figure 3 *Linear Regression*

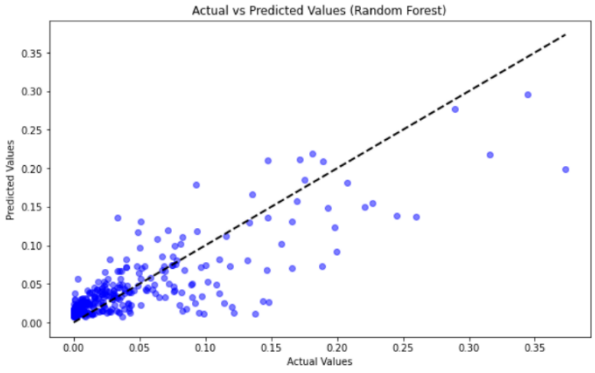
2. Random Forest: An ensemble learning technique that constructs multiple decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees, offering robustness and accuracy in handling complex datasets.



Random Forest

Figure 4 *Random Forest*

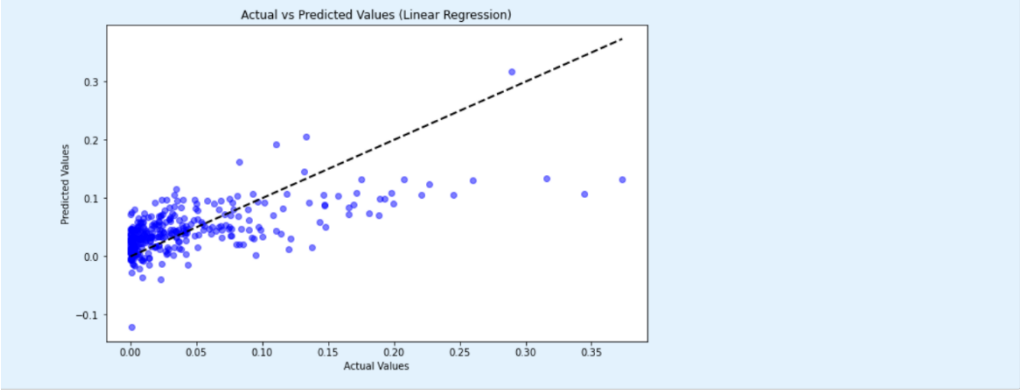
3. XGBoost: An optimized gradient boosting algorithm known for its efficiency and speed in building decision tree ensembles, making it a popular choice for predictive modeling tasks due to its high performance and scalability.



XGBoost

Figure 5 *XGBoostt*

4. Multilayer Perceptron (MLP): A type of feedforward artificial neural network with multiple layers of nodes, capable of learning complex nonlinear relationships between input and output data, often used for classification and regression tasks in machine learning.



Multi Level Perceptron

Figure 6 *MultiLevel Perceptron*

5. MLP Regressor: A variant of the multilayer perceptron specifically designed for regression tasks, utilizing multiple layers of nodes with nonlinear activation functions to predict continuous target variables based on input features, offering flexibility and adaptability in modeling diverse datasets.

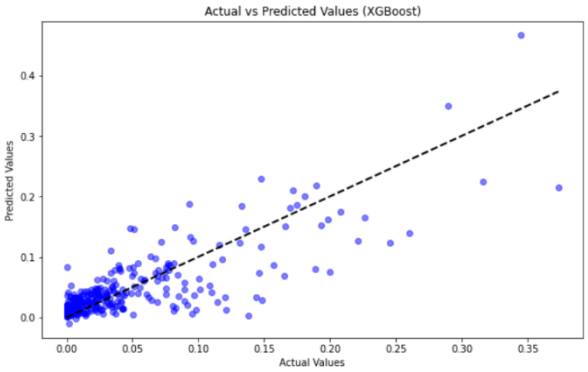


Figure 7 *MLP Regressor*

The choice of models will be based on their suitability for handling the characteristics of the Twitter data and their ability to predict consumer behavior accurately.

Evaluation Metrics:

To evaluate the performance of the models, we will use metrics such as accuracy, precision, recall, F1-score, and area under the Receiver Operating Characteristic (ROC) curve. Additionally, we will employ cross-validation techniques to assess the generalization ability of the models.

Experimental Design:

The experimental design will involve the following steps:

1. Data splitting: The dataset will be divided into training, validation, and test sets.

2. Model training: We will train the selected machine learning models on the training data using various combinations of features.

3. Hyperparameter tuning: Grid search or random search techniques will be employed to optimize the hyperparameters of the models.

4. Model evaluation: The trained models will be evaluated on the validation set to select the best-performing model.

5. Performance assessment: The final model will be evaluated on the test set to assess its performance in predicting consumer behavior based on social media influence.

CHAPTER-4

RESULT ANALYSIS AND VALIDATION

4.1 Implementation of Design using Modern Engineering Tools in Analysis

Implementing the selected design for analyzing social media influence on consumer behavior using modern engineering tools and analysis involves a systematic process to translate the design into a functional system. This process is crucial for ensuring that the system meets the defined objectives and specifications. Here is a step-by-step overview of the implementation process:

1. Data Collection and Compilation:

- Begin by collecting a diverse and comprehensive dataset of social media interactions and consumer behavior. Utilize APIs and data scraping techniques to gather data from various social media platforms.
- Ensure that the collected data aligns with the defined objectives and covers a wide range of user demographics, content types, and engagement metrics.

2. Data Preprocessing:

- Clean and preprocess the collected data to handle missing values, remove duplicates, and standardize formats. Perform text preprocessing techniques such as tokenization, stemming, and lemmatization to prepare the data for analysis.
- Apply data augmentation techniques if necessary to enhance the robustness of the dataset.

3. Software Development:

- Develop or configure the software required for data processing, analysis, and visualization. This software should support real-time data processing, sentiment analysis, and behavioral tracking.

- Implement a user-friendly interface for data exploration and visualization, allowing users to interact with the data effectively.

4. Model Selection and Development:

- Select appropriate machine learning algorithms and statistical models for analyzing social media data and consumer behavior. Consider techniques such as sentiment analysis, topic modeling, and predictive modeling.

- Develop and train machine learning models using the preprocessed data to extract insights into consumer behavior patterns, sentiment trends, and content preferences.

5. Quality Control and Testing:

- Conduct rigorous testing and quality control checks to verify the functionality and accuracy of the developed software and models. Test various components, including data processing pipelines, machine learning algorithms, and user interfaces.

- Identify and address any issues or discrepancies in the data or software implementation.

6. Data Analysis and Insights Generation:

- Utilize modern data analysis tools and techniques to analyze the processed data and generate insights into social media influence on consumer behavior. Explore patterns, correlations, and trends in user engagement, sentiment dynamics, and content preferences.

- Develop algorithms and models to identify key factors driving consumer behavior on social media platforms, such as user engagement metrics, sentiment analysis, and behavioral patterns.

7. Real-time Monitoring and Visualization:

- Implement real-time monitoring and visualization tools to provide users with immediate insights into social media trends and consumer behavior patterns. Develop dashboards and visualization techniques to present the data in an intuitive and actionable format.

- Ensure that the system provides real-time alerts and notifications for significant changes or events in social media activity.

8. User Training and Documentation:

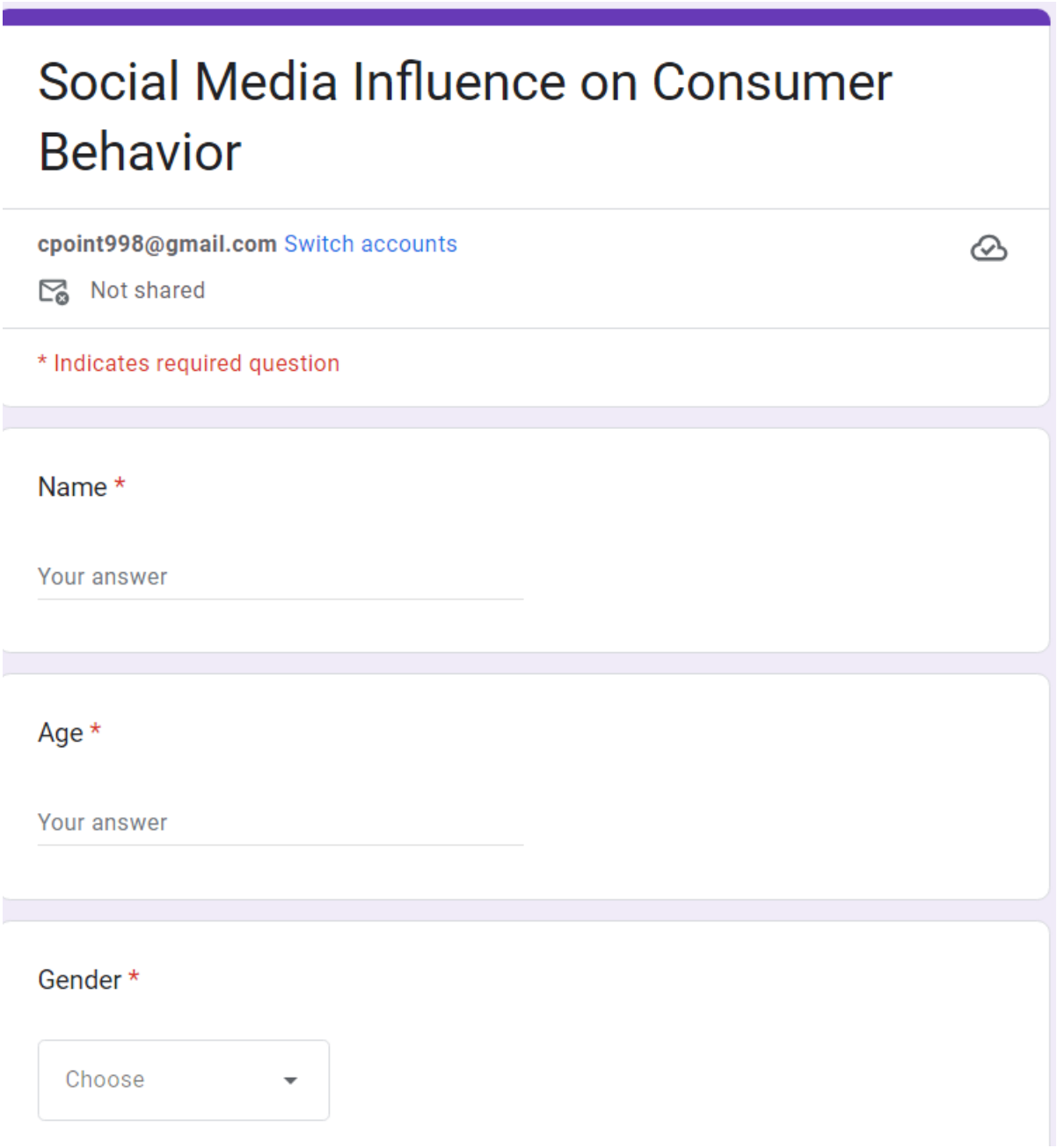
- Conduct training sessions for system users, including marketing professionals, brand managers, and market researchers. Provide detailed documentation on system operation, data interpretation, and insights generation.

- Offer ongoing support and training resources to help users effectively leverage the system for strategic decision-making and campaign optimization.

By following this implementation process, the selected design for analyzing social media influence on consumer behavior can effectively collect, process, and analyze data to provide actionable insights for businesses and marketers.

4.2 Design Drawings/Schemantics

GOOGLE FORM FOR DATA:



Hometown (City & State) *

Your answer

What category of content do you usually consume on social media? *

- [ ] Fashion & Lifestyle
- [ ] Sports
- [ ] Education
- [ ] Entertainment
- [ ] Food and Cooking
- [ ] Other:

What type of advertisement do you encounters the most? *

- ◯ pop-ups
- ◯ Short video
- ◯ Memes
- ◯ Paid promotions
- ◯ Other:

What type of advertisement attracts you the most? *

- ◯ pop-ups
- ◯ Short Video
- ◯ Memes
- ◯ Paid promotions
- ◯ Other:

What type of advertisement you dislike the most? *

- ◯ Pop-ups
- ◯ Short Video
- ◯ Memes
- ◯ Paid promotion
- ◯ Other:

To what extent does social media influence your purchasing decisions? *

- ◯ Strongly Influential
- ◯ Moderately Influential
- ◯ Slightly Influential
- ◯ Not Influential at All

Which social media platforms do you use most frequently for product recommendations and reviews? *

- ◯ Instagram
- ◯ Facebook
- ◯ Twitter
- ◯ Other

How often do you make a purchase based on a product or service recommendation you saw on social media? *

- ◯ Always
- ◯ Often
- ◯ Occasionally
- ◯ Rarely/Never

Do you follow brands or influencers on social media for information about new products or promotions? *

- ◯ Yes, frequently
- ◯ Yes, occasionally
- ◯ No, rarely
- ◯ No, never

In your opinion, how trustworthy are product reviews and recommendations on social media? *

- ◯ Very Trustworthy
- ◯ Somewhat Trustworthy
- ◯ Neutral
- ◯ Not Trustworthy at All

Have you ever changed your perception of a brand based on its social media content? *

- ◯ Yes, positively
- ◯ Yes, negatively
- ◯ No, it hasn't changed
- ◯ Not applicable

Do you engage with brands on social media through comments, likes, or shares? *

- ◯ Frequently
- ◯ Occasionally
- ◯ Rarely
- ◯ Never

How likely are you to participate in social media contests or giveaways hosted by brands? *

- ◯ Very Likely
- ◯ Likely
- ◯ Unlikely
- ◯ Very Unlikely

To what extent does the number of social media followers influence your perception of a brand's credibility? *

1 2 3 4

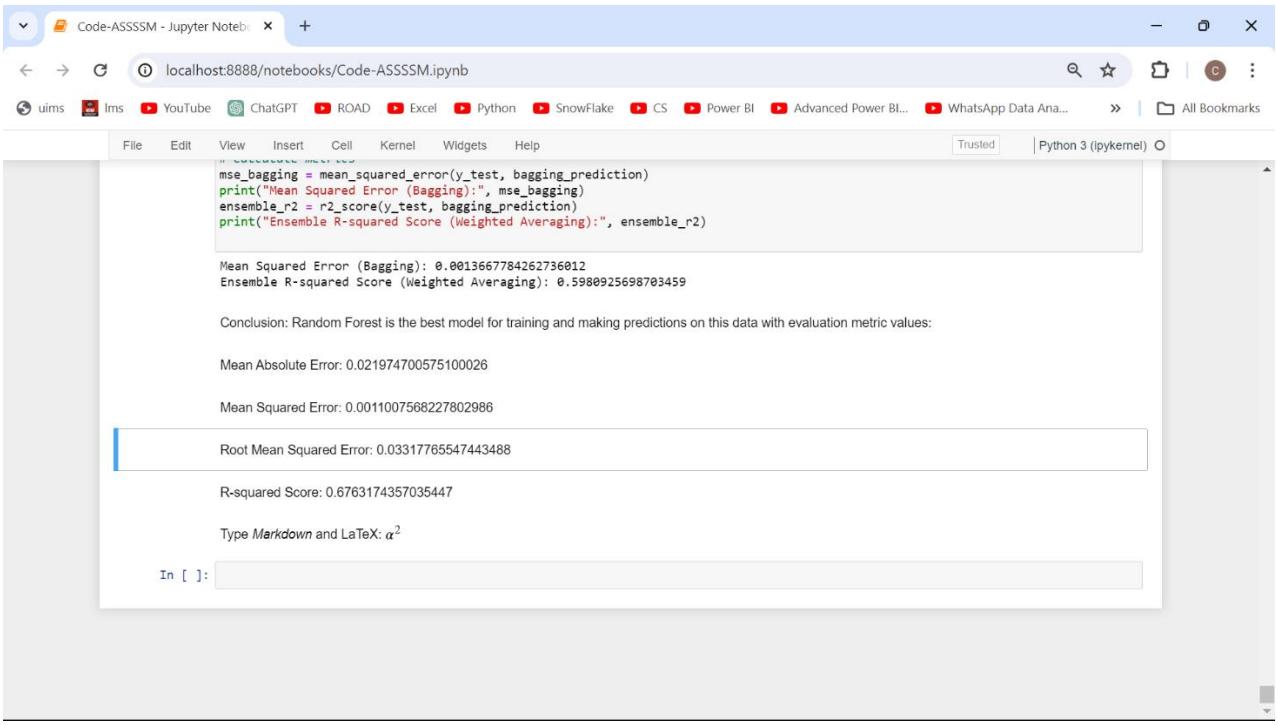
Would you be more inclined to make a purchase if a friend shared their positive experience with a product on social media? *

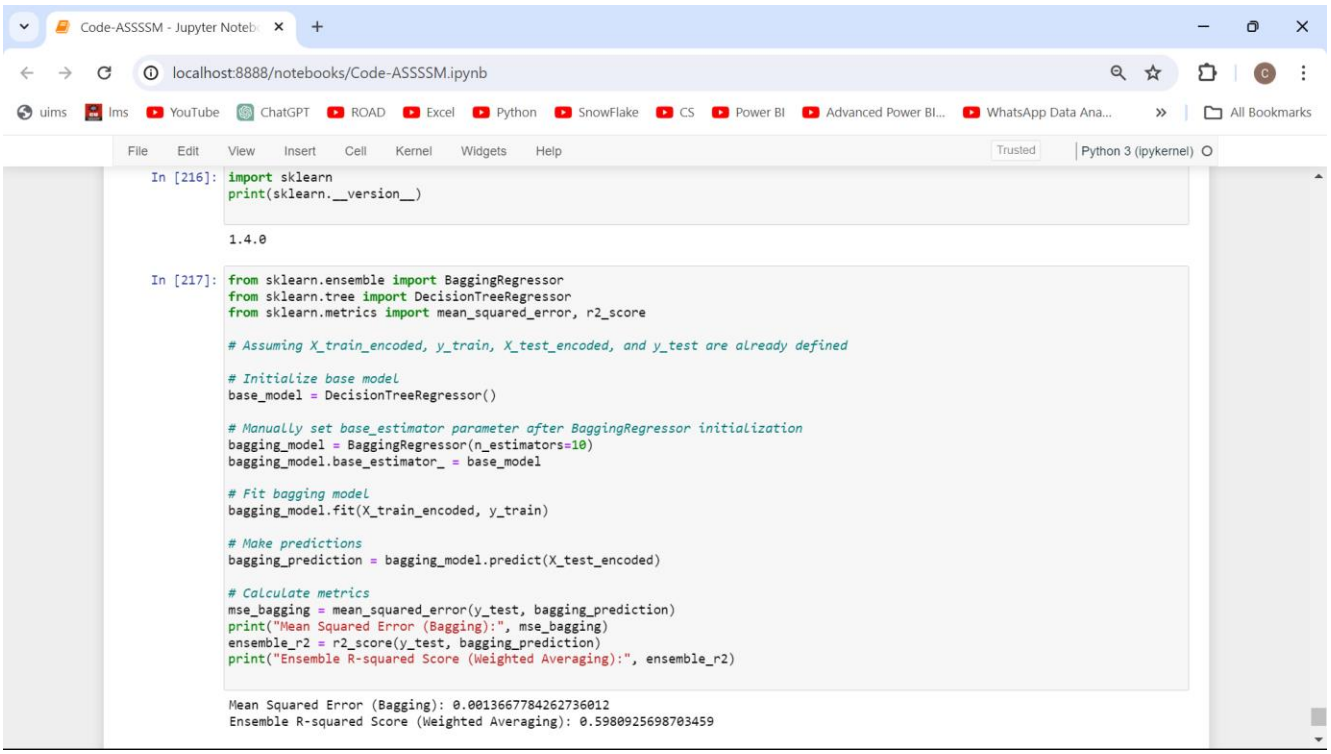
- Definitely
- Probably
- Not sure
- Definitely not

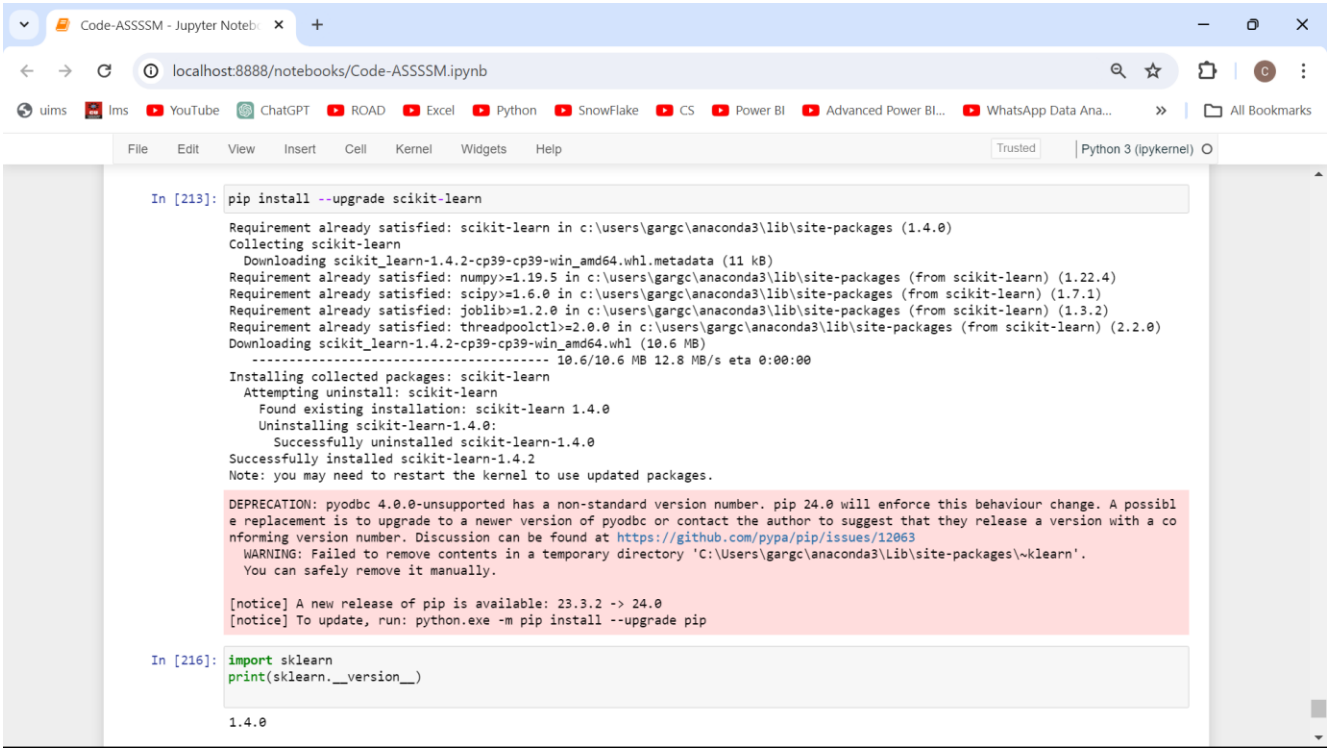
Submit Clear form

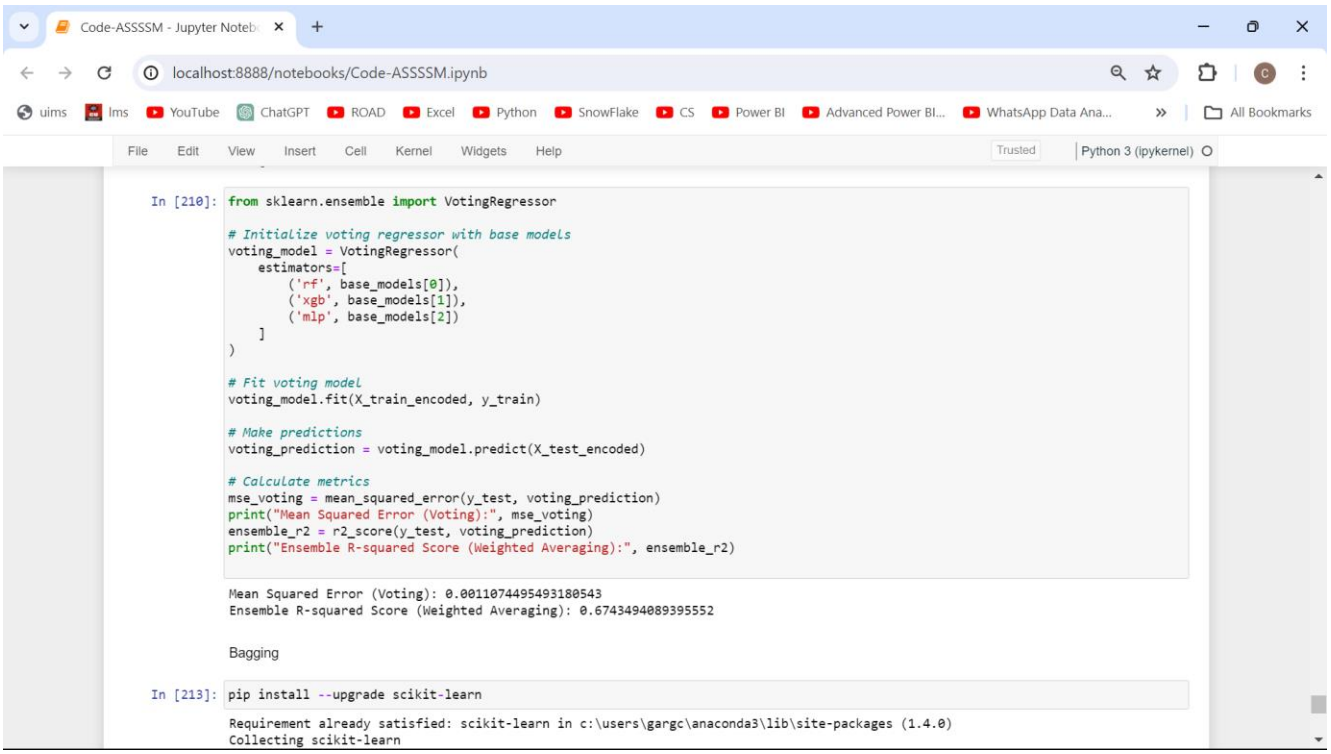
CODE :

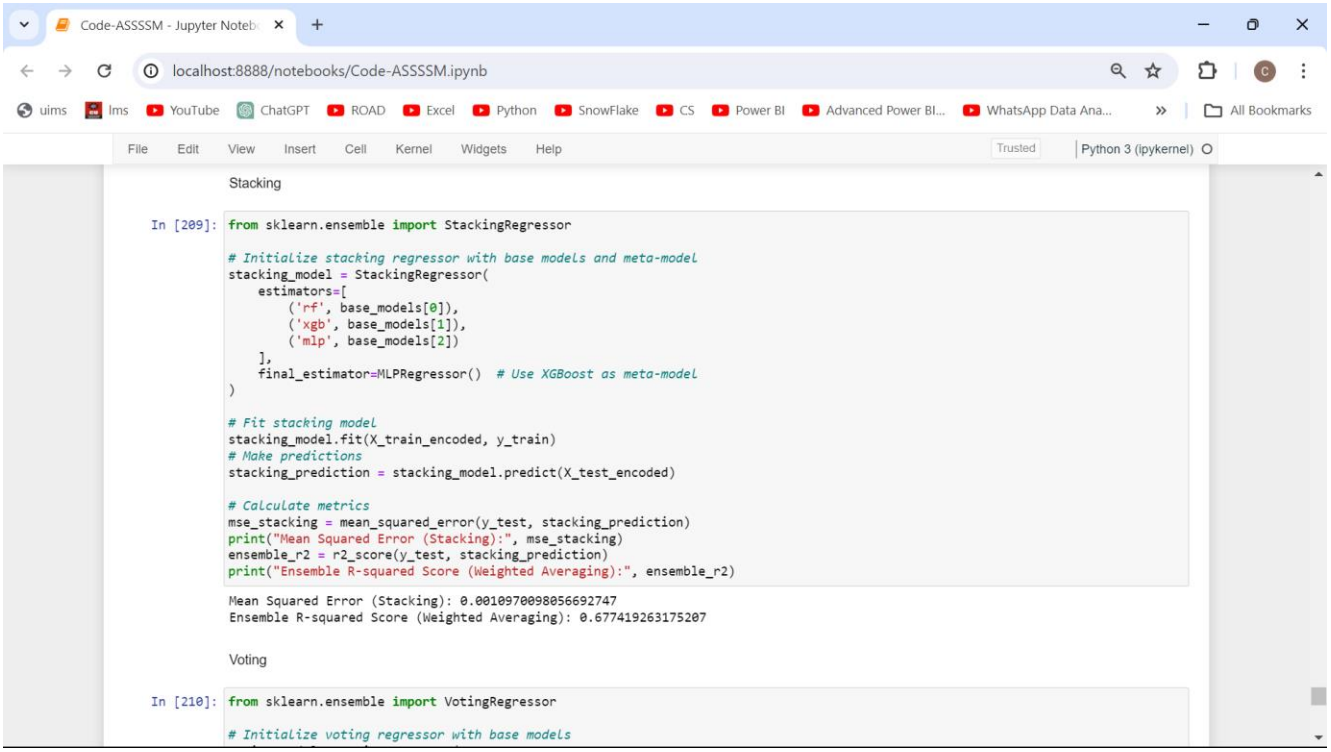
Link: https://github.com/Riviii/Unveiling-Social-Media-s-Impact-Comparative-ML-Analysis

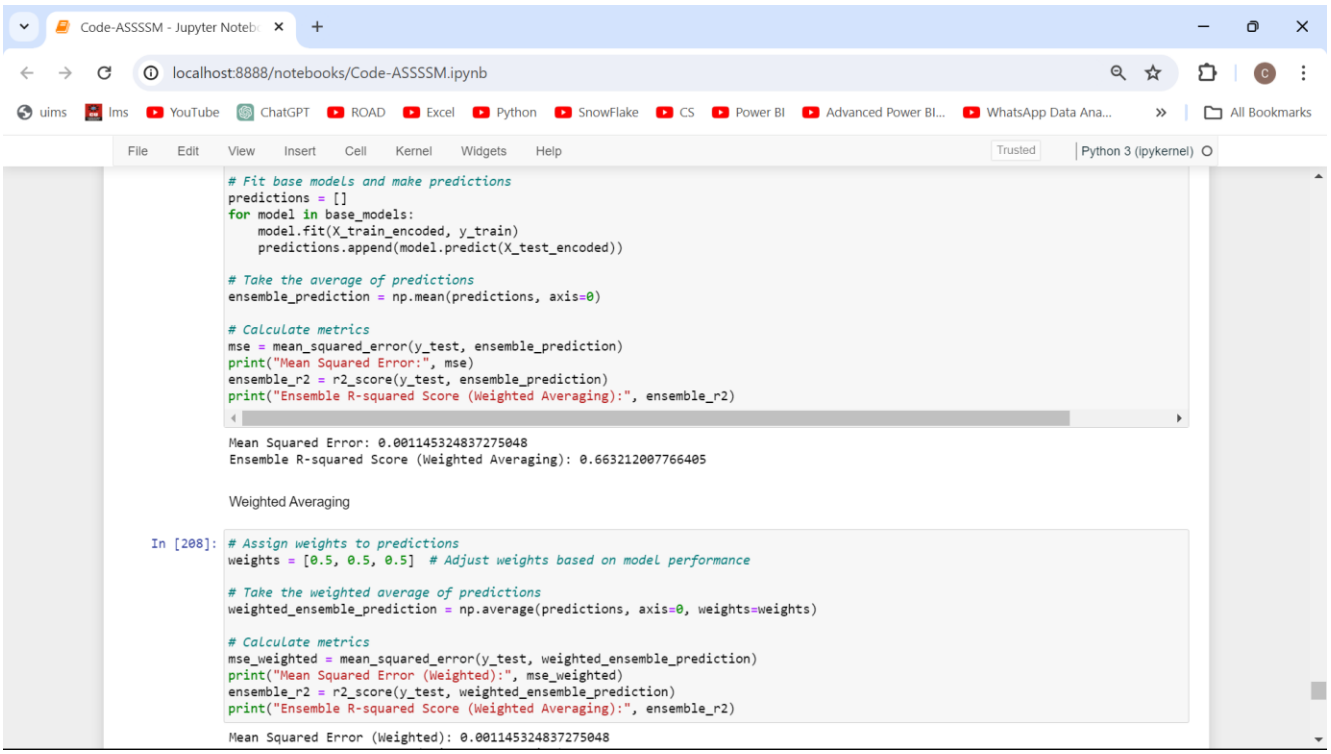


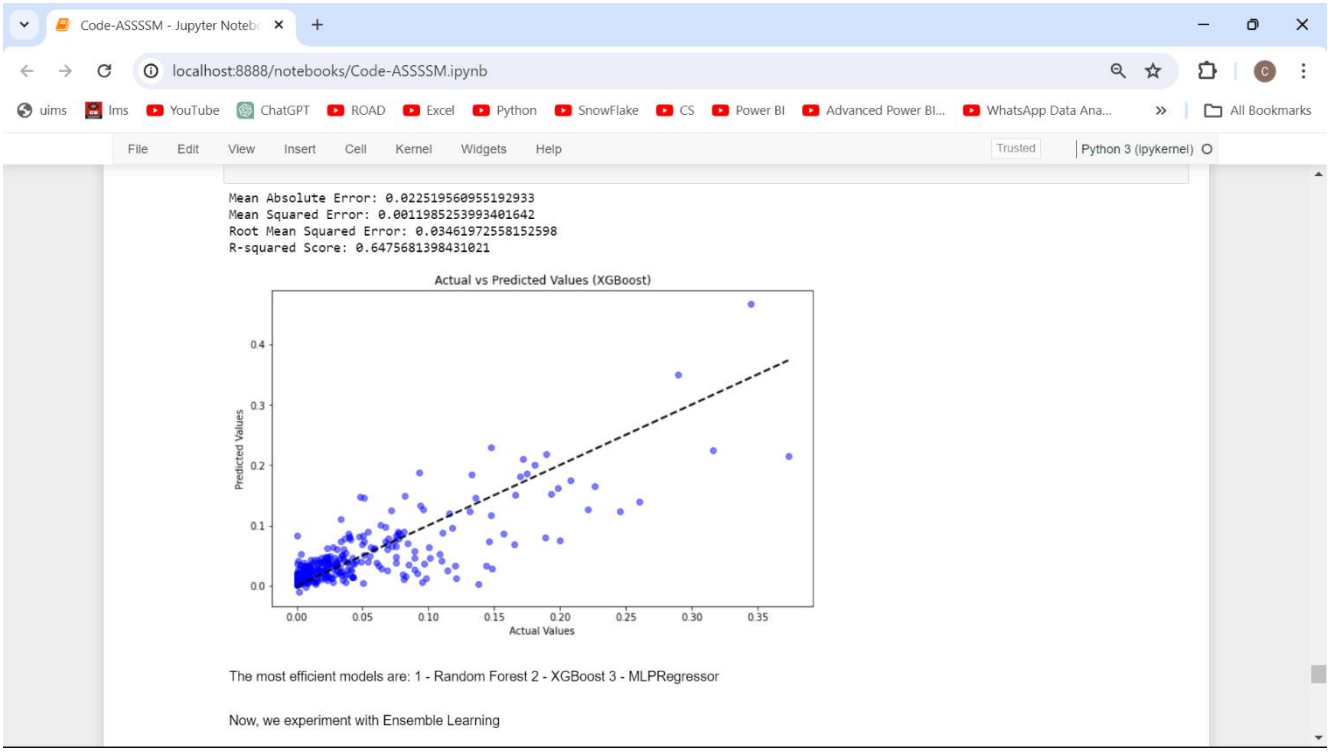

Code-ASSSSM - Jupyter Noteb
localhost:8888/notebooks/Code-ASSSSM.ipynb
uims
lms
YouTube
ChatGPT
ROAD
Excel
Python
SnowFlake
CS
Power BI
Advanced Power BI...
WhatsApp Data Ana...
All Bookmarks
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3 (ipykernel)
In [216]:
import sklearn
print(sklearn.__version__)
1.4.0
In [217]:
from sklearn.ensemble import BaggingRegressor
from sklearn.tree import DecisionTreeRegressor
from sklearn.metrics import mean_squared_error, r2_score
# Assuming X_train_encoded, y_train, X_test_encoded, and y_test are already defined
# Initialize base model
base_model = DecisionTreeRegressor()
# Manually set base_estimator parameter after BaggingRegressor initialization
bagging_model = BaggingRegressor(n_estimators=10)
bagging_model.base_estimator_ = base_model
# Fit bagging model
bagging_model.fit(X_train_encoded, y_train)
# Make predictions
bagging_prediction = bagging_model.predict(X_test_encoded)
# Calculate metrics
mse_bagging = mean_squared_error(y_test, bagging_prediction)
print("Mean Squared Error (Bagging):", mse_bagging)
ensemble_r2 = r2_score(y_test, bagging_prediction)
print("Ensemble R-squared Score (Weighted Averaging):", ensemble_r2)
Mean Squared Error (Bagging): 0.0013667784262736012
Ensemble R-squared Score (Weighted Averaging): 0.5980925698703459

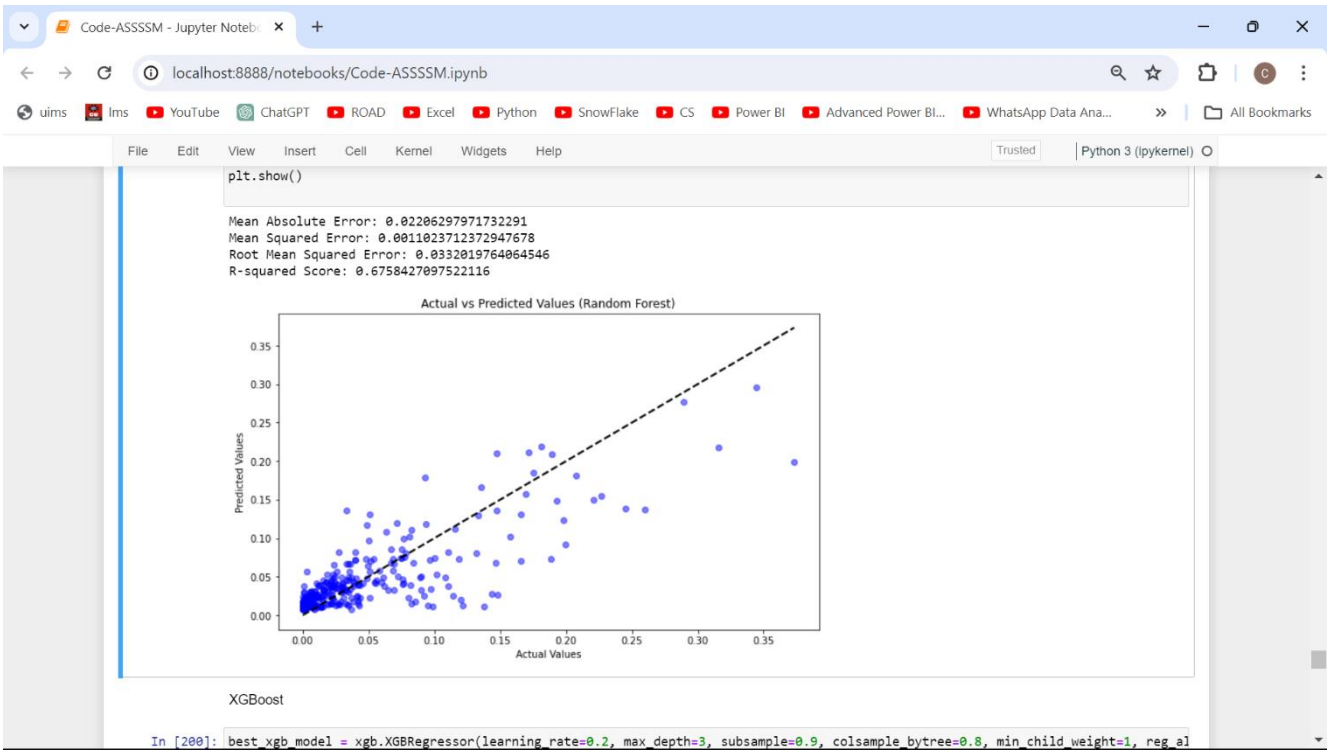

Code-ASSSSM - Jupyter Noteb
localhost:8888/notebooks/Code-ASSSSM.ipynb
uims
lms
YouTube
ChatGPT
ROAD
Excel
Python
SnowFlake
CS
Power BI
Advanced Power BI...
WhatsApp Data Ana...
All Bookmarks
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3 (ipykernel)
In [213]:
pip install --upgrade scikit-learn
Requirement already satisfied: scikit-learn in c:\users\gargc\anaconda3\lib\site-packages (1.4.0)
Collecting scikit-learn
Downloading scikit_learn-1.4.2-cp39-cp39-win_amd64.whl.metadata (11 kB)
Requirement already satisfied: numpy>=1.19.5 in c:\users\gargc\anaconda3\lib\site-packages (from scikit-learn) (1.22.4)
Requirement already satisfied: scipy>=1.6.0 in c:\users\gargc\anaconda3\lib\site-packages (from scikit-learn) (1.7.1)
Requirement already satisfied: joblib>=1.2.0 in c:\users\gargc\anaconda3\lib\site-packages (from scikit-learn) (1.3.2)
Requirement already satisfied: threadpoolctl>=2.0.0 in c:\users\gargc\anaconda3\lib\site-packages (from scikit-learn) (2.2.0)
Downloading scikit_learn-1.4.2-cp39-cp39-win_amd64.whl (10.6 MB)
10.6/10.6 MB 12.8 MB/s eta 0:00:00
Installing collected packages: scikit-learn
Attempting uninstall: scikit-learn
Found existing installation: scikit-learn 1.4.0
Uninstalling scikit-learn-1.4.0:
Successfully uninstalled scikit-learn-1.4.0
Successfully installed scikit-learn-1.4.2
Note: you may need to restart the kernel to use updated packages.
DEPRECATION: pyodbc 4.0.0-unsupported has a non-standard version number. pip 24.0 will enforce this behaviour change. A possible replacement is to upgrade to a newer version of pyodbc or contact the author to suggest that they release a version with a conforming version number. Discussion can be found at https://github.com/pypa/pip/issues/12063
WARNING: Failed to remove contents in a temporary directory 'C:\Users\gargc\anaconda3\Lib\site-packages\~klearn'.
You can safely remove it manually.
[notice] A new release of pip is available: 23.3.2 -> 24.0
[notice] To update, run: python.exe -m pip install --upgrade pip
In [216]:
import sklearn
print(sklearn.__version__)
1.4.0

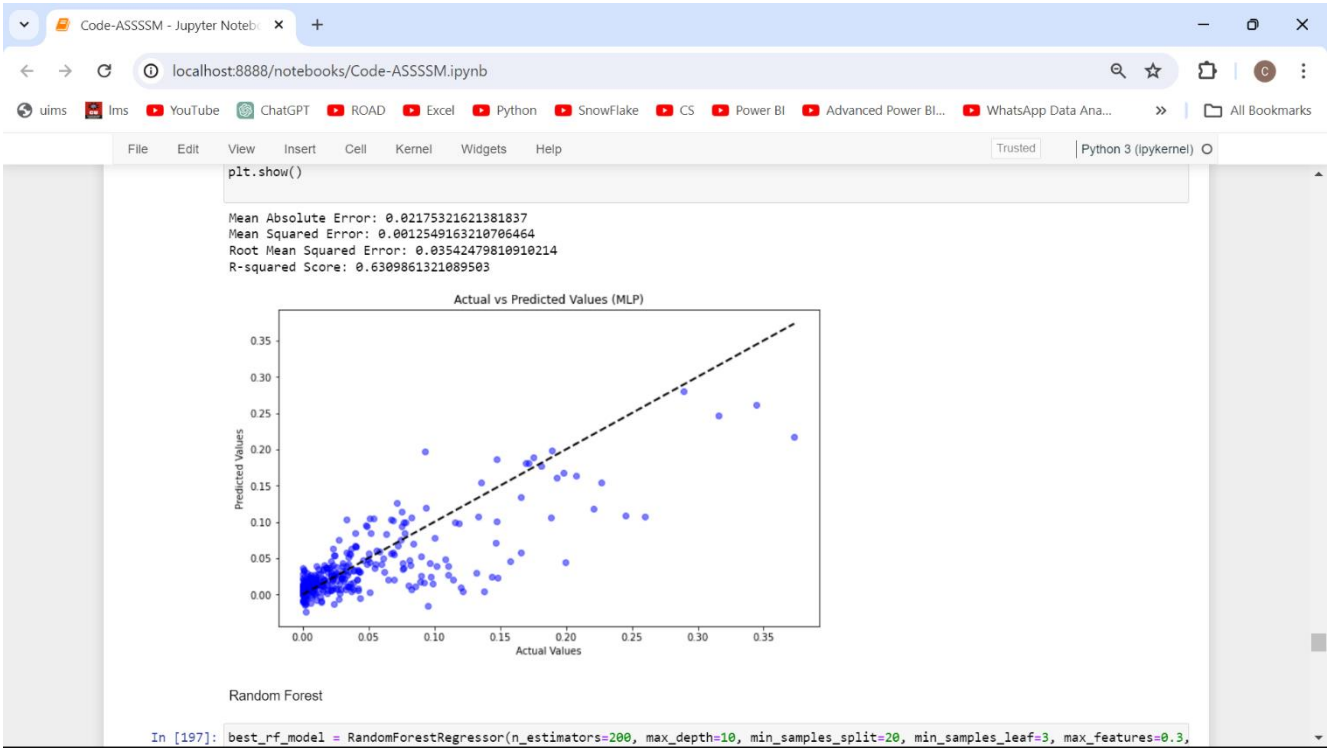

Code-ASSSSM - Jupyter Noteb
localhost:8888/notebooks/Code-ASSSSM.ipynb
uims
lms
YouTube
ChatGPT
ROAD
Excel
Python
SnowFlake
CS
Power BI
Advanced Power BI...
WhatsApp Data Ana...
All Bookmarks
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3 (ipykernel)
In [210]:
from sklearn.ensemble import VotingRegressor
# Initialize voting regressor with base models
voting_model = VotingRegressor(
estimators=[
('rf', base_models[0]),
('xgb', base_models[1]),
('mlp', base_models[2])
]
)
# Fit voting model
voting_model.fit(X_train_encoded, y_train)
# Make predictions
voting_prediction = voting_model.predict(X_test_encoded)
# Calculate metrics
mse_voting = mean_squared_error(y_test, voting_prediction)
print("Mean Squared Error (Voting):", mse_voting)
ensemble_r2 = r2_score(y_test, voting_prediction)
print("Ensemble R-squared Score (Weighted Averaging):", ensemble_r2)
Mean Squared Error (Voting): 0.0011074495493180543
Ensemble R-squared Score (Weighted Averaging): 0.6743494089395552
Bagging
In [213]:
pip install --upgrade scikit-learn
Requirement already satisfied: scikit-learn in c:\users\gargc\anaconda3\lib\site-packages (1.4.0)
Collecting scikit-learn

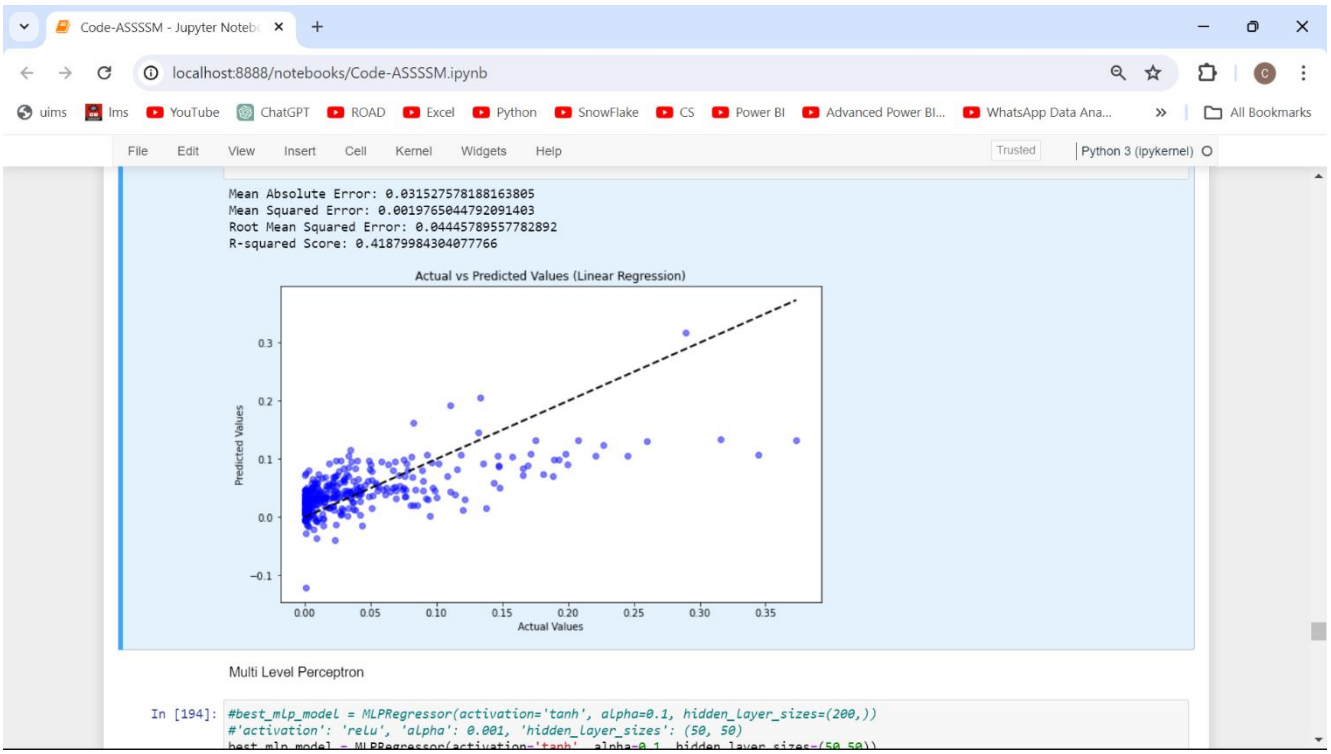

Code-ASSSSM - Jupyter Noteb
localhost:8888/notebooks/Code-ASSSSM.ipynb
uims
lms
YouTube
ChatGPT
ROAD
Excel
Python
SnowFlake
CS
Power BI
Advanced Power BI...
WhatsApp Data Ana...
All Bookmarks
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3 (ipykernel)
Stacking
In [209]:
from sklearn.ensemble import StackingRegressor
# Initialize stacking regressor with base models and meta-model
stacking_model = StackingRegressor(
estimators=[
('rf', base_models[0]),
('xgb', base_models[1]),
('mlp', base_models[2])
],
final_estimator=MLPRegressor() # Use XGBoost as meta-model
)
# Fit stacking model
stacking_model.fit(X_train_encoded, y_train)
# Make predictions
stacking_prediction = stacking_model.predict(X_test_encoded)
# Calculate metrics
mse_stacking = mean_squared_error(y_test, stacking_prediction)
print("Mean Squared Error (Stacking):", mse_stacking)
ensemble_r2 = r2_score(y_test, stacking_prediction)
print("Ensemble R-squared Score (Weighted Averaging):", ensemble_r2)
Mean Squared Error (Stacking): 0.0010970098056692747
Ensemble R-squared Score (Weighted Averaging): 0.677419263175207
Voting
In [210]:
from sklearn.ensemble import VotingRegressor
# Initialize voting regressor with base models


Code-ASSSSM - Jupyter Noteb
localhost:8888/notebooks/Code-ASSSSM.ipynb
uims
lms
YouTube
ChatGPT
ROAD
Excel
Python
SnowFlake
CS
Power BI
Advanced Power BI...
WhatsApp Data Ana...
All Bookmarks
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3 (ipykernel)
# Fit base models and make predictions
predictions = []
for model in base_models:
model.fit(X_train_encoded, y_train)
predictions.append(model.predict(X_test_encoded))
# Take the average of predictions
ensemble_prediction = np.mean(predictions, axis=0)
# Calculate metrics
mse = mean_squared_error(y_test, ensemble_prediction)
print("Mean Squared Error:", mse)
ensemble_r2 = r2_score(y_test, ensemble_prediction)
print("Ensemble R-squared Score (Weighted Averaging):", ensemble_r2)
Mean Squared Error: 0.001145324837275048
Ensemble R-squared Score (Weighted Averaging): 0.663212007766405
Weighted Averaging
In [208]:
# Assign weights to predictions
weights = [0.5, 0.5, 0.5] # Adjust weights based on model performance
# Take the weighted average of predictions
weighted_ensemble_prediction = np.average(predictions, axis=0, weights=weights)
# Calculate metrics
mse_weighted = mean_squared_error(y_test, weighted_ensemble_prediction)
print("Mean Squared Error (Weighted):", mse_weighted)
ensemble_r2 = r2_score(y_test, weighted_ensemble_prediction)
print("Ensemble R-squared Score (Weighted Averaging):", ensemble_r2)
Mean Squared Error (Weighted): 0.001145324837275048

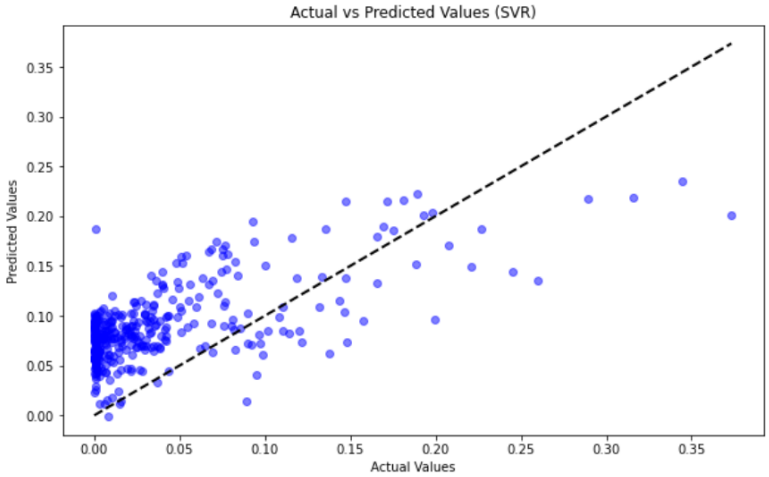

Code-ASSSSM - Jupyter Noteb
localhost:8888/notebooks/Code-ASSSSM.ipynb
uims
lms
YouTube
ChatGPT
ROAD
Excel
Python
SnowFlake
CS
Power BI
Advanced Power BI...
WhatsApp Data Ana...
All Bookmarks
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3 (ipykernel)
Mean Absolute Error: 0.022519560955192933
Mean Squared Error: 0.0011985253993401642
Root Mean Squared Error: 0.03461972558152598
R-squared Score: 0.6475681398431021
Actual vs Predicted Values (XGBoost)
Predicted Values
Actual Values
0.0
0.1
0.2
0.3
0.4
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
The most efficient models are: 1 - Random Forest 2 - XGBoost 3 - MLPRegressor
Now, we experiment with Ensemble Learning


Code-ASSSSM - Jupyter Noteb
localhost:8888/notebooks/Code-ASSSSM.ipynb
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
plt.show()
Mean Absolute Error: 0.02206297971732291
Mean Squared Error: 0.0011023712372947678
Root Mean Squared Error: 0.0332019764064546
R-squared Score: 0.6758427097522116
Actual vs Predicted Values (Random Forest)
Predicted Values
Actual Values
XGBoost
In [200]: best_xgb_model = xgb.XGBRegressor(learning_rate=0.2, max_depth=3, subsample=0.9, colsample_bytree=0.8, min_child_weight=1, reg_al

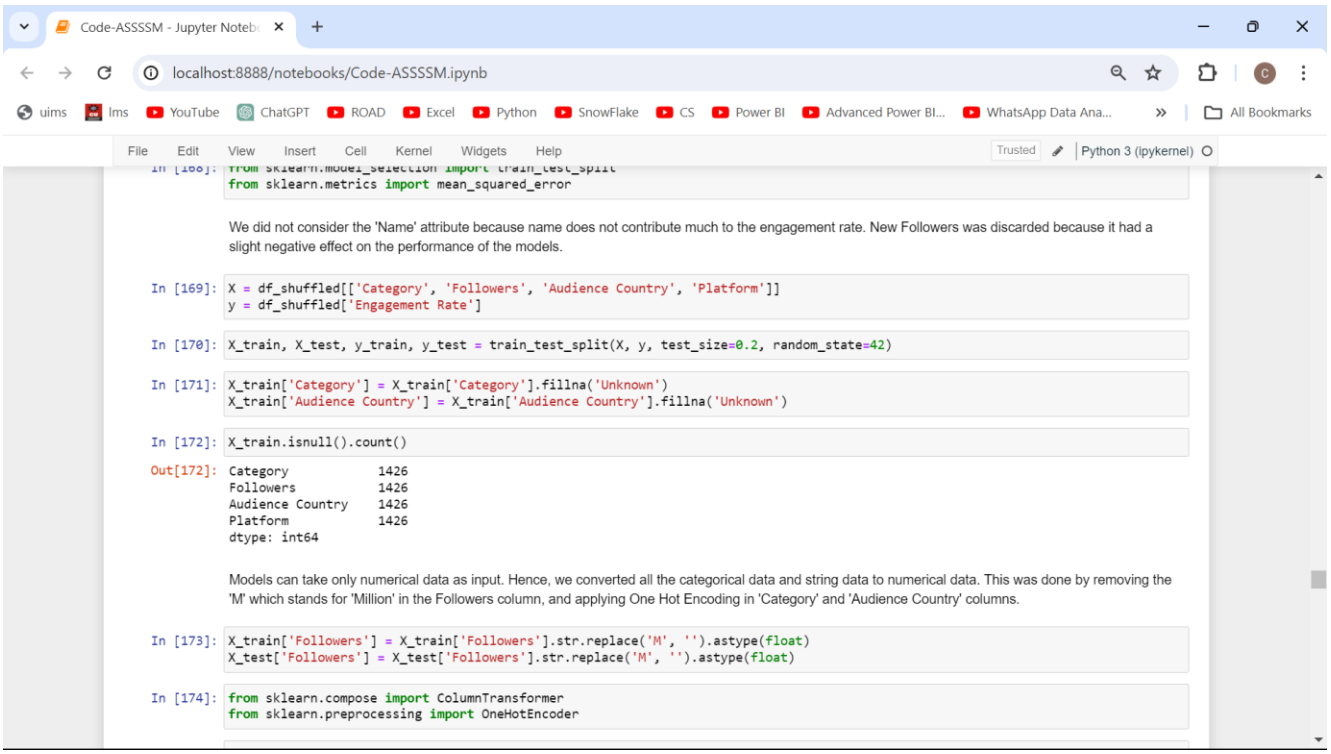

Code-ASSSSM - Jupyter Noteb
localhost:8888/notebooks/Code-ASSSSM.ipynb
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
plt.show()
Mean Absolute Error: 0.02175321621381837
Mean Squared Error: 0.0012549163210706464
Root Mean Squared Error: 0.03542479810910214
R-squared Score: 0.6309861321089503
Actual vs Predicted Values (MLP)
Predicted Values
Actual Values
Random Forest
In [197]: best_rf_model = RandomForestRegressor(n_estimators=200, max_depth=10, min_samples_split=20, min_samples_leaf=3, max_features=0.3,

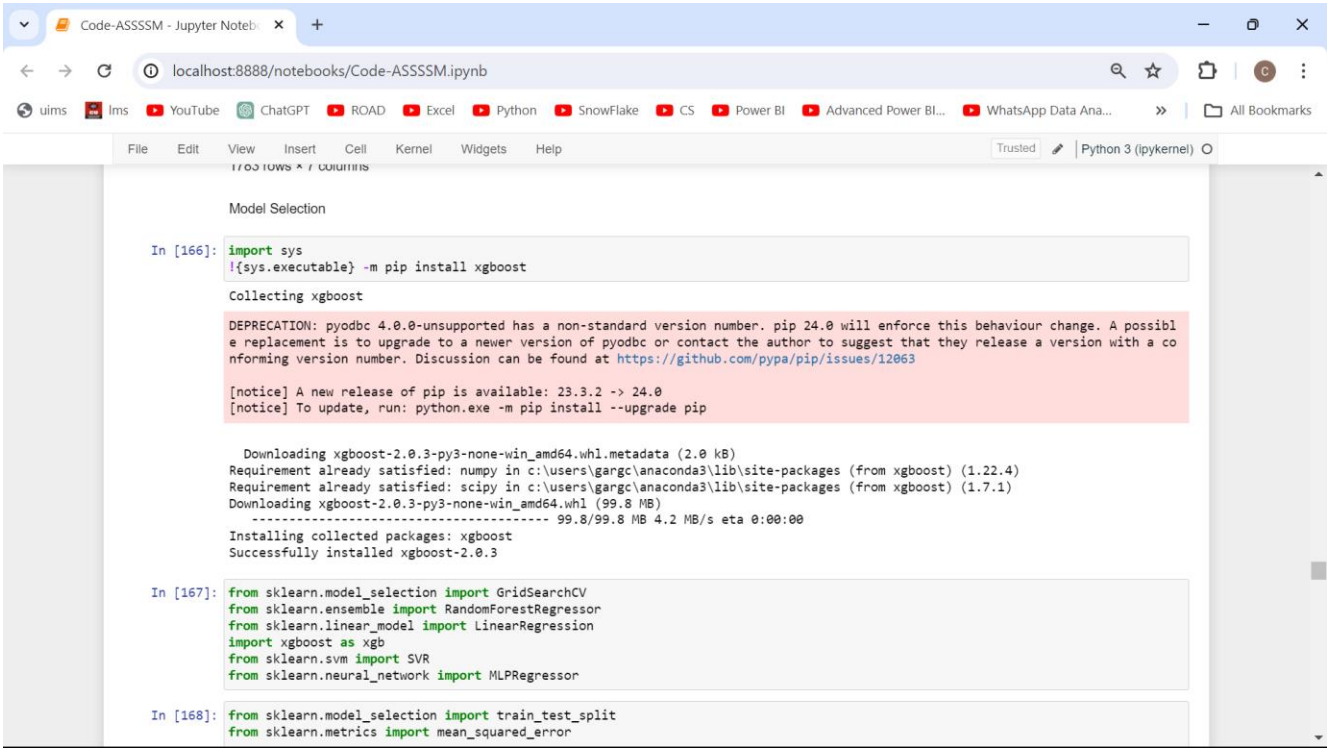


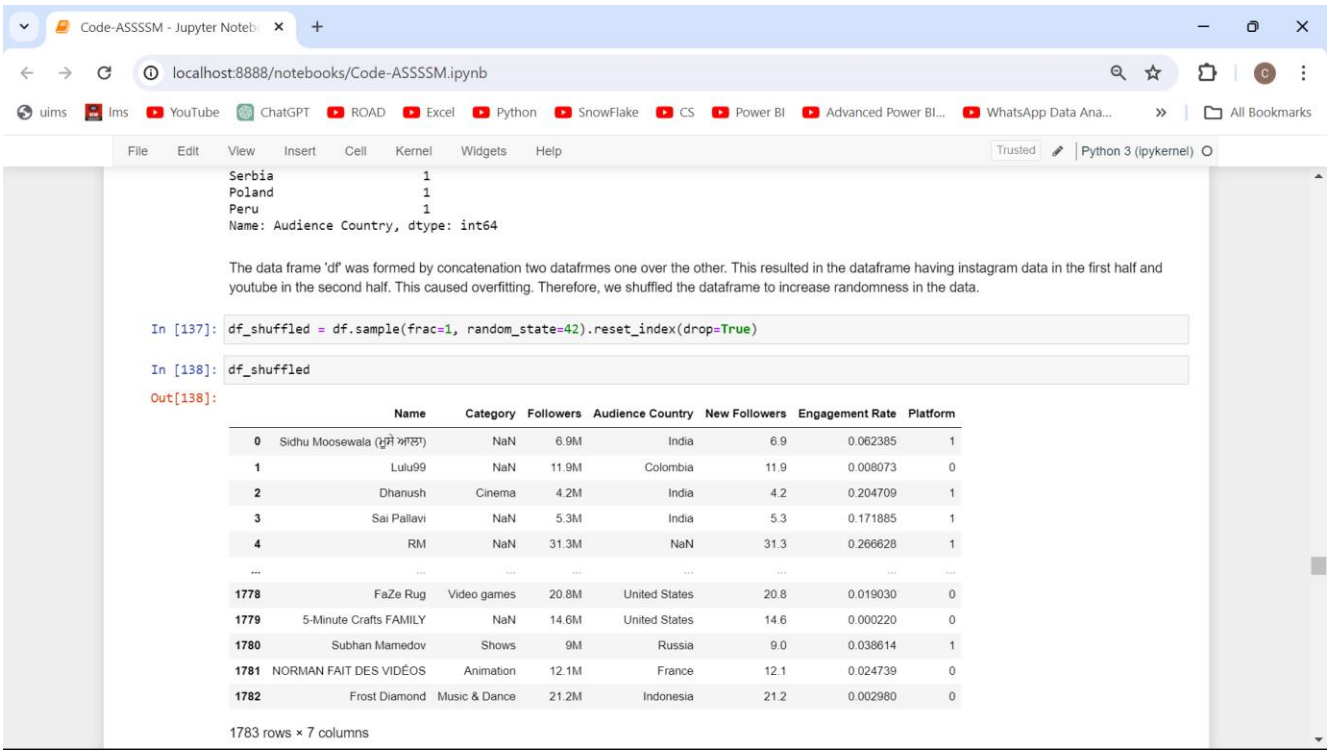

```
Mean Absolute Error: 0.06013982981618487
Mean Squared Error: 0.004377141682951148
Root Mean Squared Error: 0.06615997039714534
R-squared Score: -0.28711847603901797
```



Linear Regression


Code-ASSSSM - Jupyter Noteb
localhost:8888/notebooks/Code-ASSSSM.ipynb
uims
lms
YouTube
ChatGPT
ROAD
Excel
Python
SnowFlake
CS
Power BI
Advanced Power BI...
WhatsApp Data Ana...
All Bookmarks
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
from sklearn.metrics import mean_squared_error
We did not consider the 'Name' attribute because name does not contribute much to the engagement rate. New Followers was discarded because it had a slight negative effect on the performance of the models.
In [169]:
X = df_shuffled[['Category', 'Followers', 'Audience Country', 'Platform']]
y = df_shuffled['Engagement Rate']
In [170]:
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
In [171]:
X_train['Category'] = X_train['Category'].fillna('Unknown')
X_train['Audience Country'] = X_train['Audience Country'].fillna('Unknown')
In [172]:
X_train.isnull().count()
Out[172]:
Category 1426
Followers 1426
Audience Country 1426
Platform 1426
dtype: int64
Models can take only numerical data as input. Hence, we converted all the categorical data and string data to numerical data. This was done by removing the 'M' which stands for 'Million' in the Followers column, and applying One Hot Encoding in 'Category' and 'Audience Country' columns.
In [173]:
X_train['Followers'] = X_train['Followers'].str.replace('M', '').astype(float)
X_test['Followers'] = X_test['Followers'].str.replace('M', '').astype(float)
In [174]:
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import OneHotEncoder


Code-ASSSSM - Jupyter Noteb
localhost:8888/notebooks/Code-ASSSSM.ipynb
uims
lms
YouTube
ChatGPT
ROAD
Excel
Python
SnowFlake
CS
Power BI
Advanced Power BI...
WhatsApp Data Ana...
All Bookmarks
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Model Selection
In [166]:
import sys
!{sys.executable} -m pip install xgboost
Collecting xgboost
DEPRECATION: pyodbc 4.0.0-unsupported has a non-standard version number. pip 24.0 will enforce this behaviour change. A possible replacement is to upgrade to a newer version of pyodbc or contact the author to suggest that they release a version with a conforming version number. Discussion can be found at https://github.com/pypa/pip/issues/12063
[notice] A new release of pip is available: 23.3.2 -> 24.0
[notice] To update, run: python.exe -m pip install --upgrade pip
Downloading xgboost-2.0.3-py3-none-win_amd64.whl.metadata (2.0 kB)
Requirement already satisfied: numpy in c:\users\gargc\anaconda3\lib\site-packages (from xgboost) (1.22.4)
Requirement already satisfied: scipy in c:\users\gargc\anaconda3\lib\site-packages (from xgboost) (1.7.1)
Downloading xgboost-2.0.3-py3-none-win_amd64.whl (99.8 MB)
99.8/99.8 MB 4.2 MB/s eta 0:00:00
Installing collected packages: xgboost
Successfully installed xgboost-2.0.3
In [167]:
from sklearn.model_selection import GridSearchCV
from sklearn.ensemble import RandomForestRegressor
from sklearn.linear_model import LinearRegression
import xgboost as xgb
from sklearn.svm import SVR
from sklearn.neural_network import MLPRegressor
In [168]:
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error



```
Serbia                  1
Poland                  1
Peru                    1
Name: Audience Country, dtype: int64
```

The data frame 'df' was formed by concatenation two datafrmes one over the other. This resulted in the dataframe having instagram data in the first half and youtube in the second half. This caused overfitting. Therefore, we shuffled the dataframe to increase randomness in the data.

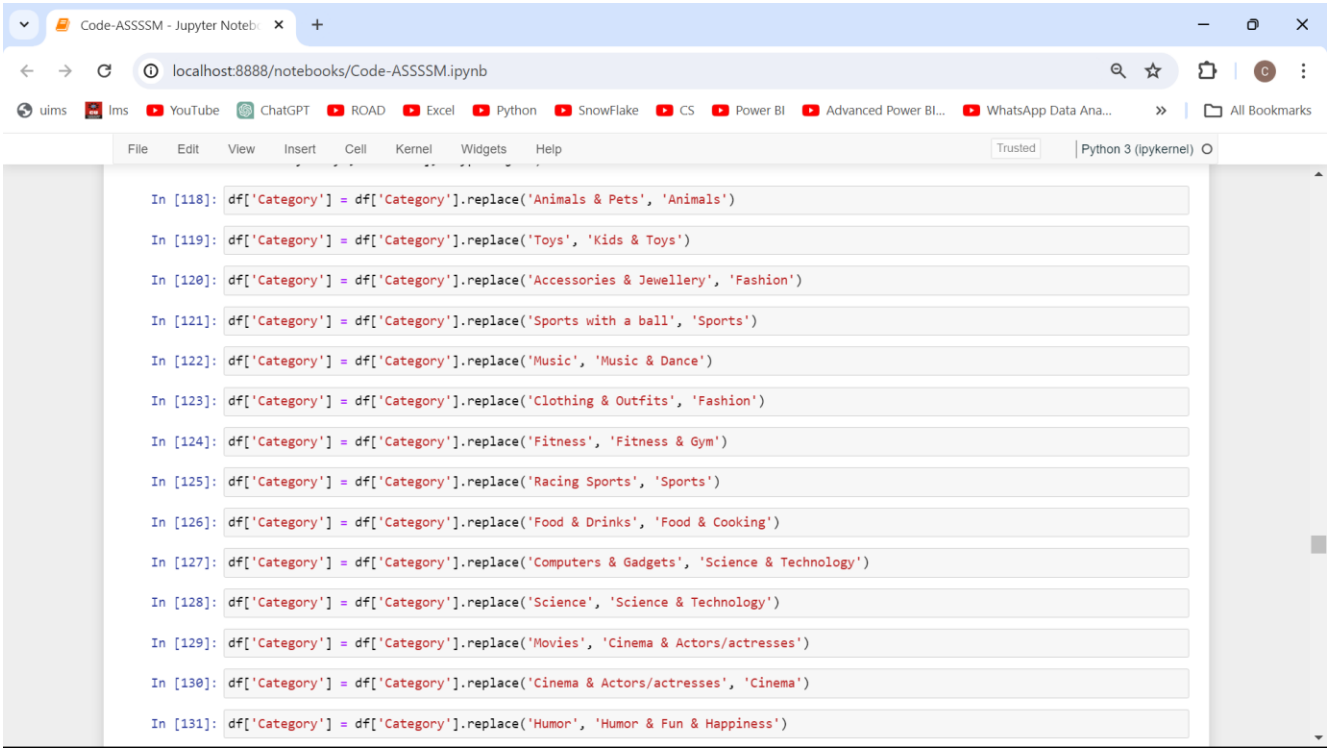
```
In [137]: df_shuffled = df.sample(frac=1, random_state=42).reset_index(drop=True)
```

```
In [138]: df_shuffled
```

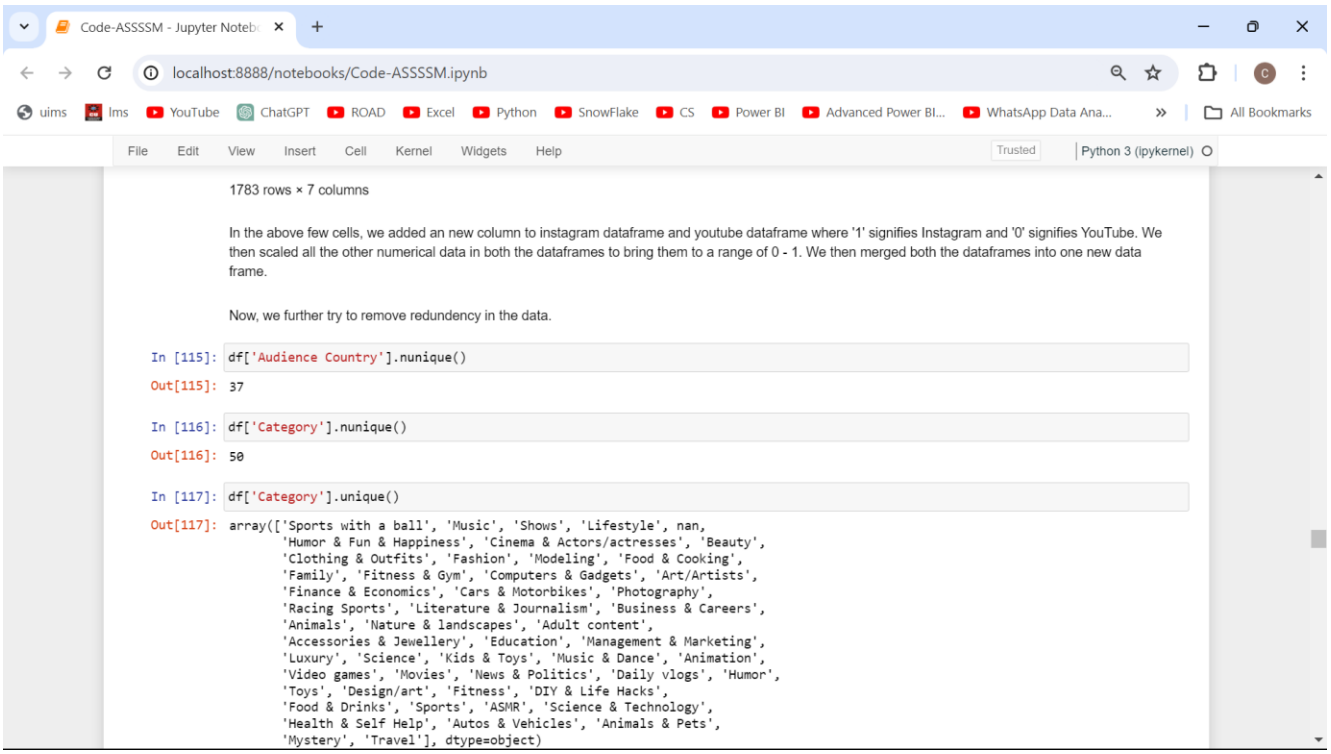
Out[138]:

	Name	Category	Followers	Audience Country	New Followers	Engagement Rate	Platform
0	Sidhu Moosewala (ਮੂਸੇ ਆਲਾ)	NaN	6.9M	India	6.9	0.062385	1
1	Lulu99	NaN	11.9M	Colombia	11.9	0.008073	0
2	Dhanush	Cinema	4.2M	India	4.2	0.204709	1
3	Sai Pallavi	NaN	5.3M	India	5.3	0.171885	1
4	RM	NaN	31.3M	NaN	31.3	0.266628	1
...	...	...	...	...	...	...	...
1778	FaZe Rug	Video games	20.8M	United States	20.8	0.019030	0
1779	5-Minute Crafts FAMILY	NaN	14.6M	United States	14.6	0.000220	0
1780	Subhan Mamedov	Shows	9M	Russia	9.0	0.038614	1
1781	NORMAN FAIT DES VIDÉOS	Animation	12.1M	France	12.1	0.024739	0
1782	Frost Diamond	Music & Dance	21.2M	Indonesia	21.2	0.002980	0

1783 rows × 7 columns



```
In [118]: df['Category'] = df['Category'].replace('Animals & Pets', 'Animals')
In [119]: df['Category'] = df['Category'].replace('Toys', 'Kids & Toys')
In [120]: df['Category'] = df['Category'].replace('Accessories & Jewellery', 'Fashion')
In [121]: df['Category'] = df['Category'].replace('Sports with a ball', 'Sports')
In [122]: df['Category'] = df['Category'].replace('Music', 'Music & Dance')
In [123]: df['Category'] = df['Category'].replace('Clothing & Outfits', 'Fashion')
In [124]: df['Category'] = df['Category'].replace('Fitness', 'Fitness & Gym')
In [125]: df['Category'] = df['Category'].replace('Racing Sports', 'Sports')
In [126]: df['Category'] = df['Category'].replace('Food & Drinks', 'Food & Cooking')
In [127]: df['Category'] = df['Category'].replace('Computers & Gadgets', 'Science & Technology')
In [128]: df['Category'] = df['Category'].replace('Science', 'Science & Technology')
In [129]: df['Category'] = df['Category'].replace('Movies', 'Cinema & Actors/actresses')
In [130]: df['Category'] = df['Category'].replace('Cinema & Actors/actresses', 'Cinema')
In [131]: df['Category'] = df['Category'].replace('Humor', 'Humor & Fun & Happiness')
```



1783 rows × 7 columns

In the above few cells, we added an new column to instagram dataframe and youtube dataframe where '1' signifies Instagram and '0' signifies YouTube. We then scaled all the other numerical data in both the dataframes to bring them to a range of 0 - 1. We then merged both the dataframes into one new data frame.

Now, we further try to remove redundency in the data.

In [115]:
```
df['Audience Country'].nunique()
```

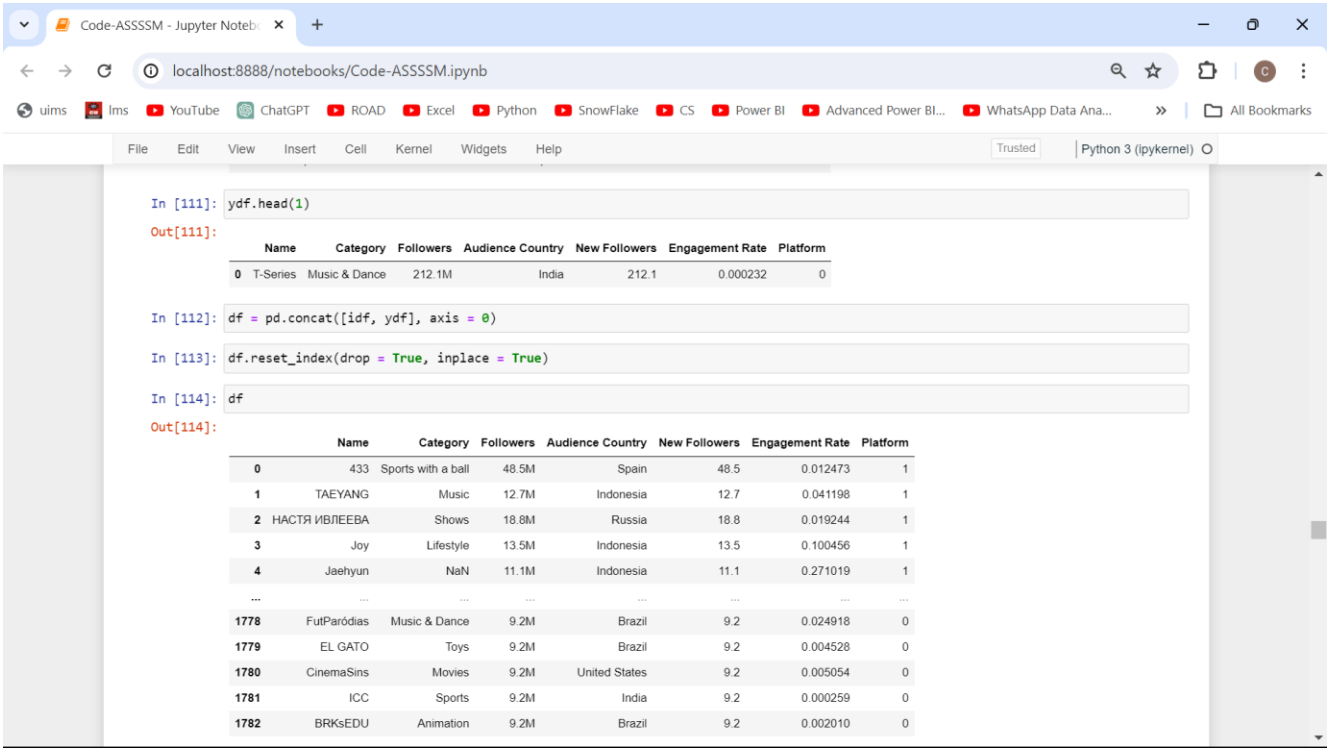
Out[115]:
```
37
```

In [116]:
```
df['Category'].nunique()
```

Out[116]:
```
50
```

In [117]:
```
df['Category'].unique()
```

Out[117]:
```
array(['Sports with a ball', 'Music', 'Shows', 'Lifestyle', nan,
       'Humor & Fun & Happiness', 'Cinema & Actors/actresses', 'Beauty',
       'Clothing & Outfits', 'Fashion', 'Modeling', 'Food & Cooking',
       'Family', 'Fitness & Gym', 'Computers & Gadgets', 'Art/Artists',
       'Finance & Economics', 'Cars & Motorbikes', 'Photography',
       'Racing Sports', 'Literature & Journalism', 'Business & Careers',
       'Animals', 'Nature & landscapes', 'Adult content',
       'Accessories & Jewellery', 'Education', 'Management & Marketing',
       'Luxury', 'Science', 'Kids & Toys', 'Music & Dance', 'Animation',
       'Video games', 'Movies', 'News & Politics', 'Daily vlogs', 'Humor',
       'Toys', 'Design/art', 'Fitness', 'DIY & Life Hacks',
       'Food & Drinks', 'Sports', 'ASMR', 'Science & Technology',
       'Health & Self Help', 'Autos & Vehicles', 'Animals & Pets',
       'Mystery', 'Travel'], dtype=object)
```



In [111]:
```
ydf.head(1)
```

Out[111]:

	Name	Category	Followers	Audience Country	New Followers	Engagement Rate	Platform
0	T-Series	Music & Dance	212.1M	India	212.1	0.000232	0

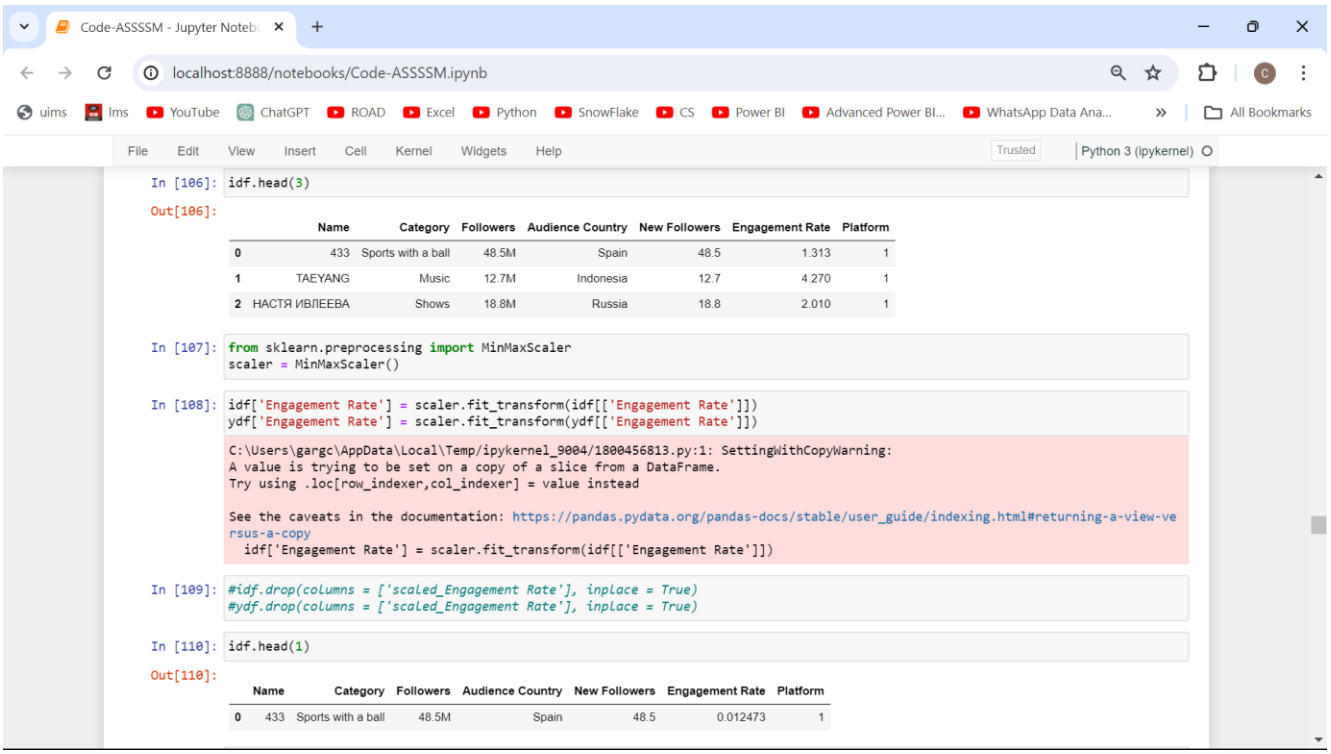
In [112]:
```
df = pd.concat([idf, ydf], axis = 0)
```

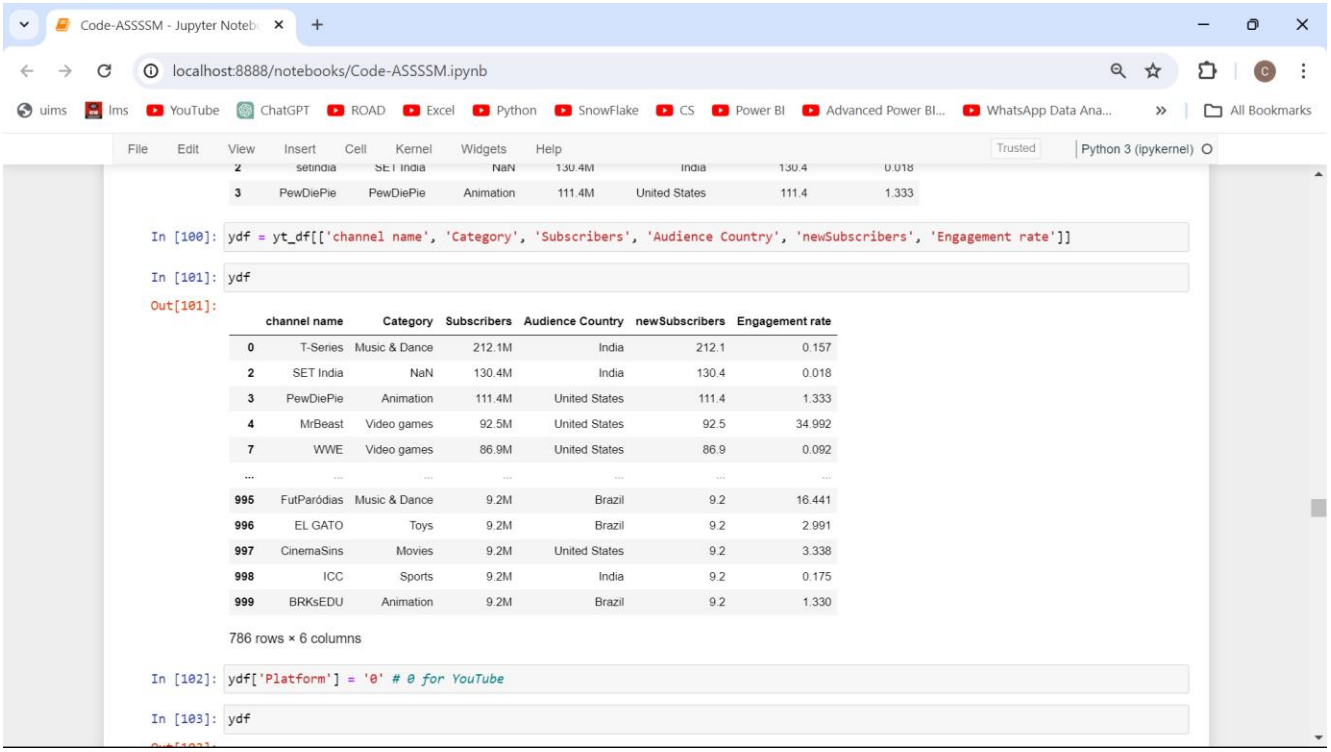
In [113]:
```
df.reset_index(drop = True, inplace = True)
```

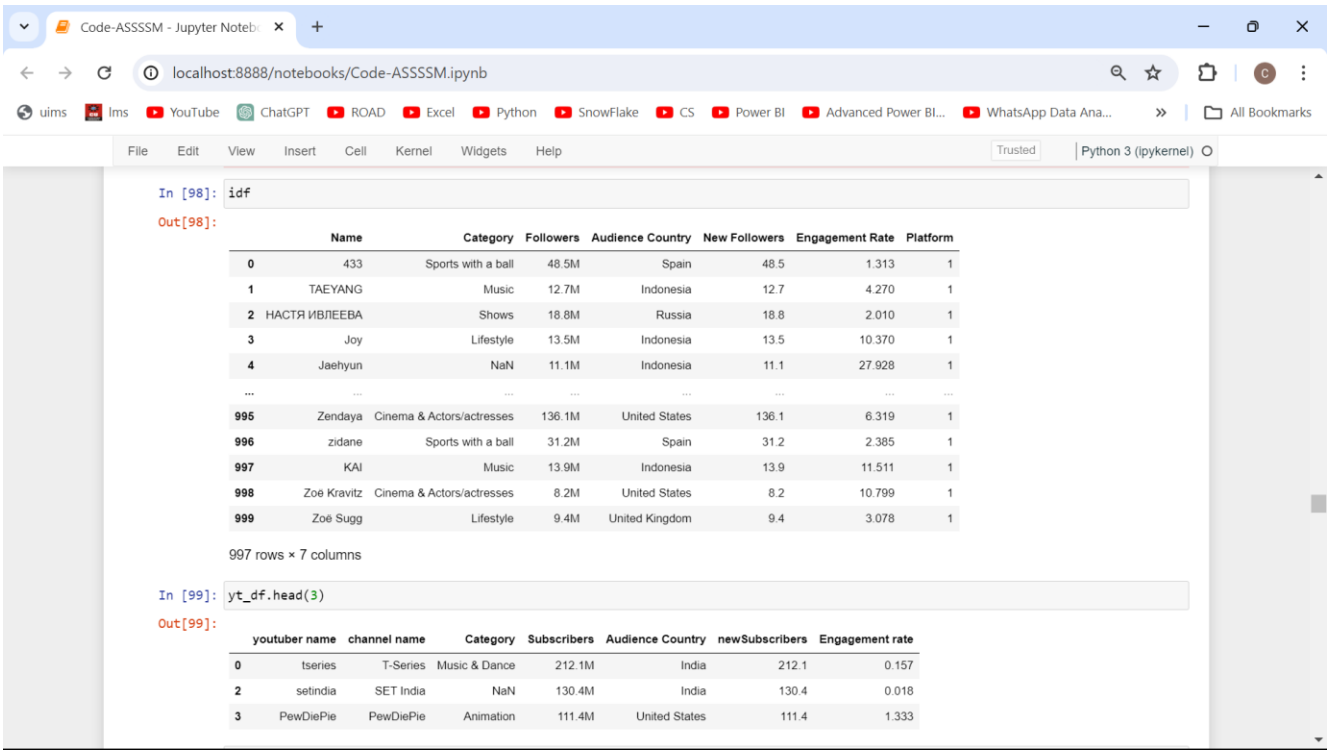
In [114]:
```
df
```

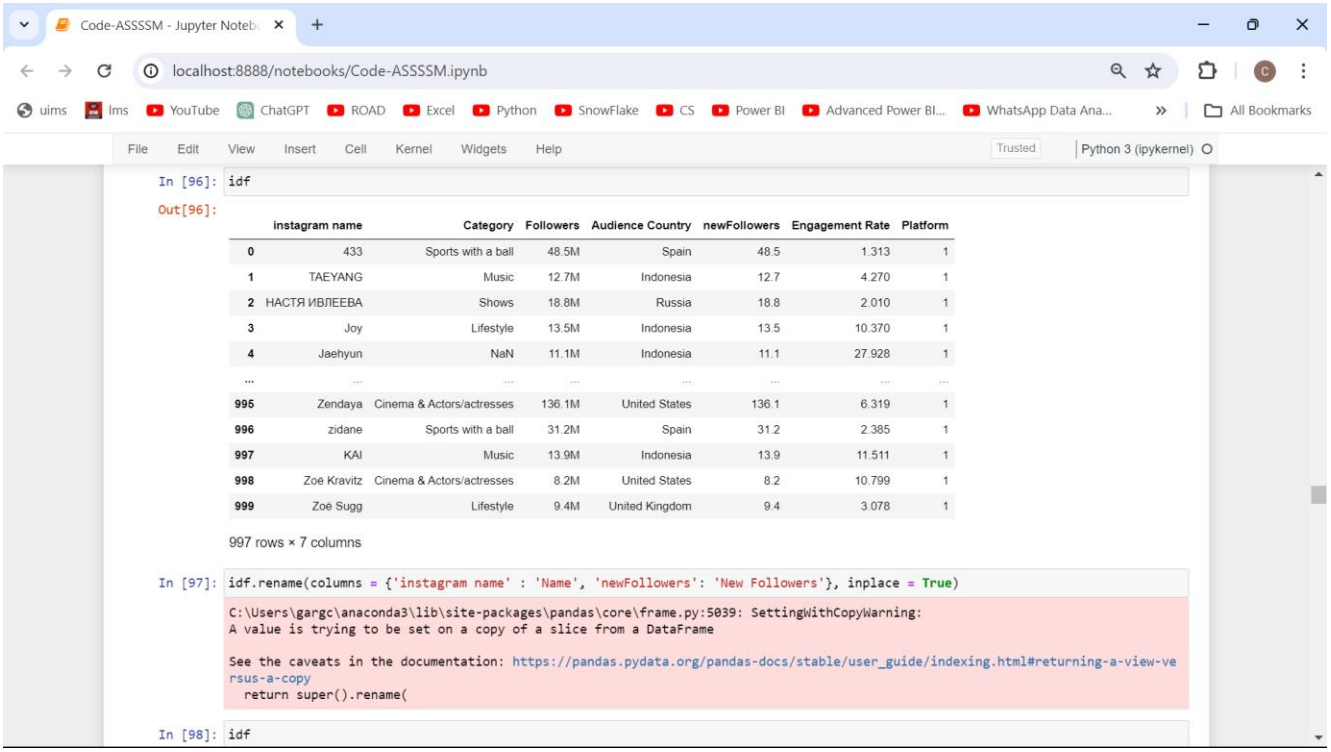
Out[114]:

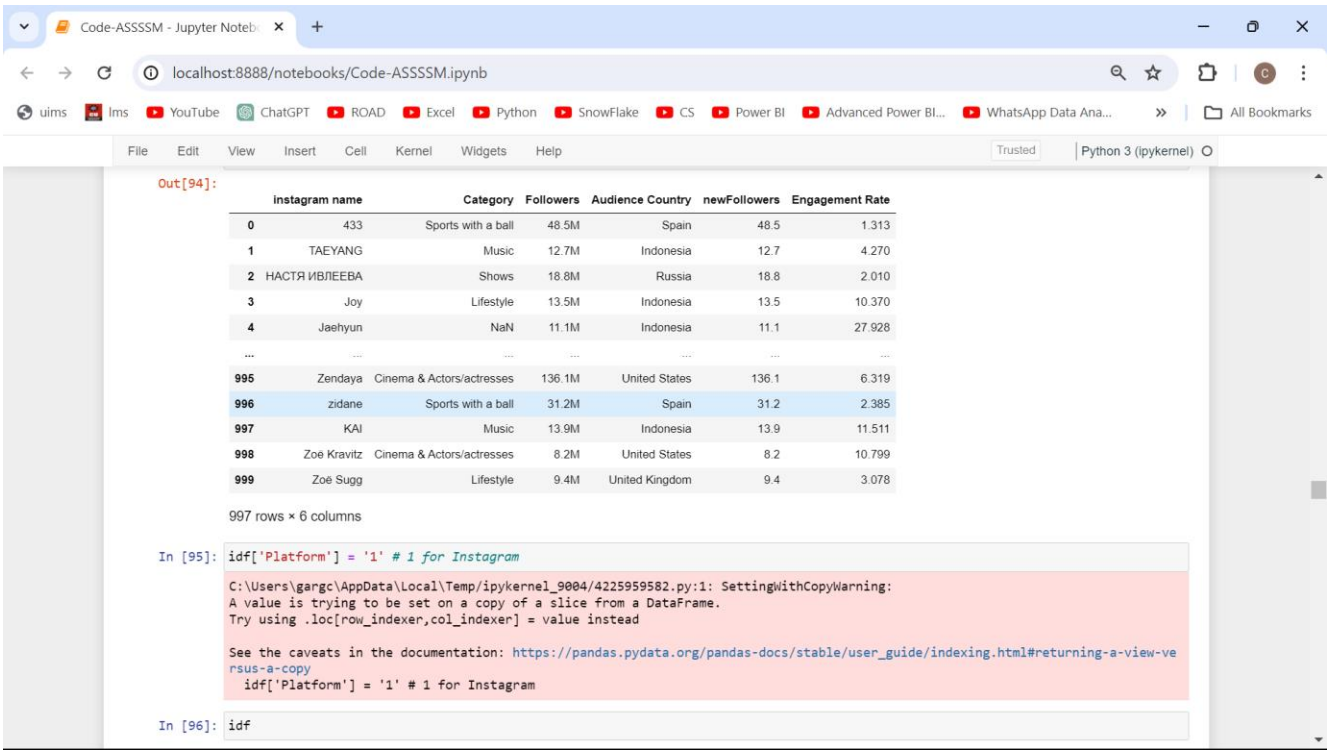
	Name	Category	Followers	Audience Country	New Followers	Engagement Rate	Platform
0	433	Sports with a ball	48.5M	Spain	48.5	0.012473	1
1	TAEYANG	Music	12.7M	Indonesia	12.7	0.041198	1
2	НАСТЯ ИВЛЕЕВА	Shows	18.8M	Russia	18.8	0.019244	1
3	Joy	Lifestyle	13.5M	Indonesia	13.5	0.100456	1
4	Jaehyun	NaN	11.1M	Indonesia	11.1	0.271019	1
...	...	...	...	...	...	...	...
1778	FutParódias	Music & Dance	9.2M	Brazil	9.2	0.024918	0
1779	EL GATO	Toys	9.2M	Brazil	9.2	0.004528	0
1780	CinemaSins	Movies	9.2M	United States	9.2	0.005054	0
1781	ICC	Sports	9.2M	India	9.2	0.000259	0
1782	BRKsEDU	Animation	9.2M	Brazil	9.2	0.002010	0

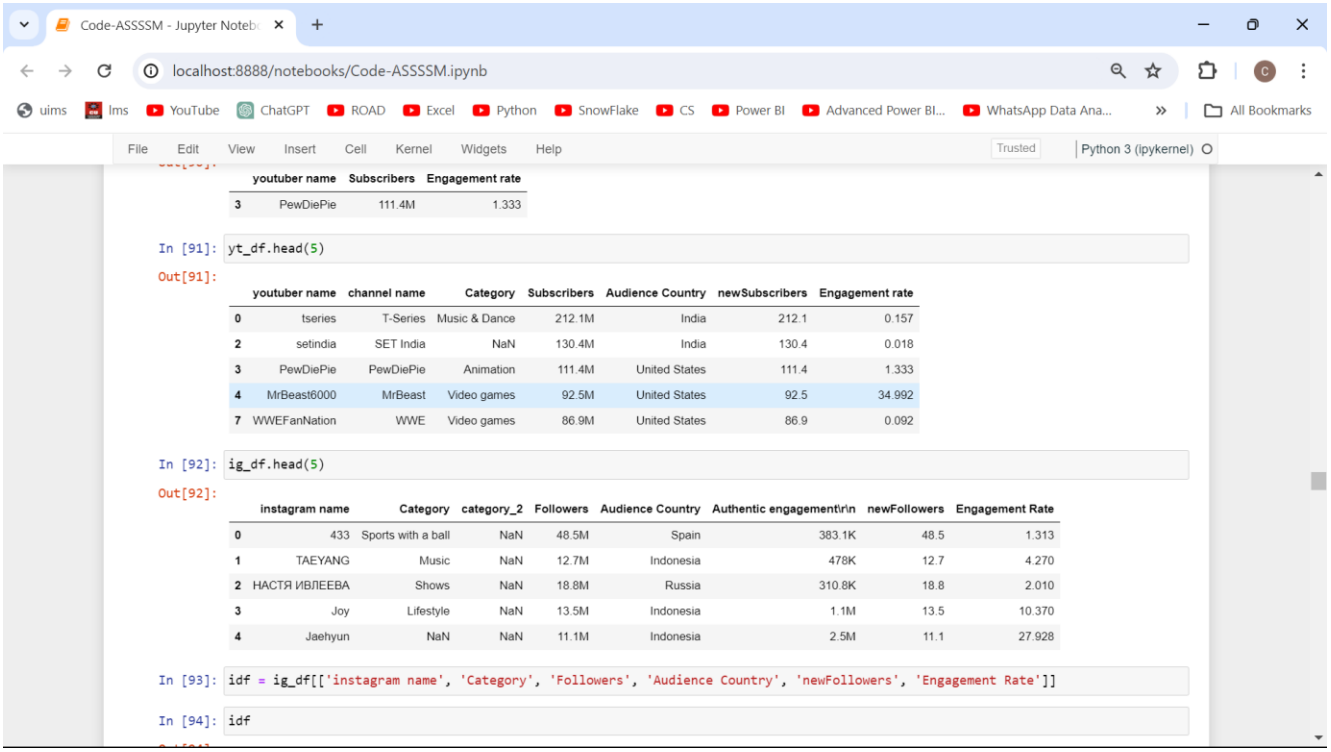

Code-ASSSSM - Jupyter Noteb
localhost:8888/notebooks/Code-ASSSSM.ipynb
uims
Ims
YouTube
ChatGPT
ROAD
Excel
Python
SnowFlake
CS
Power BI
Advanced Power BI...
WhatsApp Data Ana...
All Bookmarks
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3 (ipykernel)
In [106]: idf.head(3)
Out[106]:
Name Category Followers Audience Country New Followers Engagement Rate Platform
0 433 Sports with a ball 48.5M Spain 48.5 1.313 1
1 TAEYANG Music 12.7M Indonesia 12.7 4.270 1
2 НАСТЯ ИВЛЕЕВА Shows 18.8M Russia 18.8 2.010 1
In [107]: from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
In [108]: idf['Engagement Rate'] = scaler.fit_transform(idf[['Engagement Rate']])
ydf['Engagement Rate'] = scaler.fit_transform(ydf[['Engagement Rate']])
C:\Users\gargc\AppData\Local\Temp/ipykernel_9004/1800456813.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
idf['Engagement Rate'] = scaler.fit_transform(idf[['Engagement Rate']])
In [109]: #idf.drop(columns = ['scaled_Engagement Rate'], inplace = True)
#ydf.drop(columns = ['scaled_Engagement Rate'], inplace = True)
In [110]: idf.head(1)
Out[110]:
Name Category Followers Audience Country New Followers Engagement Rate Platform
0 433 Sports with a ball 48.5M Spain 48.5 0.012473 1

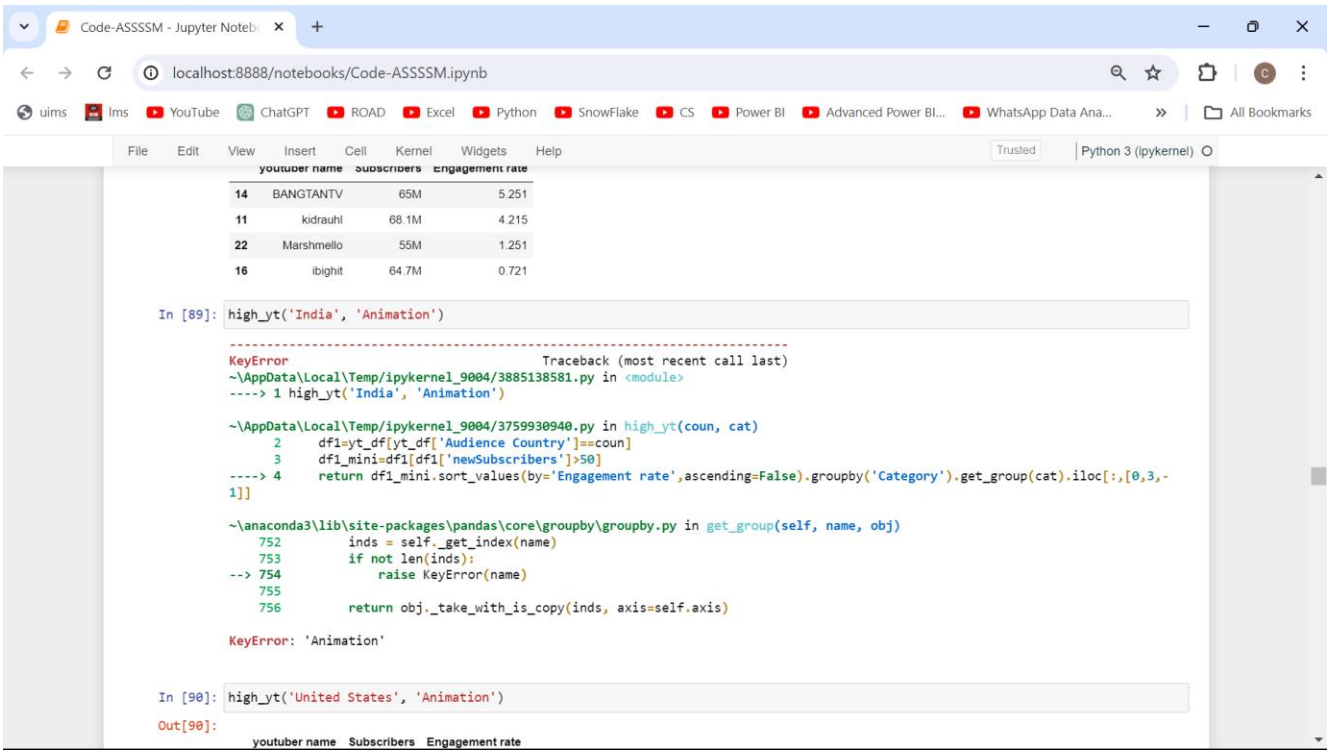

Code-ASSSSM - Jupyter Noteb
localhost:8888/notebooks/Code-ASSSSM.ipynb
uims
Ims
YouTube
ChatGPT
ROAD
Excel
Python
SnowFlake
CS
Power BI
Advanced Power BI...
WhatsApp Data Ana...
All Bookmarks
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3 (ipykernel)
3 PewDiePie PewDiePie Animation 111.4M United States 111.4 1.333
In [100]: ydf = yt_df[['channel name', 'Category', 'Subscribers', 'Audience Country', 'newSubscribers', 'Engagement rate']]
In [101]: ydf
Out[101]:
channel name Category Subscribers Audience Country newSubscribers Engagement rate
0 T-Series Music & Dance 212.1M India 212.1 0.157
2 SET India NaN 130.4M India 130.4 0.018
3 PewDiePie Animation 111.4M United States 111.4 1.333
4 MrBeast Video games 92.5M United States 92.5 34.992
7 WWE Video games 86.9M United States 86.9 0.092
995 FutParódias Music & Dance 9.2M Brazil 9.2 16.441
996 EL GATO Toys 9.2M Brazil 9.2 2.991
997 CinemaSins Movies 9.2M United States 9.2 3.338
998 ICC Sports 9.2M India 9.2 0.175
999 BRKsEDU Animation 9.2M Brazil 9.2 1.330
786 rows × 6 columns
In [102]: ydf['Platform'] = '0' # 0 for YouTube
In [103]: ydf

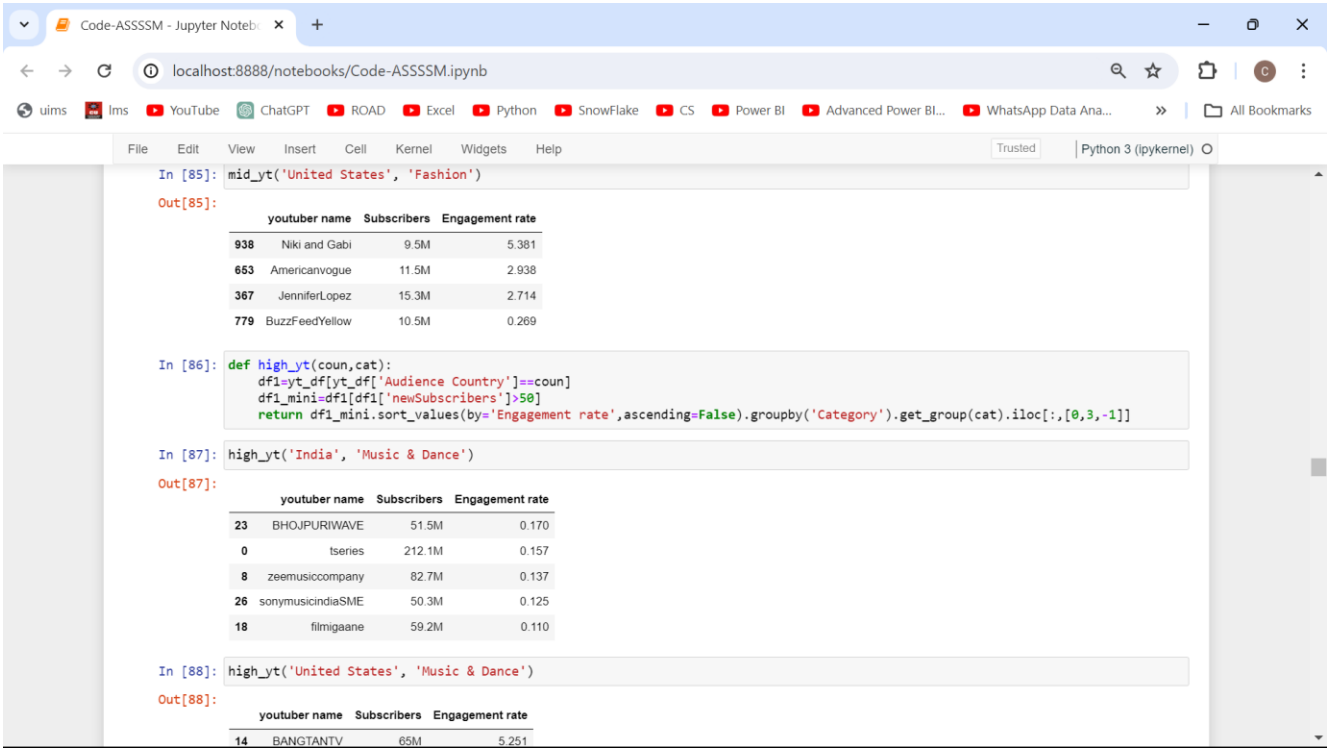

Code-ASSSSM - Jupyter Noteb
localhost:8888/notebooks/Code-ASSSSM.ipynb
uims
lms
YouTube
ChatGPT
ROAD
Excel
Python
SnowFlake
CS
Power BI
Advanced Power BI...
WhatsApp Data Ana...
All Bookmarks
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3 (ipykernel)
In [98]: idf
Out[98]:
Name Category Followers Audience Country New Followers Engagement Rate Platform
0 433 Sports with a ball 48.5M Spain 48.5 1.313 1
1 TAEYANG Music 12.7M Indonesia 12.7 4.270 1
2 НАСТЯ ИВЛЕЕВА Shows 18.8M Russia 18.8 2.010 1
3 Joy Lifestyle 13.5M Indonesia 13.5 10.370 1
4 Jaehyun NaN 11.1M Indonesia 11.1 27.928 1
... ...
995 Zendaya Cinema & Actors/actresses 136.1M United States 136.1 6.319 1
996 zidane Sports with a ball 31.2M Spain 31.2 2.385 1
997 KAI Music 13.9M Indonesia 13.9 11.511 1
998 Zoë Kravitz Cinema & Actors/actresses 8.2M United States 8.2 10.799 1
999 Zoë Sugg Lifestyle 9.4M United Kingdom 9.4 3.078 1
997 rows × 7 columns
In [99]: yt_df.head(3)
Out[99]:
youtuber name channel name Category Subscribers Audience Country newSubscribers Engagement rate
0 tseries T-Series Music & Dance 212.1M India 212.1 0.157
2 setindia SET India NaN 130.4M India 130.4 0.018
3 PewDiePie PewDiePie Animation 111.4M United States 111.4 1.333

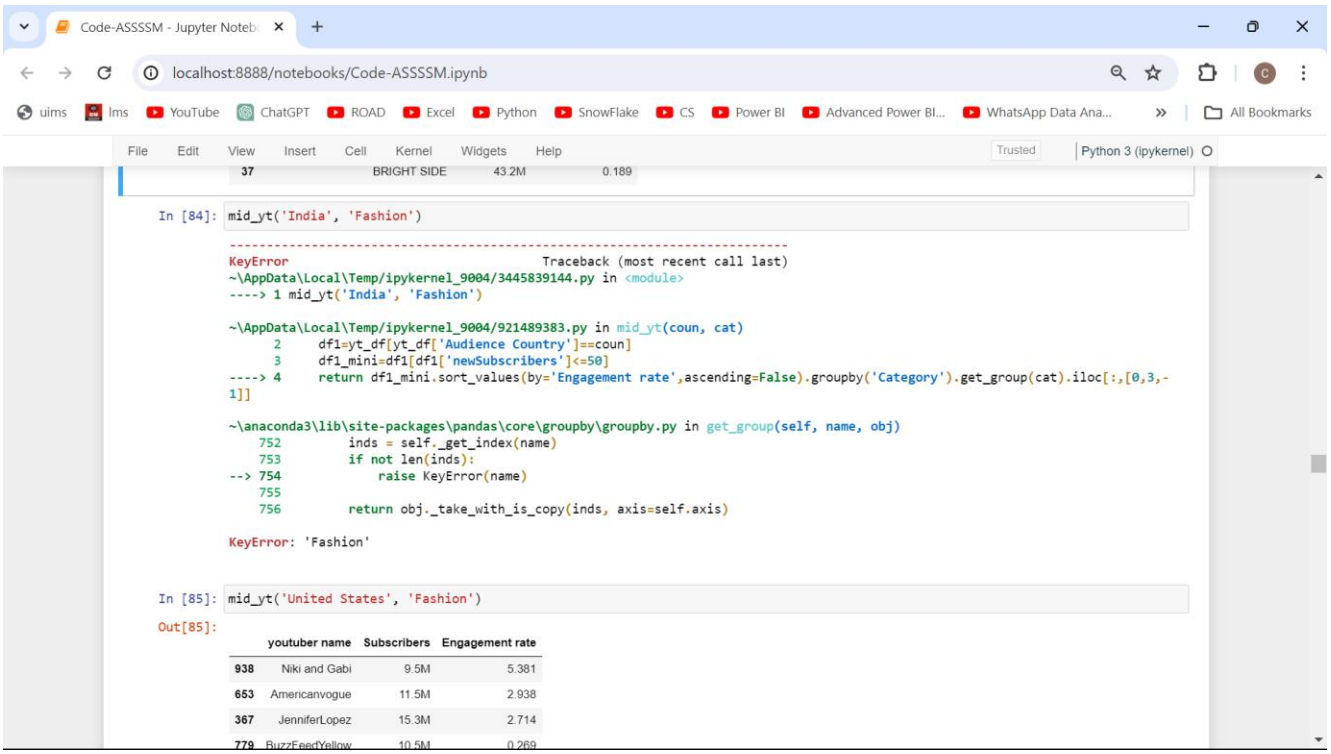

Code-ASSSSM - Jupyter Noteb
localhost:8888/notebooks/Code-ASSSSM.ipynb
uims
lms
YouTube
ChatGPT
ROAD
Excel
Python
SnowFlake
CS
Power BI
Advanced Power BI...
WhatsApp Data Ana...
All Bookmarks
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3 (ipykernel)
In [96]: idf
Out[96]:
instagram name Category Followers Audience Country newFollowers Engagement Rate Platform
0 433 Sports with a ball 48.5M Spain 48.5 1.313 1
1 TAEYANG Music 12.7M Indonesia 12.7 4.270 1
2 НАСТЯ ИВЛЕЕВА Shows 18.8M Russia 18.8 2.010 1
3 Joy Lifestyle 13.5M Indonesia 13.5 10.370 1
4 Jaehyun NaN 11.1M Indonesia 11.1 27.928 1
... ...
995 Zendaya Cinema & Actors/actresses 136.1M United States 136.1 6.319 1
996 zidane Sports with a ball 31.2M Spain 31.2 2.385 1
997 KAI Music 13.9M Indonesia 13.9 11.511 1
998 Zoë Kravitz Cinema & Actors/actresses 8.2M United States 8.2 10.799 1
999 Zoë Sugg Lifestyle 9.4M United Kingdom 9.4 3.078 1
997 rows × 7 columns
In [97]: idf.rename(columns = {'instagram name' : 'Name', 'newFollowers': 'New Followers'}, inplace = True)
C:\Users\gargc\anaconda3\lib\site-packages\pandas\core\frame.py:5039: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame
See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
return super().rename(
In [98]: idf

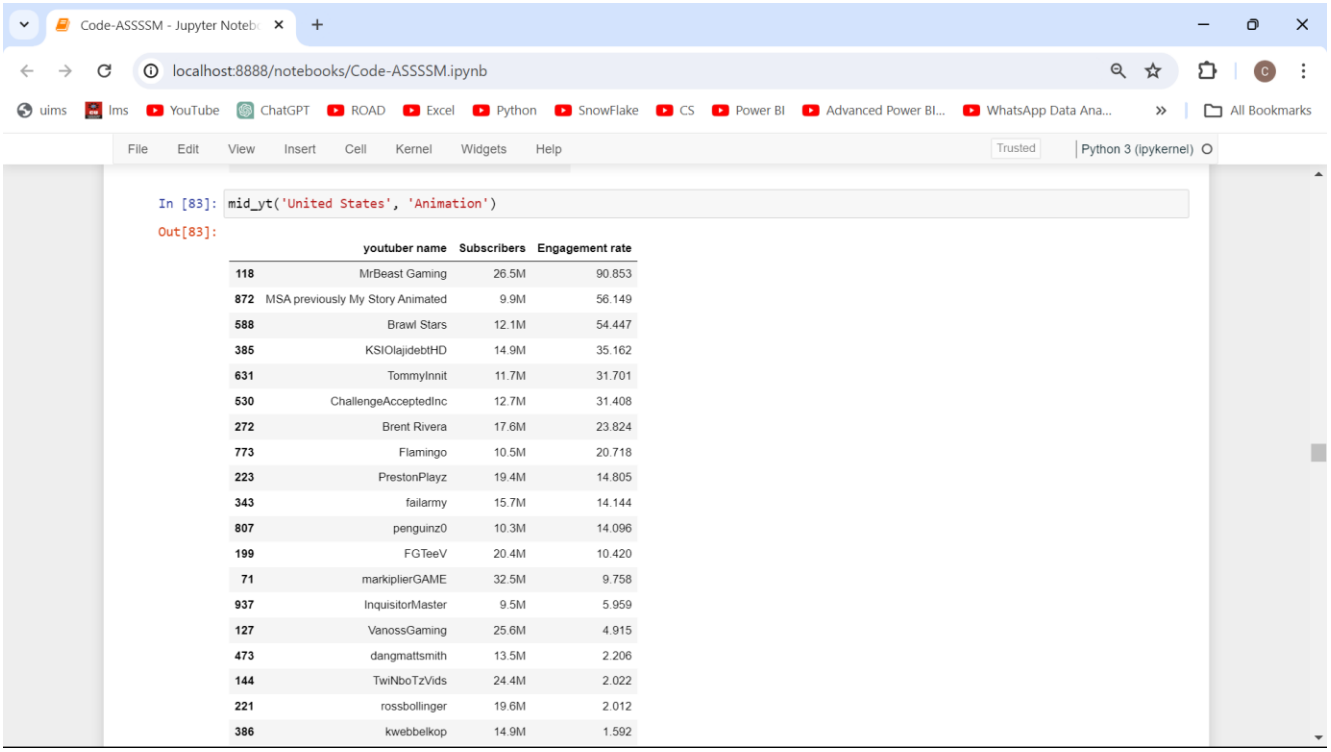

Code-ASSSSM - Jupyter Noteb
localhost:8888/notebooks/Code-ASSSSM.ipynb
uims
Ims
YouTube
ChatGPT
ROAD
Excel
Python
SnowFlake
CS
Power BI
Advanced Power BI...
WhatsApp Data Ana...
All Bookmarks
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3 (ipykernel)
Out[94]:
instagram name
Category
Followers
Audience Country
newFollowers
Engagement Rate
0 433 Sports with a ball 48.5M Spain 48.5 1.313
1 TAEYANG Music 12.7M Indonesia 12.7 4.270
2 НАСТЯ ИВЛЕЕВА Shows 18.8M Russia 18.8 2.010
3 Joy Lifestyle 13.5M Indonesia 13.5 10.370
4 Jaehyun NaN 11.1M Indonesia 11.1 27.928
995 Zendaya Cinema & Actors/actresses 136.1M United States 136.1 6.319
996 zidane Sports with a ball 31.2M Spain 31.2 2.385
997 KAI Music 13.9M Indonesia 13.9 11.511
998 Zoe Kravitz Cinema & Actors/actresses 8.2M United States 8.2 10.799
999 Zoe Sugg Lifestyle 9.4M United Kingdom 9.4 3.078
997 rows × 6 columns
In [95]:
idf['Platform'] = '1' # 1 for Instagram
C:\Users\gargc\AppData\Local\Temp/ipykernel_9004/4225959582.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
idf['Platform'] = '1' # 1 for Instagram
In [96]:
idf

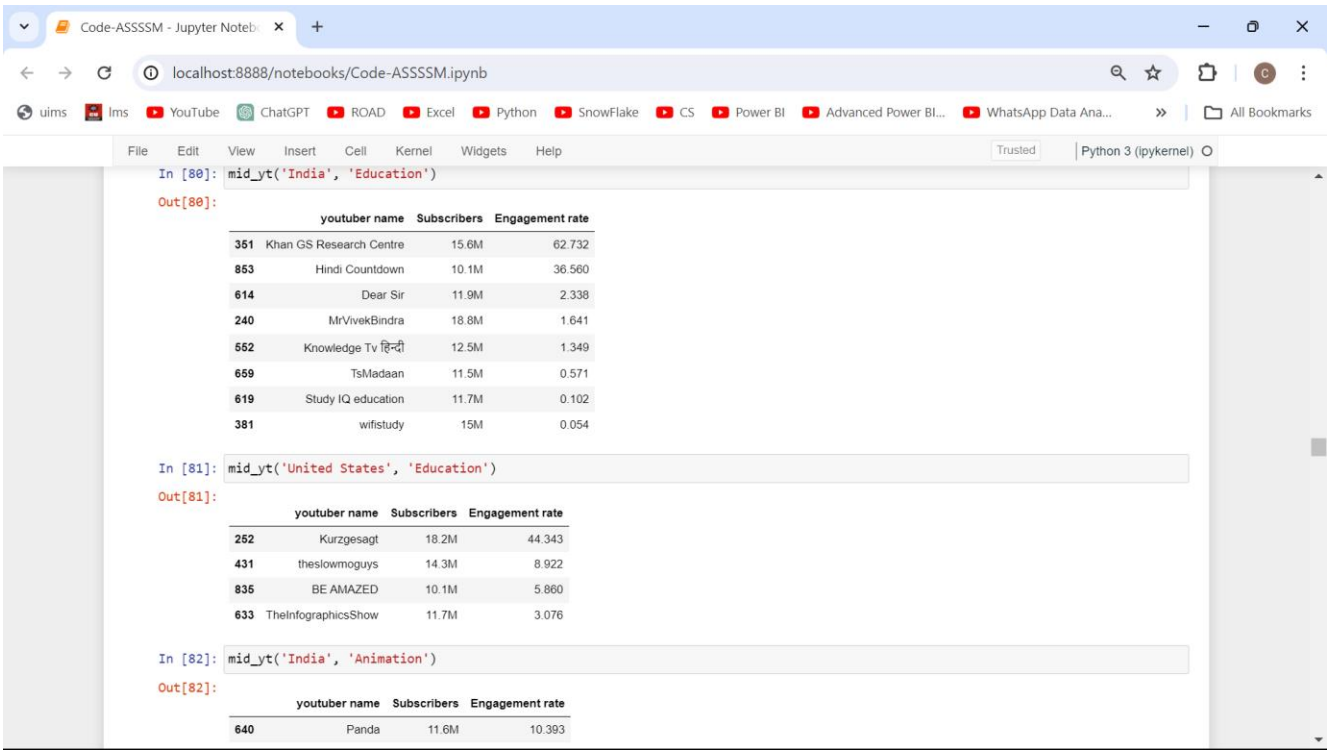

Code-ASSSSM - Jupyter Noteb
localhost:8888/notebooks/Code-ASSSSM.ipynb
uims
Ims
YouTube
ChatGPT
ROAD
Excel
Python
SnowFlake
CS
Power BI
Advanced Power BI...
WhatsApp Data Ana...
All Bookmarks
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3 (ipykernel)
youtuber name
Subscribers
Engagement rate
3 PewDiePie 111.4M 1.333
In [91]:
yt_df.head(5)
Out[91]:
youtuber name
channel name
Category
Subscribers
Audience Country
newSubscribers
Engagement rate
0 tseries T-Series Music & Dance 212.1M India 212.1 0.157
2 setindia SET India NaN 130.4M India 130.4 0.018
3 PewDiePie PewDiePie Animation 111.4M United States 111.4 1.333
4 MrBeast6000 MrBeast Video games 92.5M United States 92.5 34.992
7 WWEFanNation WWE Video games 86.9M United States 86.9 0.092
In [92]:
ig_df.head(5)
Out[92]:
instagram name
Category
category_2
Followers
Audience Country
Authentic engagement\r\n
newFollowers
Engagement Rate
0 433 Sports with a ball NaN 48.5M Spain 383.1K 48.5 1.313
1 TAEYANG Music NaN 12.7M Indonesia 478K 12.7 4.270
2 НАСТЯ ИВЛЕЕВА Shows NaN 18.8M Russia 310.8K 18.8 2.010
3 Joy Lifestyle NaN 13.5M Indonesia 1.1M 13.5 10.370
4 Jaehyun NaN NaN 11.1M Indonesia 2.5M 11.1 27.928
In [93]:
idf = ig_df[['instagram name', 'Category', 'Followers', 'Audience Country', 'newFollowers', 'Engagement Rate']]
In [94]:
idf

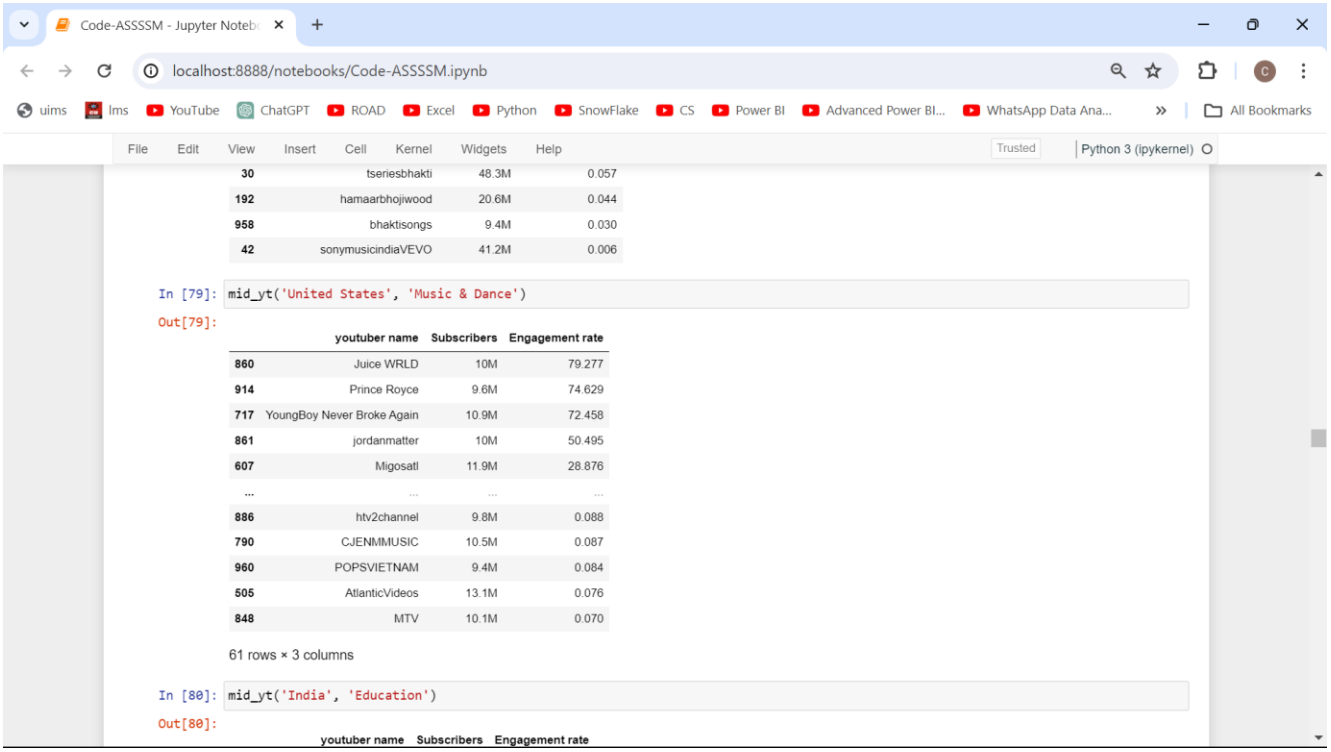

Code-ASSSSM - Jupyter Noteb
localhost:8888/notebooks/Code-ASSSSM.ipynb
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
14 BANGTANTV 65M 5.251
11 kidrauhl 68.1M 4.215
22 Marshmello 55M 1.251
16 ibighit 64.7M 0.721
In [89]: high_yt('India', 'Animation')
KeyError Traceback (most recent call last)
~\AppData\Local\Temp/ipykernel_9004/3885138581.py in <module>
----> 1 high_yt('India', 'Animation')
~\AppData\Local\Temp/ipykernel_9004/3759930940.py in high_yt(coun, cat)
2 df1=yt_df[yt_df['Audience Country']==coun]
3 df1_mini=df1[df1['newSubscribers']>50]
----> 4 return df1_mini.sort_values(by='Engagement rate',ascending=False).groupby('Category').get_group(cat).iloc[:,[0,3,-1]]
~\anaconda3\lib\site-packages\pandas\core\groupby\groupby.py in get_group(self, name, obj)
752 inds = self._get_index(name)
753 if not len(inds):
--> 754 raise KeyError(name)
755
756 return obj._take_with_is_copy(inds, axis=self.axis)
KeyError: 'Animation'
In [90]: high_yt('United States', 'Animation')
Out[90]:
youtuber name Subscribers Engagement rate

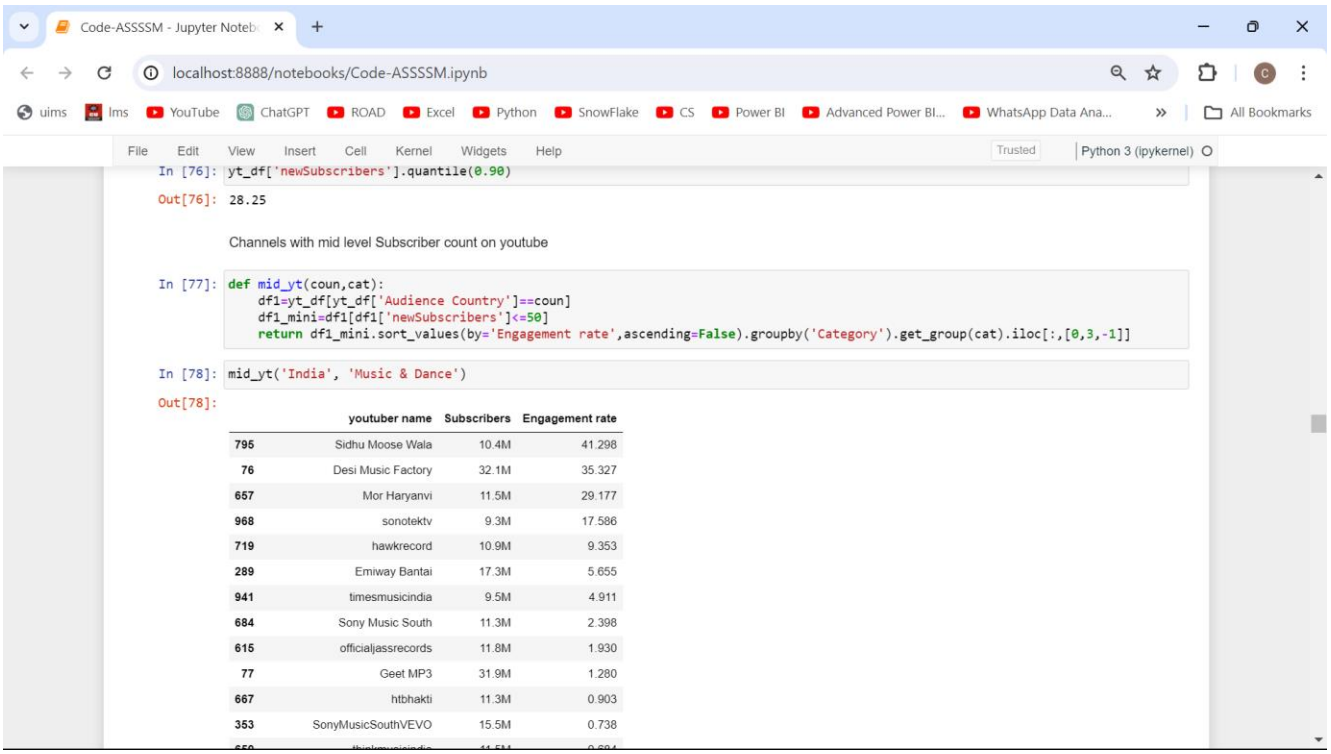

Code-ASSSSM - Jupyter Noteb
localhost:8888/notebooks/Code-ASSSSM.ipynb
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
In [85]: mid_yt('United States', 'Fashion')
Out[85]:
youtuber name Subscribers Engagement rate
938 Niki and Gabi 9.5M 5.381
653 Americanvogue 11.5M 2.938
367 JenniferLopez 15.3M 2.714
779 BuzzFeedYellow 10.5M 0.269
In [86]: def high_yt(coun,cat):
df1=yt_df[yt_df['Audience Country']==coun]
df1_mini=df1[df1['newSubscribers']>50]
return df1_mini.sort_values(by='Engagement rate',ascending=False).groupby('Category').get_group(cat).iloc[:,[0,3,-1]]
In [87]: high_yt('India', 'Music & Dance')
Out[87]:
youtuber name Subscribers Engagement rate
23 BHOJPURIWAVE 51.5M 0.170
0 tseries 212.1M 0.157
8 zeemusiccompany 82.7M 0.137
26 sonymusicindiaSME 50.3M 0.125
18 filmigaane 59.2M 0.110
In [88]: high_yt('United States', 'Music & Dance')
Out[88]:
youtuber name Subscribers Engagement rate
14 BANGTANTV 65M 5.251


Code-ASSSSM - Jupyter Noteb
localhost:8888/notebooks/Code-ASSSSM.ipynb
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
37 BRIGHT SIDE 43.2M 0.189
In [84]: mid_yt('India', 'Fashion')
KeyError Traceback (most recent call last)
~\AppData\Local\Temp/ipykernel_9004/3445839144.py in <module>
----> 1 mid_yt('India', 'Fashion')
~\AppData\Local\Temp/ipykernel_9004/921489383.py in mid_yt(coun, cat)
2 df1=yt_df[yt_df['Audience Country']==coun]
3 df1_mini=df1[df1['newSubscribers']<=50]
----> 4 return df1_mini.sort_values(by='Engagement rate',ascending=False).groupby('Category').get_group(cat).iloc[:,[0,3,-1]]
~\anaconda3\lib\site-packages\pandas\core\groupby\groupby.py in get_group(self, name, obj)
752 inds = self._get_index(name)
753 if not len(inds):
--> 754 raise KeyError(name)
755
756 return obj._take_with_is_copy(inds, axis=self.axis)
KeyError: 'Fashion'
In [85]: mid_yt('United States', 'Fashion')
Out[85]:
youtuber name Subscribers Engagement rate
938 Niki and Gabi 9.5M 5.381
653 Americanvogue 11.5M 2.938
367 JenniferLopez 15.3M 2.714
779 BuzzFeedYellow 10.5M 0.269

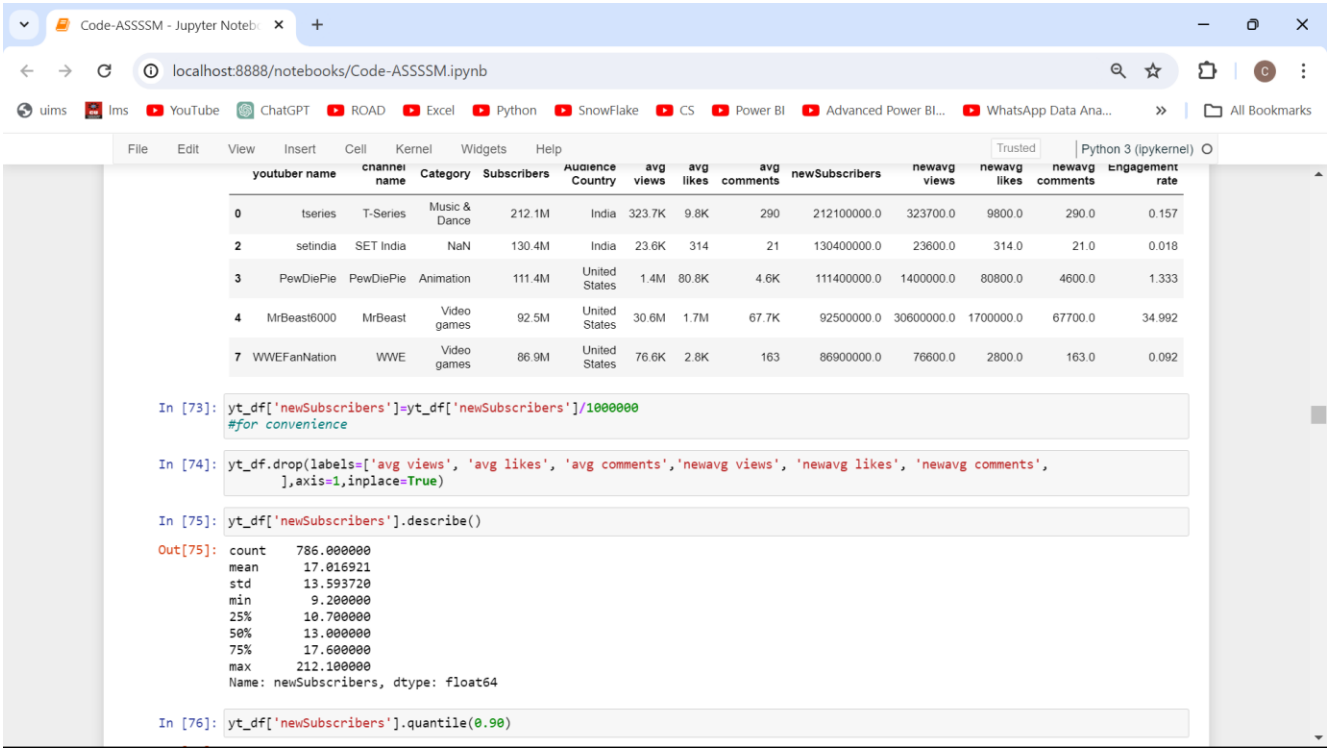

Code-ASSSSM - Jupyter Noteb
localhost:8888/notebooks/Code-ASSSSM.ipynb
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
In [83]: mid_yt('United States', 'Animation')
Out[83]:
youtuber name Subscribers Engagement rate
118 MrBeast Gaming 26.5M 90.853
872 MSA previously My Story Animated 9.9M 56.149
588 Brawl Stars 12.1M 54.447
385 KSIOlajidebtHD 14.9M 35.162
631 TommyInnit 11.7M 31.701
530 ChallengeAcceptedInc 12.7M 31.408
272 Brent Rivera 17.6M 23.824
773 Flamingo 10.5M 20.718
223 PrestonPlayz 19.4M 14.805
343 failarmy 15.7M 14.144
807 penguinz0 10.3M 14.096
199 FGTeeV 20.4M 10.420
71 markiplierGAME 32.5M 9.758
937 InquisitorMaster 9.5M 5.959
127 VanossGaming 25.6M 4.915
473 dangmattsmith 13.5M 2.206
144 TwiNboTzVids 24.4M 2.022
221 rossbollinger 19.6M 2.012
386 kwebbelkop 14.9M 1.592


Code-ASSSSM - Jupyter Noteb
localhost:8888/notebooks/Code-ASSSSM.ipynb
uims
lms
YouTube
ChatGPT
ROAD
Excel
Python
SnowFlake
CS
Power BI
Advanced Power BI...
WhatsApp Data Ana...
All Bookmarks
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3 (ipykernel)
In [80]: mid_yt('India', 'Education')
Out[80]:
youtuber name Subscribers Engagement rate
351 Khan GS Research Centre 15.6M 62.732
853 Hindi Countdown 10.1M 36.560
614 Dear Sir 11.9M 2.338
240 MrVivekBindra 18.8M 1.641
552 Knowledge Tv हिन्दी 12.5M 1.349
659 TsMadaan 11.5M 0.571
619 Study IQ education 11.7M 0.102
381 wifistudy 15M 0.054
In [81]: mid_yt('United States', 'Education')
Out[81]:
youtuber name Subscribers Engagement rate
252 Kurzgesagt 18.2M 44.343
431 theslowmoguys 14.3M 8.922
835 BE AMAZED 10.1M 5.860
633 TheInfographicsShow 11.7M 3.076
In [82]: mid_yt('India', 'Animation')
Out[82]:
youtuber name Subscribers Engagement rate
640 Panda 11.6M 10.393

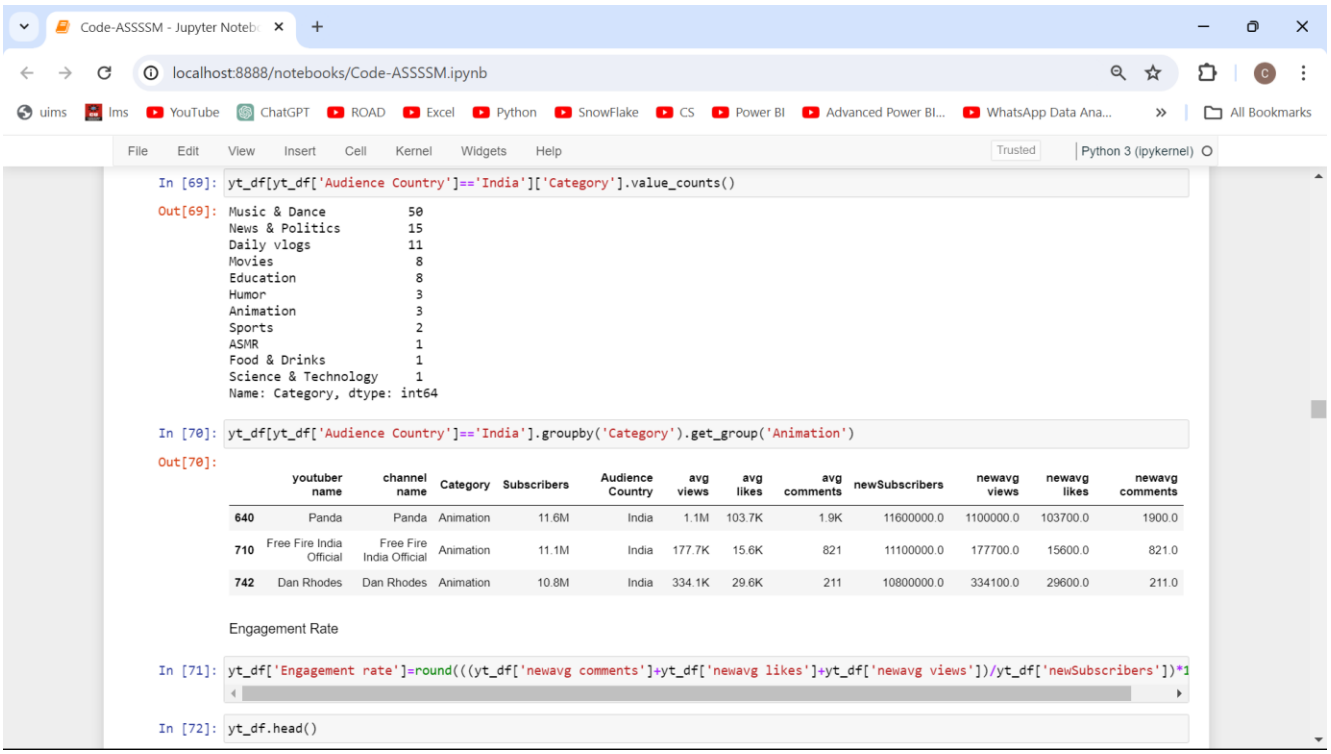

Code-ASSSSM - Jupyter Noteb
localhost:8888/notebooks/Code-ASSSSM.ipynb
uims
lms
YouTube
ChatGPT
ROAD
Excel
Python
SnowFlake
CS
Power BI
Advanced Power BI...
WhatsApp Data Ana...
All Bookmarks
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3 (ipykernel)
30 tseriesbhakti 48.3M 0.057
192 hamaarbhojiwood 20.6M 0.044
958 bhaktisongs 9.4M 0.030
42 sonymusicindiaVEVO 41.2M 0.006
In [79]: mid_yt('United States', 'Music & Dance')
Out[79]:
youtuber name Subscribers Engagement rate
860 Juice WRLD 10M 79.277
914 Prince Royce 9.6M 74.629
717 YoungBoy Never Broke Again 10.9M 72.458
861 jordanmatter 10M 50.495
607 Migosatl 11.9M 28.876
... ...
886 htv2channel 9.8M 0.088
790 CJENMMUSIC 10.5M 0.087
960 POPSVIETNAM 9.4M 0.084
505 AtlanticVideos 13.1M 0.076
848 MTV 10.1M 0.070
61 rows × 3 columns
In [80]: mid_yt('India', 'Education')
Out[80]:
youtuber name Subscribers Engagement rate



	youtuber name	Subscribers	Engagement rate
795	Sidhu Moose Wala	10.4M	41.298
76	Desi Music Factory	32.1M	35.327
657	Mor Haryanvi	11.5M	29.177
968	sonotektv	9.3M	17.586
719	hawkrecord	10.9M	9.353
289	Emiway Bantai	17.3M	5.655
941	timesmusicindia	9.5M	4.911
684	Sony Music South	11.3M	2.398
615	officialjassrecords	11.8M	1.930
77	Geet MP3	31.9M	1.280
667	htbhakti	11.3M	0.903
353	SonyMusicSouthVEVO	15.5M	0.738



	youtuber name	channel name	Category	Subscribers	Audience Country	avg views	avg likes	avg comments	newSubscribers	newavg views	newavg likes	newavg comments	Engagement rate
0	tseries	T-Series	Music & Dance	212.1M	India	323.7K	9.8K	290	212100000.0	323700.0	9800.0	290.0	0.157
2	setindia	SET India	NaN	130.4M	India	23.6K	314	21	130400000.0	23600.0	314.0	21.0	0.018
3	PewDiePie	PewDiePie	Animation	111.4M	United States	1.4M	80.8K	4.6K	111400000.0	1400000.0	80800.0	4600.0	1.333
4	MrBeast6000	MrBeast	Video games	92.5M	United States	30.6M	1.7M	67.7K	92500000.0	30600000.0	1700000.0	67700.0	34.992
7	WWEFanNation	WWE	Video games	86.9M	United States	76.6K	2.8K	163	86900000.0	76600.0	2800.0	163.0	0.092




```
In [69]: yt_df[yt_df['Audience Country']=='India']['Category'].value_counts()
```

```
Out[69]: Music & Dance          50
         News & Politics        15
         Daily vlogs            11
         Movies                  8
         Education               8
         Humor                   3
         Animation               3
         Sports                  2
         ASMR                    1
         Food & Drinks           1
         Science & Technology    1
         Name: Category, dtype: int64
```

```
In [70]: yt_df[yt_df['Audience Country']=='India'].groupby('Category').get_group('Animation')
```

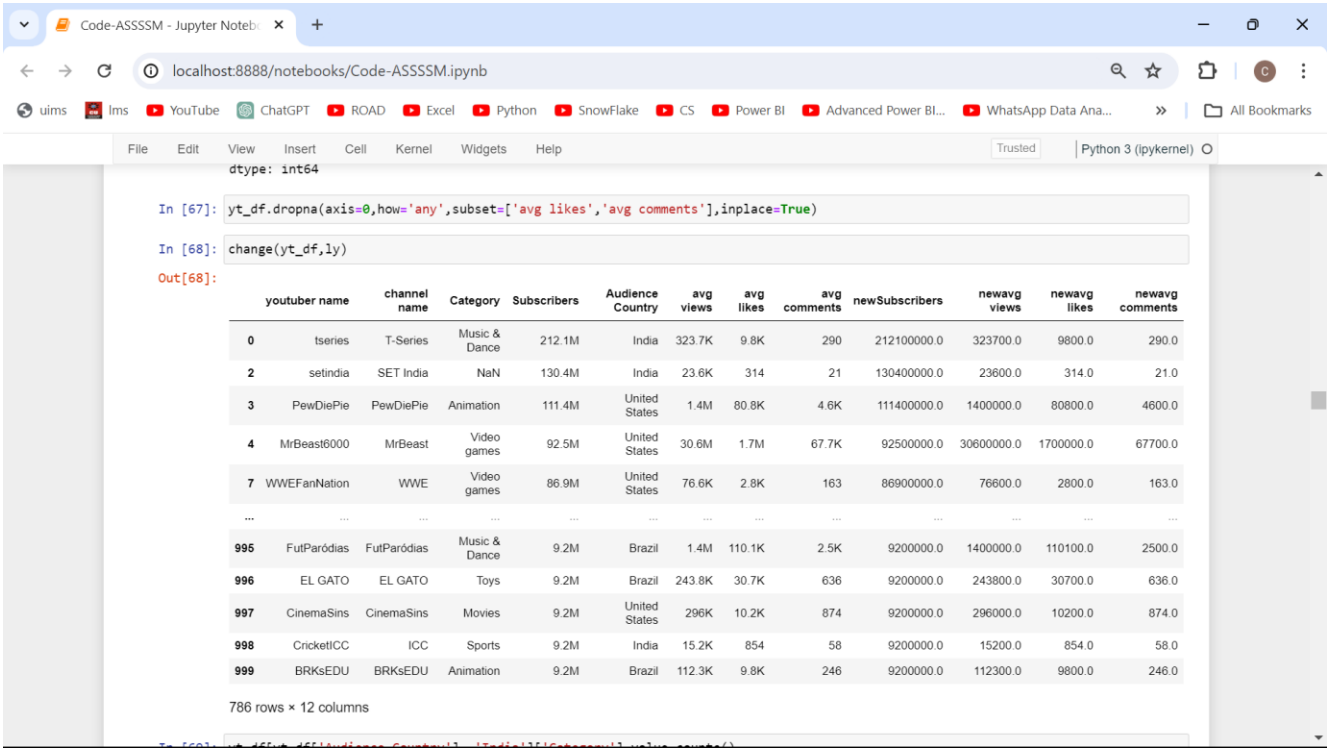
Out[70]:

	youtuber name	channel name	Category	Subscribers	Audience Country	avg views	avg likes	avg comments	newSubscribers	newavg views	newavg likes	newavg comments
640	Panda	Panda	Animation	11.6M	India	1.1M	103.7K	1.9K	11600000.0	1100000.0	103700.0	1900.0
710	Free Fire India Official	Free Fire India Official	Animation	11.1M	India	177.7K	15.6K	821	11100000.0	177700.0	15600.0	821.0
742	Dan Rhodes	Dan Rhodes	Animation	10.8M	India	334.1K	29.6K	211	10800000.0	334100.0	29600.0	211.0

Engagement Rate

```
In [71]: yt_df['Engagement rate']=round(((yt_df['newavg comments']+yt_df['newavg likes']+yt_df['newavg views'])/yt_df['newSubscribers'])*1
```

```
In [72]: yt_df.head()
```




```
dtype: int64
```

```
In [67]: yt_df.dropna(axis=0,how='any',subset=['avg likes','avg comments'],inplace=True)
```

```
In [68]: change(yt_df,1y)
```

Out[68]:

	youtuber name	channel name	Category	Subscribers	Audience Country	avg views	avg likes	avg comments	newSubscribers	newavg views	newavg likes	newavg comments
0	tseries	T-Series	Music & Dance	212.1M	India	323.7K	9.8K	290	212100000.0	323700.0	9800.0	290.0
2	setindia	SET India	NaN	130.4M	India	23.6K	314	21	130400000.0	23600.0	314.0	21.0
3	PewDiePie	PewDiePie	Animation	111.4M	United States	1.4M	80.8K	4.6K	111400000.0	1400000.0	80800.0	4600.0
4	MrBeast6000	MrBeast	Video games	92.5M	United States	30.6M	1.7M	67.7K	92500000.0	30600000.0	1700000.0	67700.0
7	WWEFanNation	WWE	Video games	86.9M	United States	76.6K	2.8K	163	86900000.0	76600.0	2800.0	163.0
...	...	...	...	...	...	...	...	...	...	...	...	...
995	FutParódias	FutParódias	Music & Dance	9.2M	Brazil	1.4M	110.1K	2.5K	9200000.0	1400000.0	110100.0	2500.0
996	EL GATO	EL GATO	Toys	9.2M	Brazil	243.8K	30.7K	636	9200000.0	243800.0	30700.0	636.0
997	CinemaSins	CinemaSins	Movies	9.2M	United States	296K	10.2K	874	9200000.0	296000.0	10200.0	874.0
998	CricketICC	ICC	Sports	9.2M	India	15.2K	854	58	9200000.0	15200.0	854.0	58.0
999	BRKsEDU	BRKsEDU	Animation	9.2M	Brazil	112.3K	9.8K	246	9200000.0	112300.0	9800.0	246.0

786 rows × 12 columns

Code-ASSSSM - Jupyter Noteb

localhost:8888/notebooks/Code-ASSSSM.ipynb

uims lms YouTube ChatGPT ROAD Excel Python SnowFlake CS Power BI Advanced Power BI... WhatsApp Data Ana... All Bookmarks

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

TOP 15 most followed channels on YouTube

```
In [64]: yt_df.iloc[0:10, [1,2,3]]
```

Out[64]:

	channel name	Category	Subscribers
0	T-Series	Music & Dance	212.1M
1	Cocomelon - Nursery Rhymes	Education	132.1M
2	SET India	NaN	130.4M
3	PewDiePie	Animation	111.4M
4	MrBeast	Video games	92.5M
5	✿ Kids Diana Show	Animation	92.4M
6	Like Nastya	Animation	90.1M
7	WWE	Video games	86.9M
8	Zee Music Company	Music & Dance	82.7M
9	Vlad and Niki	Toys	80.4M

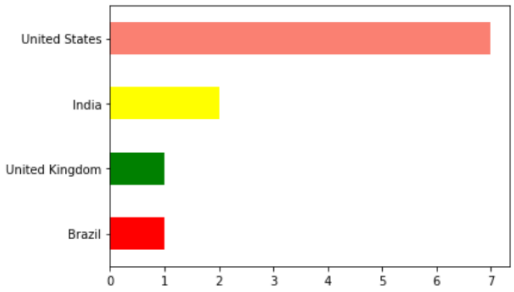
```
In [65]: ly=['Subscribers','avg views', 'avg likes', 'avg comments']
```

```
In [66]: yt_df.dropna(axis=0,how='any',subset=['avg likes','avg comments']).isnull().sum()
```

```
Out[66]: youtuber name          0
         channel name           0
         Category             215
         Subscribers            0
         Audience Country       0
         avg views              0
         avg likes              0
         avg comments           0
```

```
In [63]: demand(yt_df, 'Sports')
```

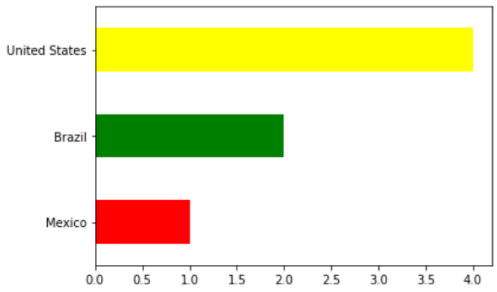
```
Out[63]: <AxesSubplot:>
```



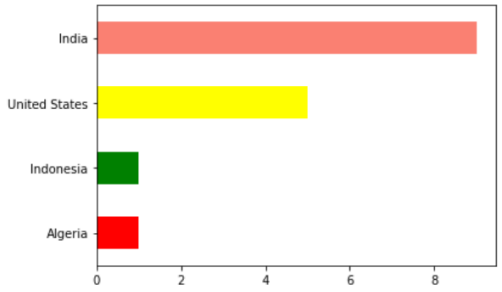
TOP 15 most followed channels on YouTube

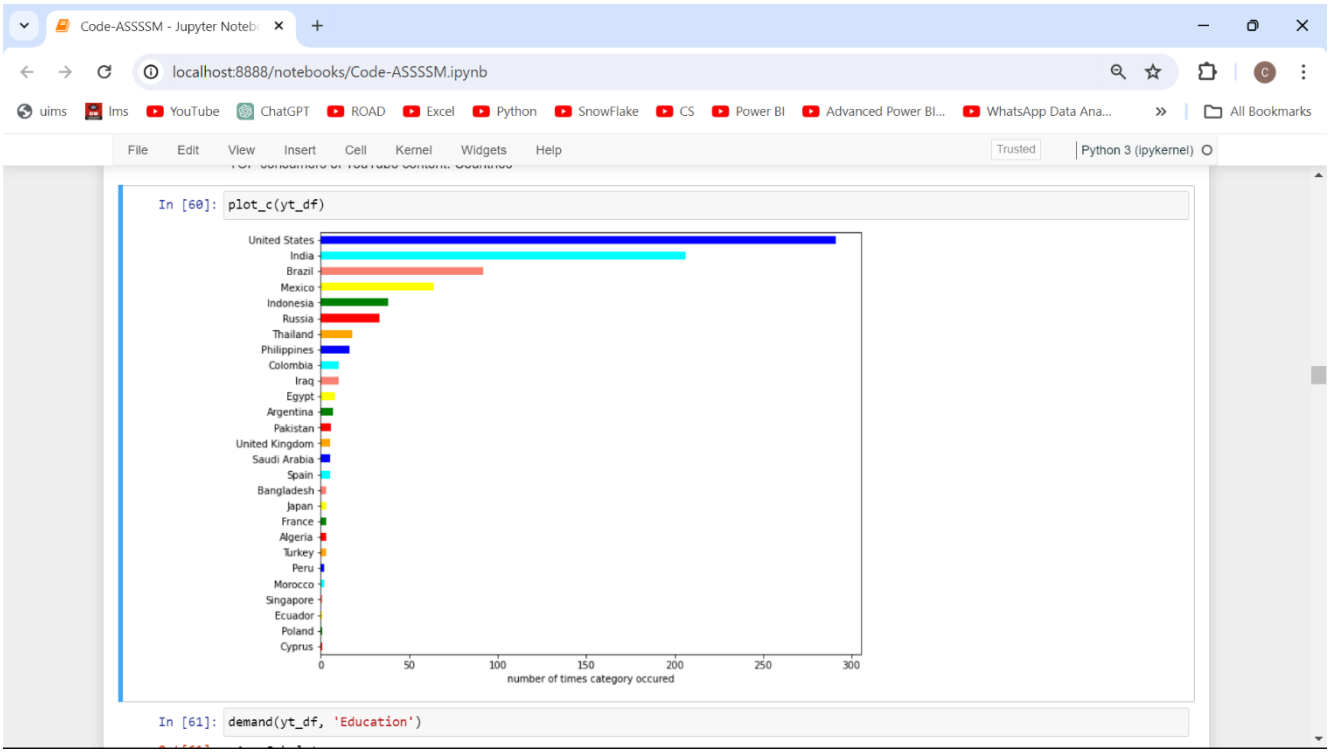
```
In [62]: demand(yt_df, 'Fashion')
```

```
Out[62]: <AxesSubplot:>
```



```
In [61]: demand(yt_df, 'Education')
Out[61]: <AxesSubplot:>
```





Code-ASSSSM - Jupyter Noteb

localhost:8888/notebooks/Code-ASSSSM.ipynb

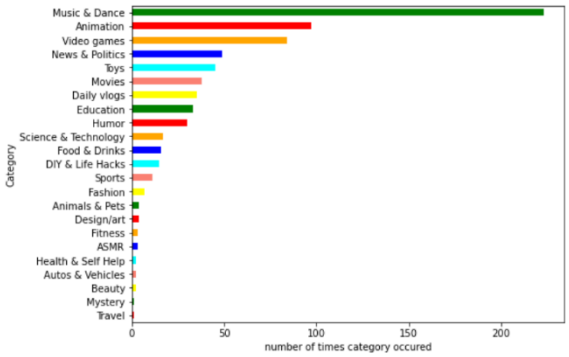
uims lms YouTube ChatGPT ROAD Excel Python SnowFlake CS Power BI Advanced Power BI... WhatsApp Data Ana... All Bookmarks

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Engagement Rate

In [58]:
```
yt_df.drop_duplicates(subset=['channel name'],inplace=True)
```

In [59]:
```
plot(yt_df)
```



TOP consumers of YouTube content: Countries

In [60]:
```
plot c(yt_df)
```

localhost:8888/notebooks/Code-ASSSSM.ipynb#

Code-ASSSSM - Jupyter Noteb

localhost:8888/notebooks/Code-ASSSSM.ipynb

uims lms YouTube ChatGPT ROAD Excel Python SnowFlake CS Power BI Advanced Power BI... WhatsApp Data Ana... All Bookmarks

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

1	checkgate	Cocomelon - Nursery Rhymes	Education	132.1M	NaN	13.8M	80.9K	NaN
2	setindia	SET India	NaN	130.4M	India	23.6K	314	21
3	PewDiePie	PewDiePie	Animation	111.4M	United States	1.4M	80.8K	4.6K
4	MrBeast6000	MrBeast	Video games	92.5M	United States	30.6M	1.7M	67.7K

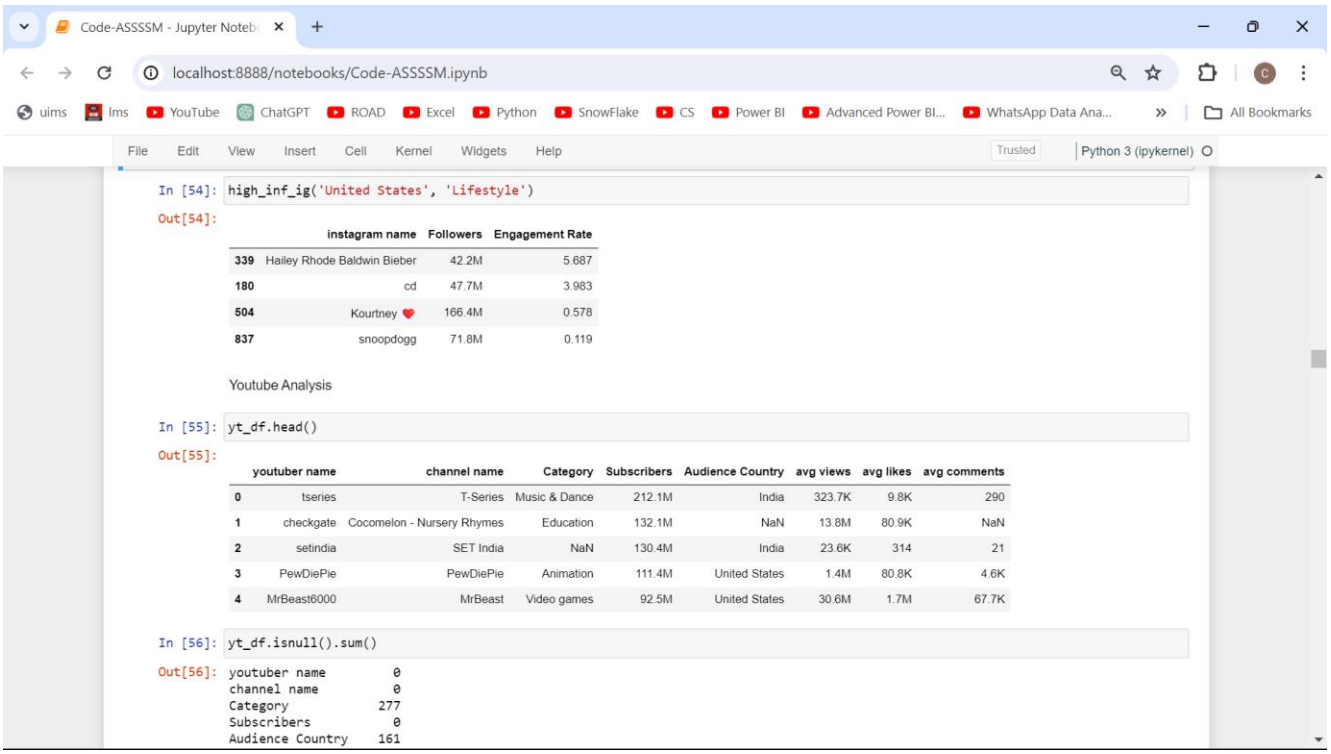
In [56]:
```
yt_df.isnull().sum()
```

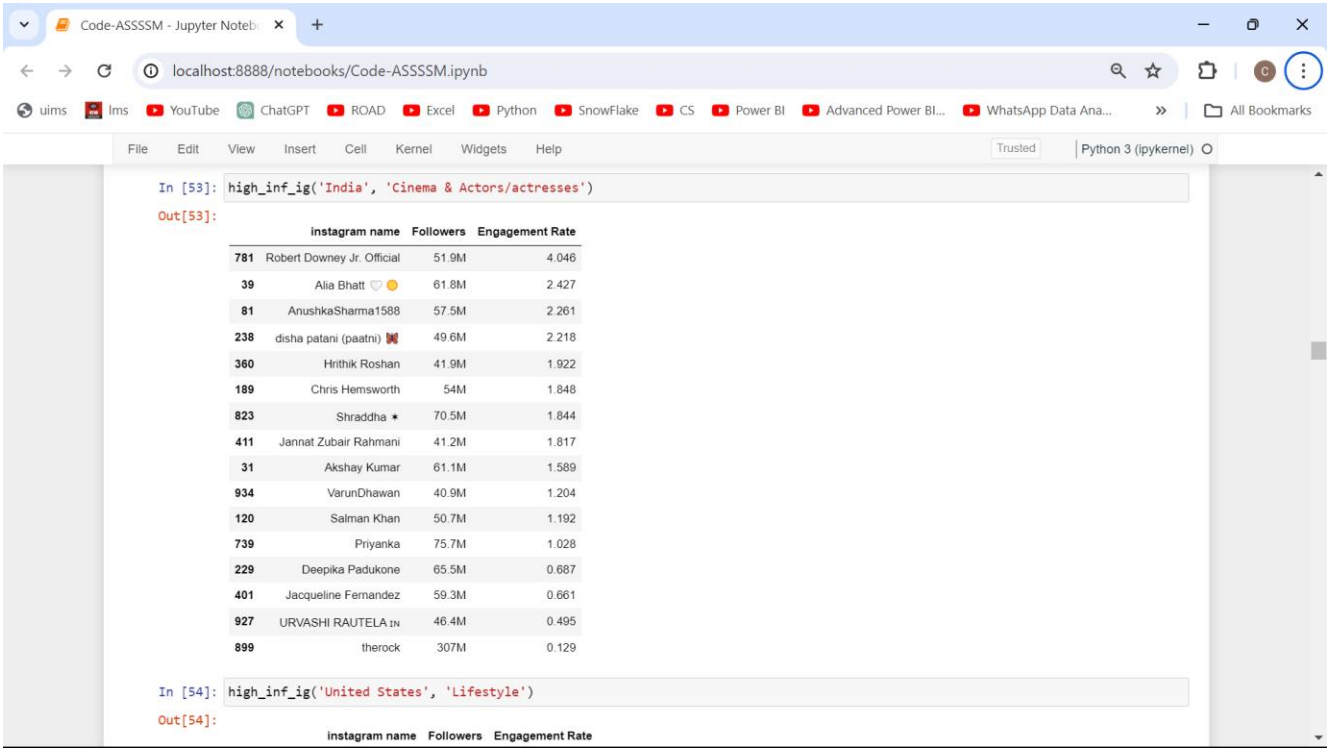
Out[56]:
```
youtuber name         0
channel name          0
Category            277
Subscribers           0
Audience Country    161
avg views             0
avg likes            38
avg comments        209
dtype: int64
```

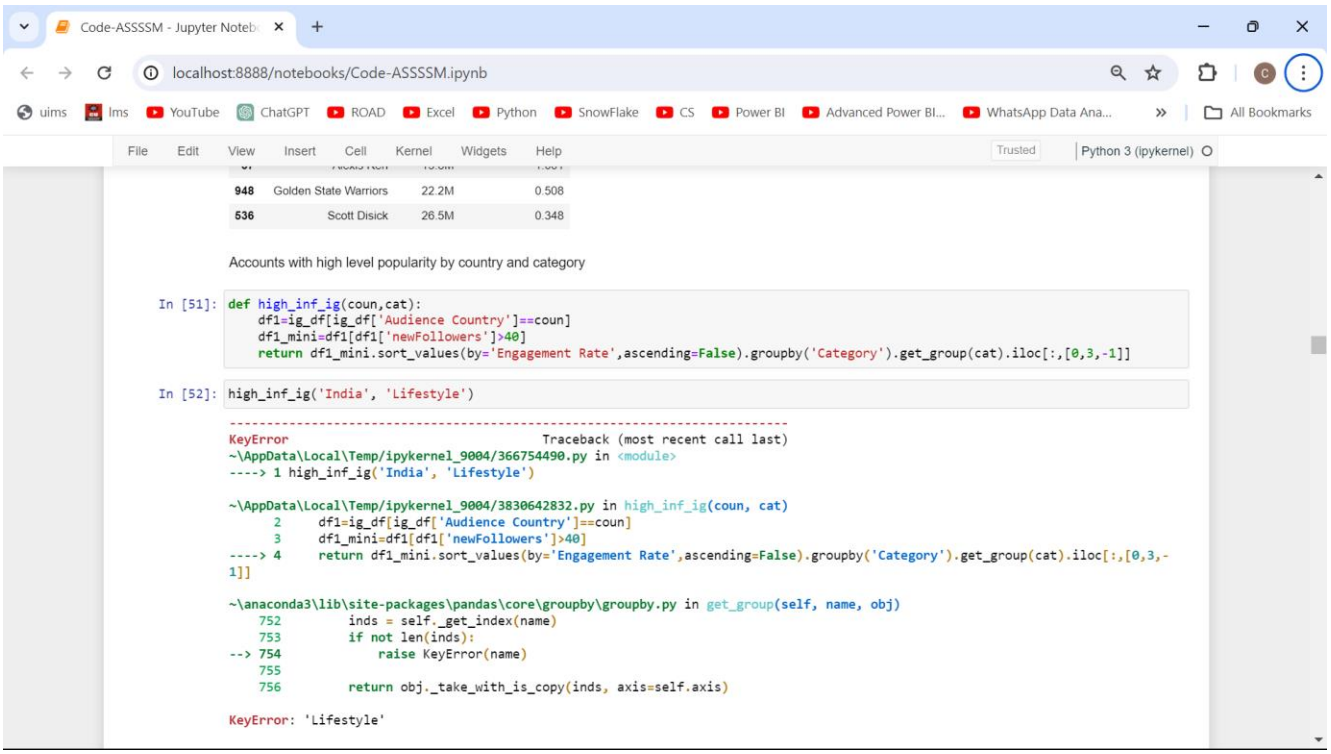
In [57]:
```
yt_df.isna().sum()
```

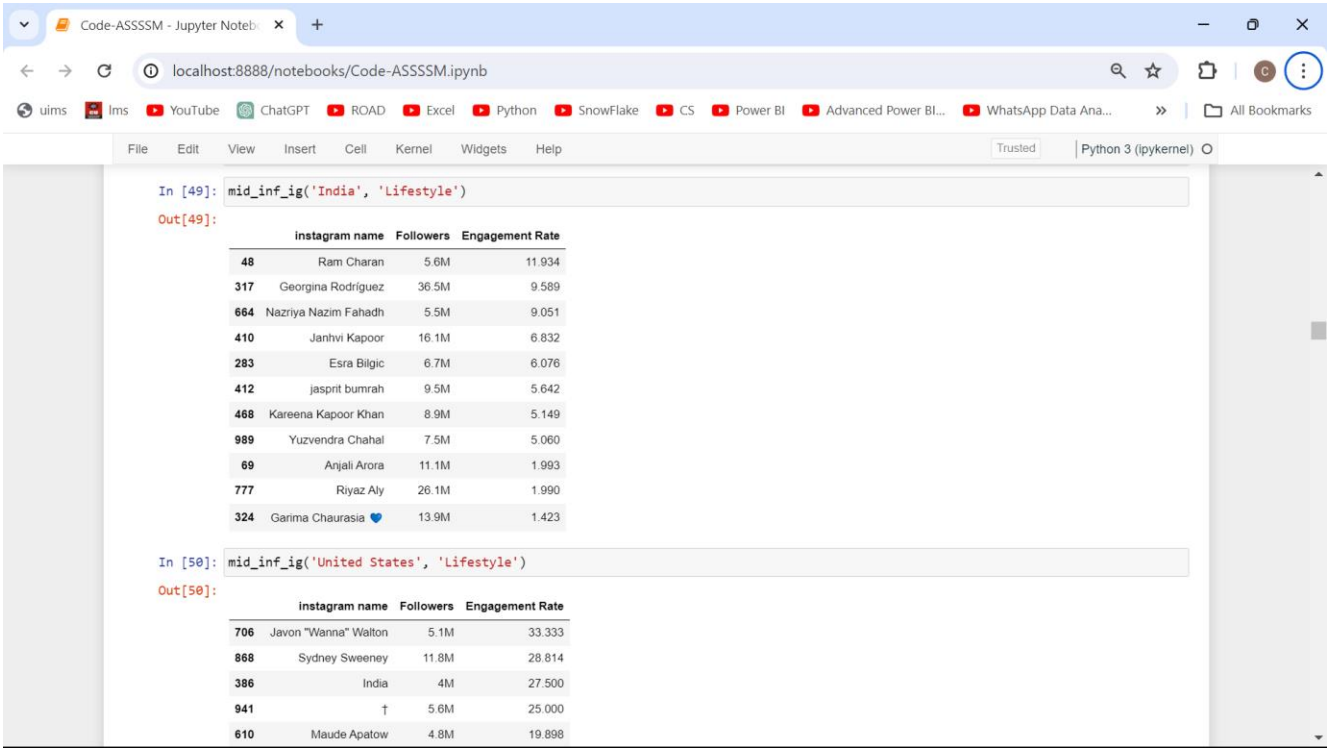
Out[57]:
```
youtuber name         0
channel name          0
Category            277
Subscribers           0
Audience Country    161
avg views             0
avg likes            38
avg comments        209
dtype: int64
```

Engagement Rate


Code-ASSSSM - Jupyter Noteb
localhost:8888/notebooks/Code-ASSSSM.ipynb
uims
lms
YouTube
ChatGPT
ROAD
Excel
Python
SnowFlake
CS
Power BI
Advanced Power BI...
WhatsApp Data Ana...
All Bookmarks
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3 (ipykernel)
In [54]: high_inf_ig('United States', 'Lifestyle')
Out[54]:
instagram name Followers Engagement Rate
339 Hailey Rhode Baldwin Bieber 42.2M 5.687
180 cd 47.7M 3.983
504 Kourtney ❤ 166.4M 0.578
837 snoopdogg 71.8M 0.119
Youtube Analysis
In [55]: yt_df.head()
Out[55]:
youtuber name channel name Category Subscribers Audience Country avg views avg likes avg comments
0 tseries T-Series Music & Dance 212.1M India 323.7K 9.8K 290
1 checkgate Cocomelon - Nursery Rhymes Education 132.1M NaN 13.8M 80.9K NaN
2 setindia SET India NaN 130.4M India 23.6K 314 21
3 PewDiePie PewDiePie Animation 111.4M United States 1.4M 80.8K 4.6K
4 MrBeast6000 MrBeast Video games 92.5M United States 30.6M 1.7M 67.7K
In [56]: yt_df.isnull().sum()
Out[56]:
youtuber name 0
channel name 0
Category 277
Subscribers 0
Audience Country 161


Code-ASSSSM - Jupyter Noteb
localhost:8888/notebooks/Code-ASSSSM.ipynb
uims
lms
YouTube
ChatGPT
ROAD
Excel
Python
SnowFlake
CS
Power BI
Advanced Power BI...
WhatsApp Data Ana...
All Bookmarks
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3 (ipykernel)
In [53]: high_inf_ig('India', 'Cinema & Actors/actresses')
Out[53]:
instagram name Followers Engagement Rate
781 Robert Downey Jr. Official 51.9M 4.046
39 Alia Bhatt 61.8M 2.427
81 AnushkaSharma1588 57.5M 2.261
238 disha patani (paatni) 49.6M 2.218
360 Hrithik Roshan 41.9M 1.922
189 Chris Hemsworth 54M 1.848
823 Shraddha ✦ 70.5M 1.844
411 Jannat Zubair Rahmani 41.2M 1.817
31 Akshay Kumar 61.1M 1.589
934 VarunDhawan 40.9M 1.204
120 Salman Khan 50.7M 1.192
739 Priyanka 75.7M 1.028
229 Deepika Padukone 65.5M 0.687
401 Jacqueline Fernandez 59.3M 0.661
927 URVASHI RAUTELA IN 46.4M 0.495
899 therock 307M 0.129
In [54]: high_inf_ig('United States', 'Lifestyle')
Out[54]:
instagram name Followers Engagement Rate


Code-ASSSSM - Jupyter Noteb
localhost:8888/notebooks/Code-ASSSSM.ipynb
uims
lms
YouTube
ChatGPT
ROAD
Excel
Python
SnowFlake
CS
Power BI
Advanced Power BI...
WhatsApp Data Ana...
All Bookmarks
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3 (ipykernel)
948 Golden State Warriors 22.2M 0.508
536 Scott Disick 26.5M 0.348
Accounts with high level popularity by country and category
In [51]:
def high_inf_ig(coun,cat):
df1=ig_df[ig_df['Audience Country']==coun]
df1_mini=df1[df1['newFollowers']>40]
return df1_mini.sort_values(by='Engagement Rate',ascending=False).groupby('Category').get_group(cat).iloc[:,[0,3,-1]]
In [52]:
high_inf_ig('India', 'Lifestyle')
KeyError Traceback (most recent call last)
~\AppData\Local\Temp/ipykernel_9004/366754490.py in <module>
----> 1 high_inf_ig('India', 'Lifestyle')
~\AppData\Local\Temp/ipykernel_9004/3830642832.py in high_inf_ig(coun, cat)
2 df1=ig_df[ig_df['Audience Country']==coun]
3 df1_mini=df1[df1['newFollowers']>40]
----> 4 return df1_mini.sort_values(by='Engagement Rate',ascending=False).groupby('Category').get_group(cat).iloc[:,[0,3,-1]]
~\anaconda3\lib\site-packages\pandas\core\groupby\groupby.py in get_group(self, name, obj)
752 inds = self._get_index(name)
753 if not len(inds):
--> 754 raise KeyError(name)
755
756 return obj._take_with_is_copy(inds, axis=self.axis)
KeyError: 'Lifestyle'


Code-ASSSSM - Jupyter Noteb
localhost:8888/notebooks/Code-ASSSSM.ipynb
uims
lms
YouTube
ChatGPT
ROAD
Excel
Python
SnowFlake
CS
Power BI
Advanced Power BI...
WhatsApp Data Ana...
All Bookmarks
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3 (ipykernel)
In [49]:
mid_inf_ig('India', 'Lifestyle')
Out[49]:
instagram name Followers Engagement Rate
48 Ram Charan 5.6M 11.934
317 Georgina Rodríguez 36.5M 9.589
664 Nazriya Nazim Fahadh 5.5M 9.051
410 Janhvi Kapoor 16.1M 6.832
283 Esra Bilgic 6.7M 6.076
412 jasprit bumrah 9.5M 5.642
468 Kareena Kapoor Khan 8.9M 5.149
989 Yuzvendra Chahal 7.5M 5.060
69 Anjali Arora 11.1M 1.993
777 Riyaz Aly 26.1M 1.990
324 Garima Chaurasia 💙 13.9M 1.423
In [50]:
mid_inf_ig('United States', 'Lifestyle')
Out[50]:
instagram name Followers Engagement Rate
706 Javon "Wanna" Walton 5.1M 33.333
868 Sydney Sweeney 11.8M 28.814
386 India 4M 27.500
941 † 5.6M 25.000
610 Maude Apatow 4.8M 19.898

Code-ASSSSM - Jupyter Noteb

localhost:8888/notebooks/Code-ASSSSM.ipynb

uims lms YouTube ChatGPT ROAD Excel Python SnowFlake CS Power BI Advanced Power BI... WhatsApp Data Ana... All Bookmarks

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

In [47]:
```
ig_df[ig_df['Audience Country']=='India']['Category'].value_counts()
```

Out[47]:
```
Cinema & Actors/actresses    69
Sports with a ball           16
Music                        13
Lifestyle                    11
Beauty                        5
Shows                         4
Modeling                      4
Fashion                       3
Humor & Fun & Happiness       2
Computers & Gadgets           2
Art/Artists                   2
Luxury                        1
Science                       1
Finance & Economics           1
Fitness & Gym                 1
Adult content                 1
Photography                   1
Animals                       1
Family                        1
Cars & Motorbikes             1
Business & Careers            1
Name: Category, dtype: int64
```

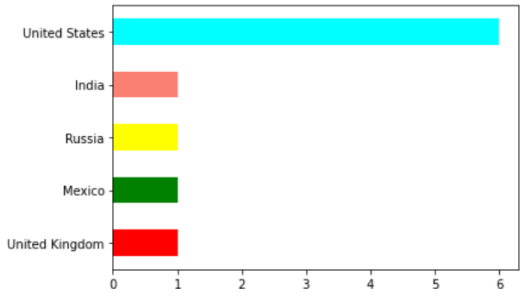
Accounts with medium level popularity by country and category

In [48]:
```
def mid_inf_ig(coun,cat):
    df1=ig_df[ig_df['Audience Country']==coun]
    df1_mini=df1[df1['newFollowers']<40]
    return df1_mini.sort_values(by='Engagement Rate',ascending=False).groupby('Category').get_group(cat).iloc[:,[0,3,-1]]
```

In [49]:
```
mid_inf_ig('India', 'Lifestyle')
```

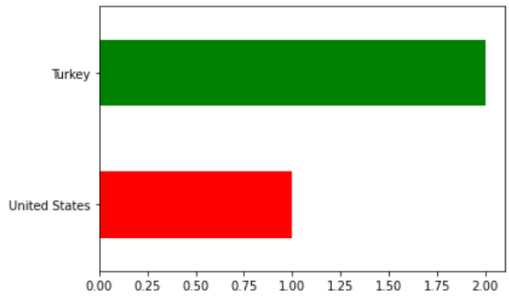
In [46]:
```
demand(ig_df,'Fitness & Gym')
```

Out[46]: <AxesSubplot:>



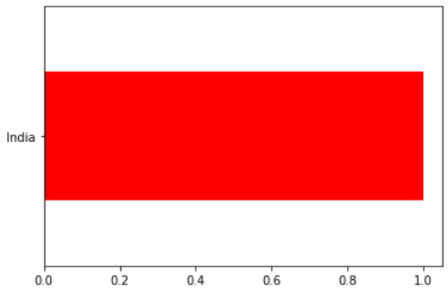
In [45]:
```
demand(ig_df,'Food & Cooking')
```

Out[45]: <AxesSubplot:>



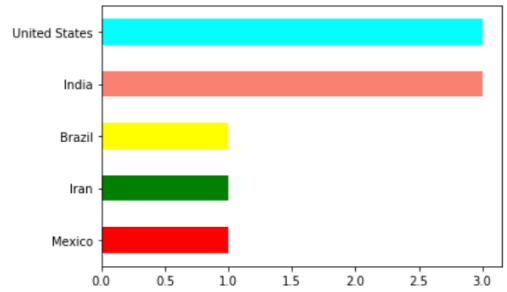
In [44]:
```
demand(ig_df,'Luxury')
```

Out[44]: <AxesSubplot:>



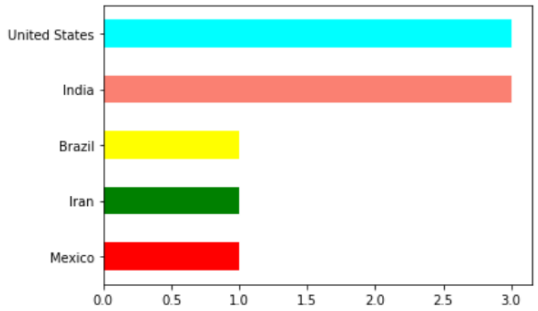
In [43]:
```
demand(ig_df,'Fashion')
```

Out[43]: <AxesSubplot:>



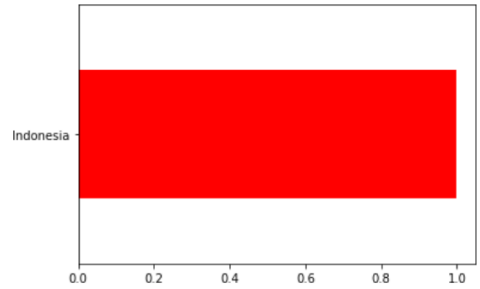
In [43]:
```
demand(ig_df,'Fashion')
```

Out[43]: <AxesSubplot:>



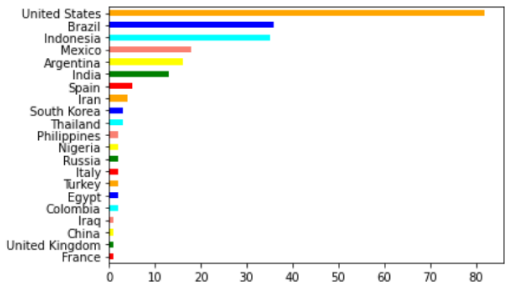
In [42]:
```
demand(ig_df,'Education')
```

Out[42]: <AxesSubplot:>



In [41]:
```
demand(ig_df,'Music')
```

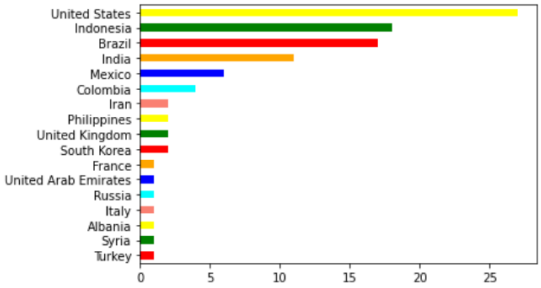
Out[41]: <AxesSubplot:>



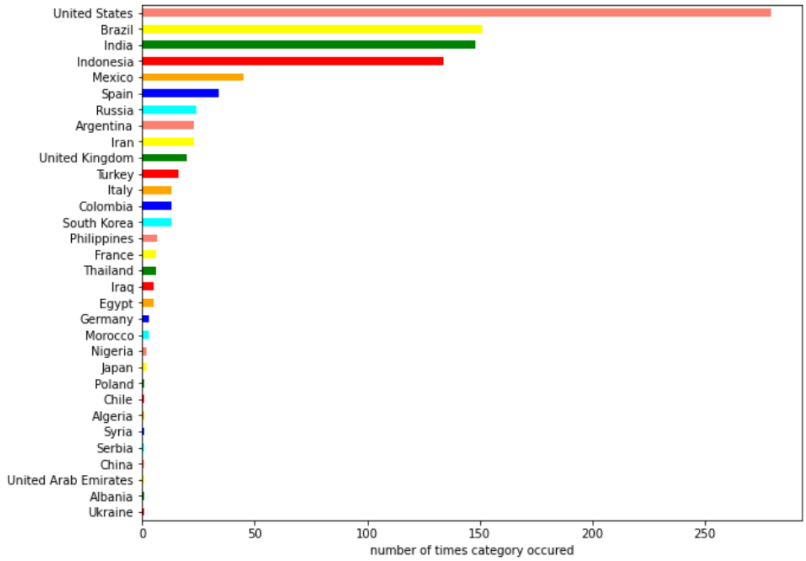
In [39]:
```
def demand(data,category):
    return data[data['Category']==category]['Audience Country'].value_counts().sort_values(ascending=True).plot.barh(color=pallet
```

In [40]:
```
demand(ig_df,'Lifestyle')
```

Out[40]: <AxesSubplot:>

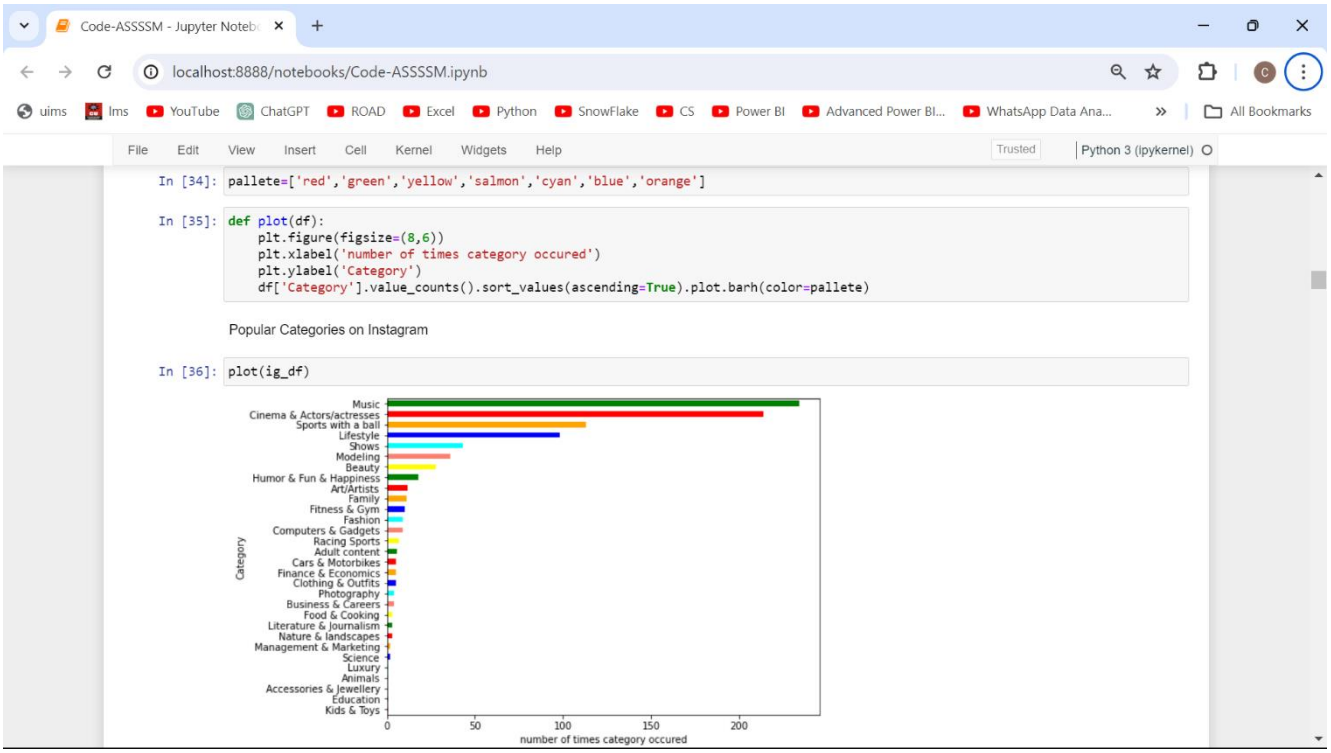


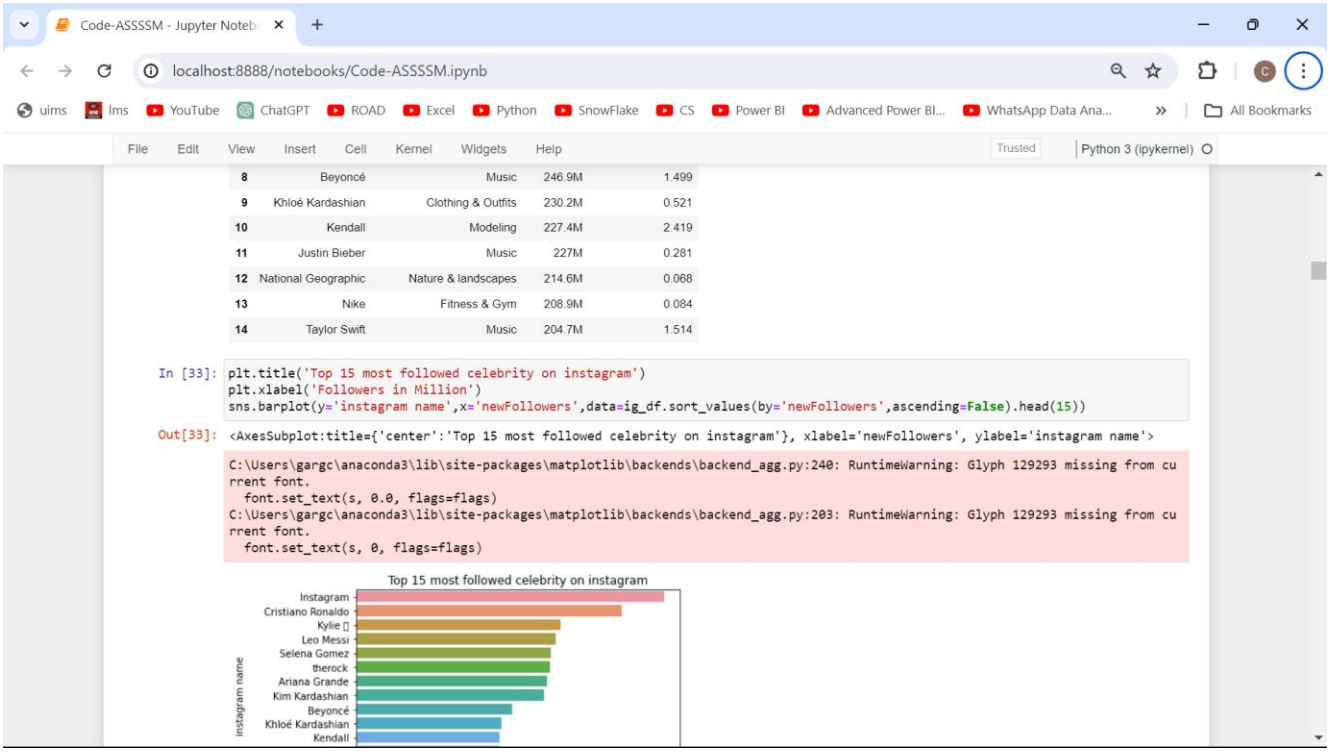
In [38]:
```
plot_c(ig_df)
```

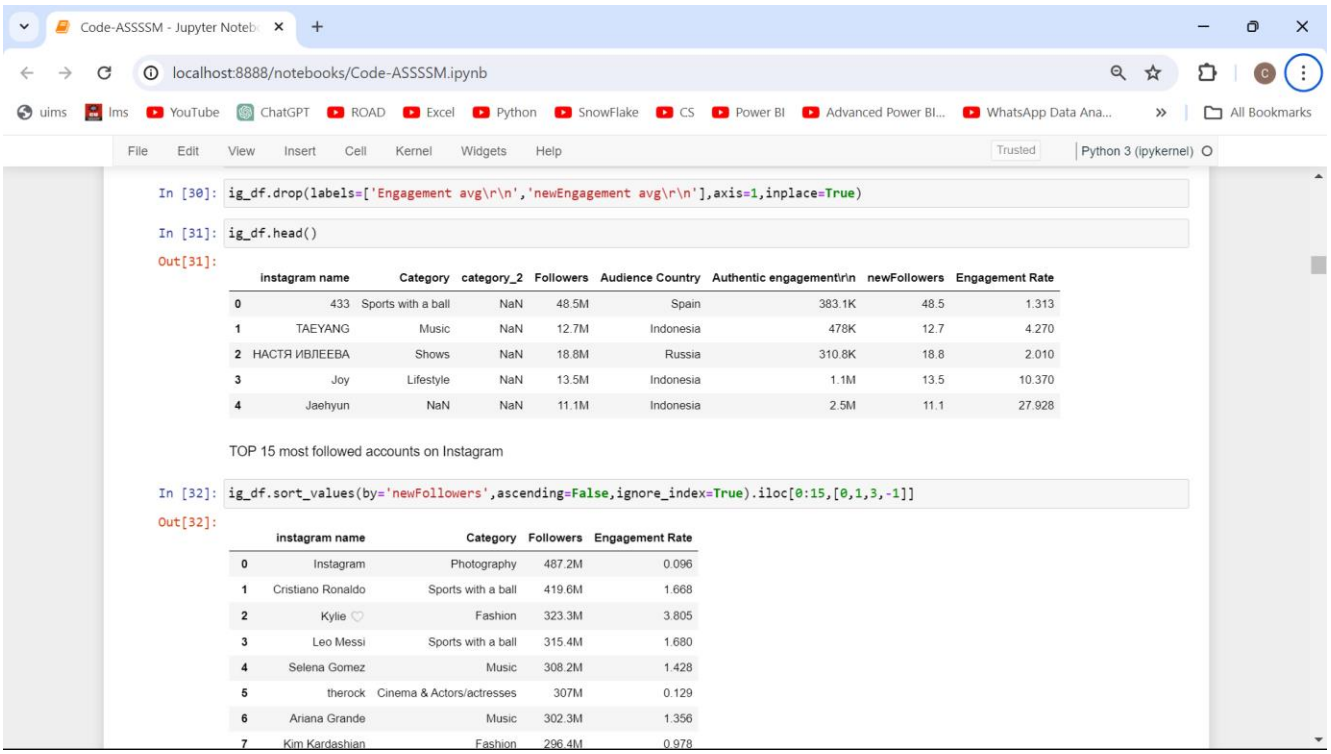


```
In [37]: def plot_c(df):
             plt.figure(figsize=(10,8))
             plt.xlabel('number of times category occured')
             df['Audience Country'].value_counts().sort_values().plot.barh(color=pallete)
```

Top consumers of Instagram content: Countries








```
In [30]: ig_df.drop(labels=['Engagement avg\r\n','newEngagement avg\r\n'],axis=1,inplace=True)
```

```
In [31]: ig_df.head()
```

Out[31]:

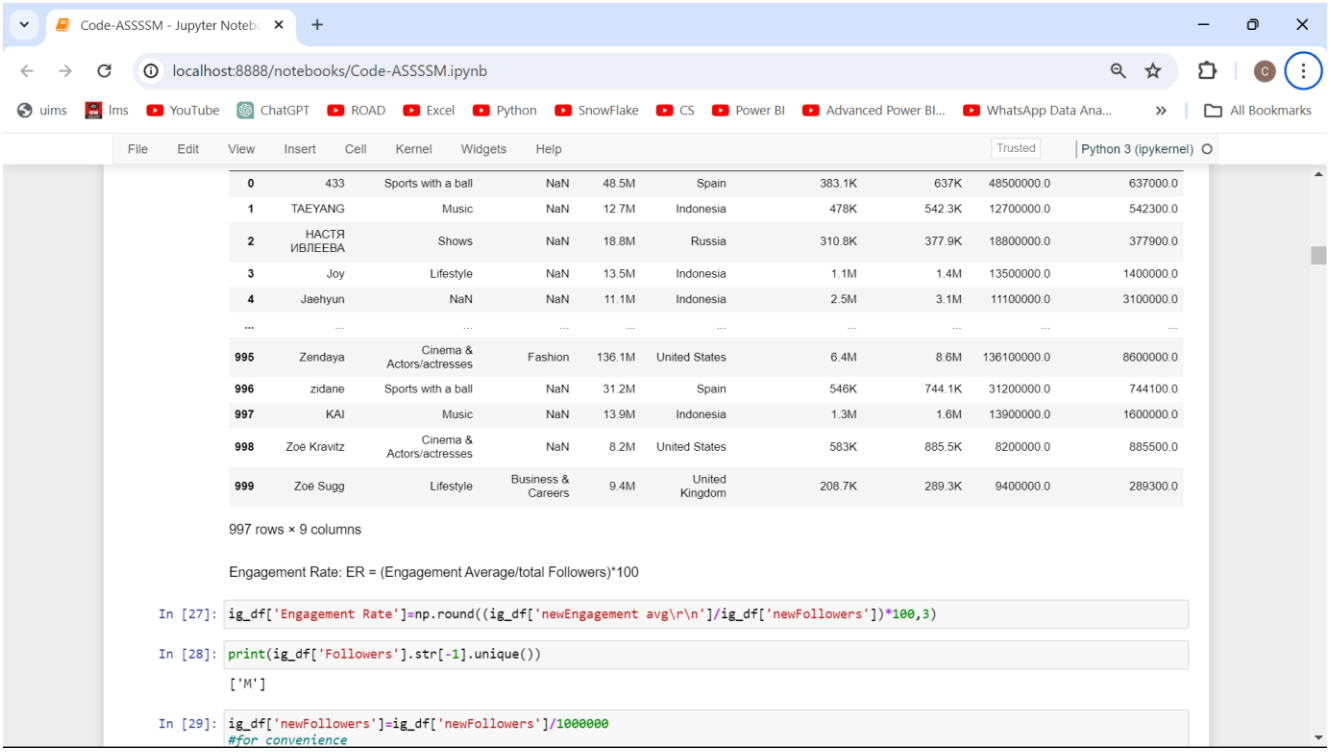
	instagram name	Category	category_2	Followers	Audience Country	Authentic engagement\r\n	newFollowers	Engagement Rate
0	433	Sports with a ball	NaN	48.5M	Spain	383.1K	48.5	1.313
1	TAEYANG	Music	NaN	12.7M	Indonesia	478K	12.7	4.270
2	НАСТЯ ИВЛЕЕВА	Shows	NaN	18.8M	Russia	310.8K	18.8	2.010
3	Joy	Lifestyle	NaN	13.5M	Indonesia	1.1M	13.5	10.370
4	Jaehyun	NaN	NaN	11.1M	Indonesia	2.5M	11.1	27.928

TOP 15 most followed accounts on Instagram

```
In [32]: ig_df.sort_values(by='newFollowers',ascending=False,ignore_index=True).iloc[0:15,[0,1,3,-1]]
```

Out[32]:

	instagram name	Category	Followers	Engagement Rate
0	Instagram	Photography	487.2M	0.096
1	Cristiano Ronaldo	Sports with a ball	419.6M	1.668
2	Kylie ♡	Fashion	323.3M	3.805
3	Leo Messi	Sports with a ball	315.4M	1.680
4	Selena Gomez	Music	308.2M	1.428
5	therock	Cinema & Actors/actresses	307M	0.129
6	Ariana Grande	Music	302.3M	1.356
7	Kim Kardashian	Fashion	296.4M	0.978



0	433	Sports with a ball	NaN	48.5M	Spain	383.1K	637K	48500000.0	637000.0
1	TAEYANG	Music	NaN	12.7M	Indonesia	478K	542.3K	12700000.0	542300.0
2	НАСТЯ ИВЛЕЕВА	Shows	NaN	18.8M	Russia	310.8K	377.9K	18800000.0	377900.0
3	Joy	Lifestyle	NaN	13.5M	Indonesia	1.1M	1.4M	13500000.0	1400000.0
4	Jaehyun	NaN	NaN	11.1M	Indonesia	2.5M	3.1M	11100000.0	3100000.0
...	...	...	...	...	...	...	...	...	...
995	Zendaya	Cinema & Actors/actresses	Fashion	136.1M	United States	6.4M	8.6M	136100000.0	8600000.0
996	zidane	Sports with a ball	NaN	31.2M	Spain	546K	744.1K	31200000.0	744100.0
997	KAI	Music	NaN	13.9M	Indonesia	1.3M	1.6M	13900000.0	1600000.0
998	Zoë Kravitz	Cinema & Actors/actresses	NaN	8.2M	United States	583K	885.5K	8200000.0	885500.0
999	Zoë Sugg	Lifestyle	Business & Careers	9.4M	United Kingdom	208.7K	289.3K	9400000.0	289300.0

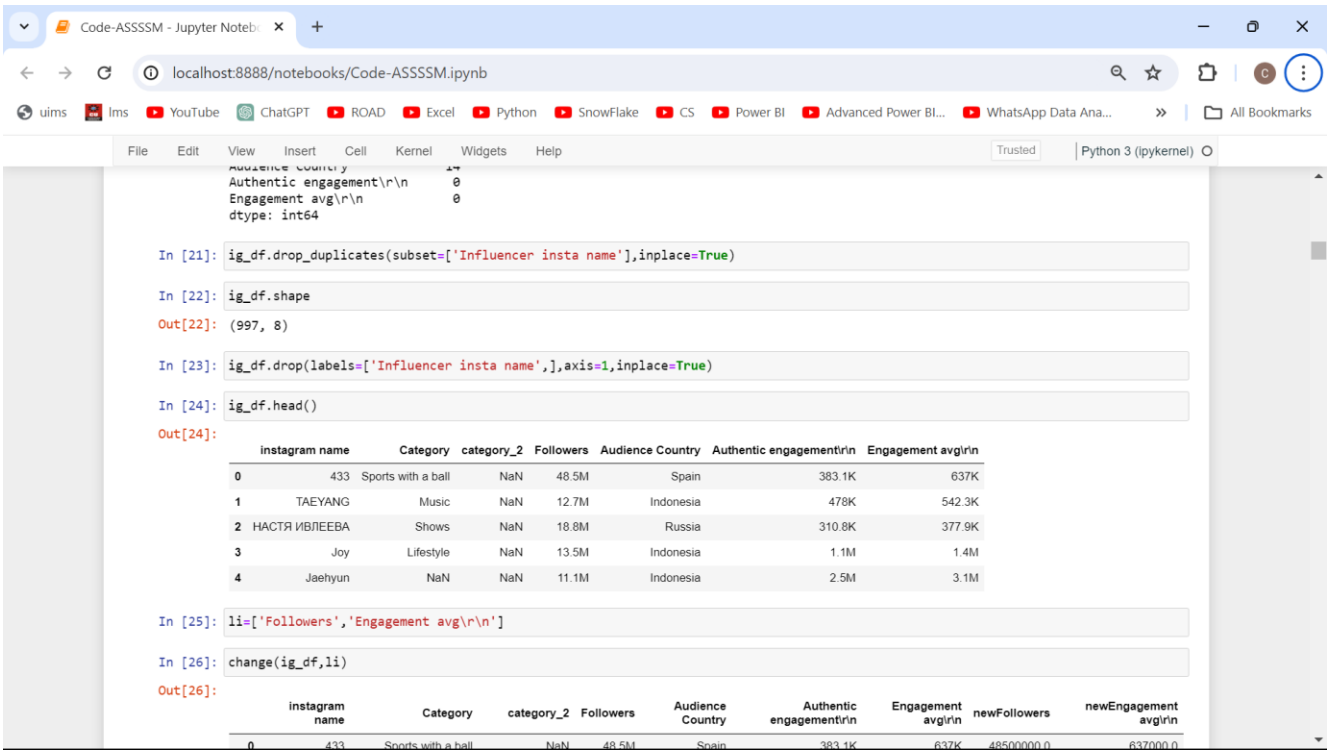
997 rows × 9 columns

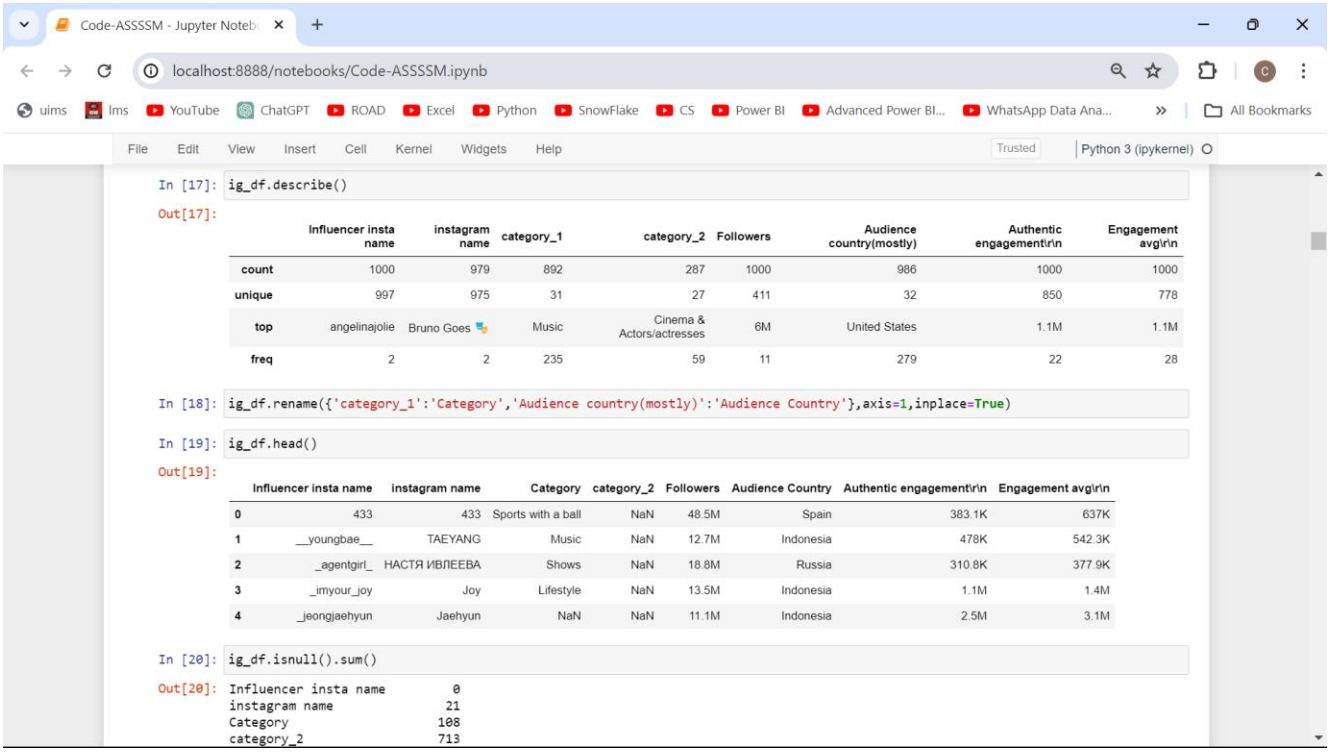
Engagement Rate: ER = (Engagement Average/total Followers)*100

```
In [27]: ig_df['Engagement Rate']=np.round((ig_df['newEngagement avg\r\n']/ig_df['newFollowers'])*100,3)
```

```
In [28]: print(ig_df['Followers'].str[-1].unique())
['M']
```

```
In [29]: ig_df['newFollowers']=ig_df['newFollowers']/1000000
#for convenience
```


Code-ASSSSM - Jupyter Noteb
localhost:8888/notebooks/Code-ASSSSM.ipynb
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Authentic engagement\r\n 0
Engagement avg\r\n 0
dtype: int64
In [21]: ig_df.drop_duplicates(subset=['Influencer insta name'],inplace=True)
In [22]: ig_df.shape
Out[22]: (997, 8)
In [23]: ig_df.drop(labels=['Influencer insta name',],axis=1,inplace=True)
In [24]: ig_df.head()
Out[24]:
instagram name Category category_2 Followers Audience Country Authentic engagement\r\n Engagement avg\r\n
0 433 Sports with a ball NaN 48.5M Spain 383.1K 637K
1 TAEYANG Music NaN 12.7M Indonesia 478K 542.3K
2 НАСТЯ ИВЛЕЕВА Shows NaN 18.8M Russia 310.8K 377.9K
3 Joy Lifestyle NaN 13.5M Indonesia 1.1M 1.4M
4 Jaehyun NaN NaN 11.1M Indonesia 2.5M 3.1M
In [25]: li=['Followers','Engagement avg\r\n']
In [26]: change(ig_df,li)
Out[26]:
instagram name Category category_2 Followers Audience Country Authentic engagement\r\n Engagement avg\r\n newFollowers newEngagement avg\r\n
0 433 Sports with a ball NaN 48.5M Spain 383.1K 637K 48500000.0 637000.0


Code-ASSSSM - Jupyter Noteb
localhost:8888/notebooks/Code-ASSSSM.ipynb
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
In [17]: ig_df.describe()
Out[17]:
Influencer insta name instagram name category_1 category_2 Followers Audience country(mostly) Authentic engagement\r\n Engagement avg\r\n
count 1000 979 892 287 1000 986 1000 1000
unique 997 975 31 27 411 32 850 778
top angelinajolie Bruno Goes Music Cinema & Actors/actresses 6M United States 1.1M 1.1M
freq 2 2 235 59 11 279 22 28
In [18]: ig_df.rename({'category_1':'Category','Audience country(mostly)':'Audience Country'},axis=1,inplace=True)
In [19]: ig_df.head()
Out[19]:
Influencer insta name instagram name Category category_2 Followers Audience Country Authentic engagement\r\n Engagement avg\r\n
0 433 433 Sports with a ball NaN 48.5M Spain 383.1K 637K
1 __youngbae__ TAEYANG Music NaN 12.7M Indonesia 478K 542.3K
2 _agentgirl_ НАСТЯ ИВЛЕЕВА Shows NaN 18.8M Russia 310.8K 377.9K
3 _imyour_joy Joy Lifestyle NaN 13.5M Indonesia 1.1M 1.4M
4 _jeongjaehyun Jaehyun NaN NaN 11.1M Indonesia 2.5M 3.1M
In [20]: ig_df.isnull().sum()
Out[20]:
Influencer insta name 0
instagram name 21
Category 108
category_2 713

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Instagram Analysis

In [15]:
```
ig_df.head()
```

Out[15]:

	Influencer insta name	instagram name	category_1	category_2	Followers	Audience country(mostly)	Authentic engagement\r\n	Engagement avg\r\n
0	433	433	Sports with a ball	NaN	48.5M	Spain	383.1K	637K
1	__youngbae__	TAEYANG	Music	NaN	12.7M	Indonesia	478K	542.3K
2	_agentgirl_	НАСТЯ ИВЛЕЕВА	Shows	NaN	18.8M	Russia	310.8K	377.9K
3	_imyour_joy	Joy	Lifestyle	NaN	13.5M	Indonesia	1.1M	1.4M
4	_jeongjaehyun	Jaehyun	NaN	NaN	11.1M	Indonesia	2.5M	3.1M

In [16]:
```
ig_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 8 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Influencer insta name      1000 non-null   object
 1   instagram name             979 non-null    object
 2   category_1                 892 non-null    object
 3   category_2                 287 non-null    object
 4   Followers                  1000 non-null   object
 5   Audience country(mostly)   986 non-null    object
 6   Authentic engagement
    1000 non-null   object
 7   Engagement avg
          1000 non-null   object
dtypes: object(8)
memory usage: 62.6+ KB
```

9	Kris HC	43.3M

In [13]:
```
tt_df.sort_values(by = 'Views avg.', ascending = False, ignore_index=True).iloc[0:10, [1,2,3]]
```

Out[13]:

	Tiktok name	Subscribers count	Views avg.
0	jajad5	630.8K	982.4K
1	Ph1LzA	2.2M	980.5K
2	Gregisms' Neighborhood	2.2M	975.5K
3	sinadeinert	5.5M	967.7K
4	LA MARA	6.1M	959.7K
5	Teamsport_Philipp	23.2K	957.2K
6	T	21.1K	951.6K
7	【コムドット】ゆうた	1M	949.1K
8	startup.mp4	98K	944.3K
9	Sayden	3.3M	938.9K

In [14]:
```
tt_df.sort_values(by = 'Shares avg', ascending = False, ignore_index=True).iloc[0:10, [1,2,6]]
```

Out[14]:

	Tiktok name	Subscribers count	Shares avg
0	Nate	306.4K	9K
1	Colson	3.4M	9K
2	pablitocastilloo	4.5M	9K
3	Game of Thrones	647.3K	9K

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

2	kiet.ac.quy	Kiệt Ák Wỷ	2.1M	20.7M	3.5M	38.8K	33.9K	2100000.0
3	charlidamelio	charli d'amelio	135.4M	18.7M	2.6M	54.7K	35.2K	135400000.0
4	luvadepedreiro	Iran Ferreira (Lai)	11.4M	24.8M	2.6M	32.7K	26.8K	11400000.0
...	...	...	...	...	...	...	...	...
995	nicolebernaal	nicolebernaal	4M	2.2M	351.3K	957	195	4000000.0
996	bellaretamosa	bella	5.1M	2.5M	340.4K	901	145	5100000.0
997	tunico80	Antonio Tonon	5.8M	1M	206.8K	2K	2.1K	5800000.0
998	armon.warren	Armoney	1.2M	1.9M	300.8K	904	630	1200000.0
999	soanhvadiephihi	Soanh x Diệp	1.6M	2.6M	273.2K	720	530	1600000.0

1000 rows × 8 columns

In [12]:
```
tt_df.sort_values(by = 'newSubscribers count', ascending = False, ignore_index=True).iloc[0:10, [1,2]]
```

Out[12]:

	Tiktok name	Subscribers count
0	charli d'amelio	135.4M
1	Khabane lame	135.2M
2	Bella Poarch	88.5M
3	Addison Rae	87.3M
4	Will Smith	67.4M
5	Kimberly Loaiza	61M
6	dixie	57.2M
7	Loren Gray	54.3M
8	Dominik	50.7M

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

In [10]:
```
tt_df.drop_duplicates()
```

Out[10]:

	Tiktoker name	Tiktok name	Subscribers count	Views avg.	Likes avg	Comments avg.	Shares avg
0	ekin.721	MOMO's	221.7K	26M	2.8M	29.4K	116.4K
1	dojacat	Doja Cat	22.2M	25.4M	5M	36.7K	46.8K
2	kiet.ac.quy	Kiệt Ák Wỷ	2.1M	20.7M	3.5M	38.8K	33.9K
3	charlidamelio	charli d'amelio	135.4M	18.7M	2.6M	54.7K	35.2K
4	luvadepedreiro	Iran Ferreira (Lai)	11.4M	24.8M	2.6M	32.7K	26.8K
...	...	...	...	...	...	...	...
995	nicolebernaal	nicolebernaal	4M	2.2M	351.3K	957	195
996	bellaretamosa	bella	5.1M	2.5M	340.4K	901	145
997	tunico80	Antonio Tonon	5.8M	1M	206.8K	2K	2.1K
998	armon.warren	Armoney	1.2M	1.9M	300.8K	904	630
999	soanhvadiephihi	Soanh x Diệp	1.6M	2.6M	273.2K	720	530

989 rows × 7 columns

In [11]:
```
change(tt_df, ['Subscribers count'])
```

Out[11]:

	Tiktoker name	Tiktok name	Subscribers count	Views avg.	Likes avg	Comments avg.	Shares avg	newSubscribers count
0	ekin.721	MOMO's	221.7K	26M	2.8M	29.4K	116.4K	221700.0
1	dojacat	Doja Cat	22.2M	25.4M	5M	36.7K	46.8K	22200000.0
2	kiet.ac.quy	Kiệt Ák Wỷ	2.1M	20.7M	3.5M	38.8K	33.9K	2100000.0
3	charlidamelio	charli d'amelio	135.4M	18.7M	2.6M	54.7K	35.2K	135400000.0

File Edit View Insert Cell Kernel Widgets Help

In [10]:
```
tt_df.drop_duplicates()
```

Out[10]:

	Tiktoker name	Tiktok name	Subscribers count	Views avg.	Likes avg	Comments avg.	Shares avg
0	ekin.721	MOMO's	221.7K	26M	2.8M	29.4K	116.4K
1	dojacat	Doja Cat	22.2M	25.4M	5M	36.7K	46.8K
2	kiet.ac.quy	Kiệt Ắk Wỷ	2.1M	20.7M	3.5M	38.8K	33.9K
3	charlidamelio	charli d'amelio	135.4M	18.7M	2.6M	54.7K	35.2K
4	luvadepedreiro	Iran Ferreira (Lai)	11.4M	24.8M	2.6M	32.7K	26.8K
...	...	...	...	...	...	...	...
995	nicolebernaal	nicolebernaal	4M	2.2M	351.3K	957	195
996	bellaretamosa	bella	5.1M	2.5M	340.4K	901	145
997	tunico80	Antonio Tonon	5.8M	1M	206.8K	2K	2.1K
998	armon.warren	Armoney	1.2M	1.9M	300.8K	904	630
999	soanhvadiephihi	Soanh x Diệp 💎	1.6M	2.6M	273.2K	720	530

989 rows × 7 columns

In [11]:
```
change(tt_df, ['Subscribers count'])
```

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Importing Libraries

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

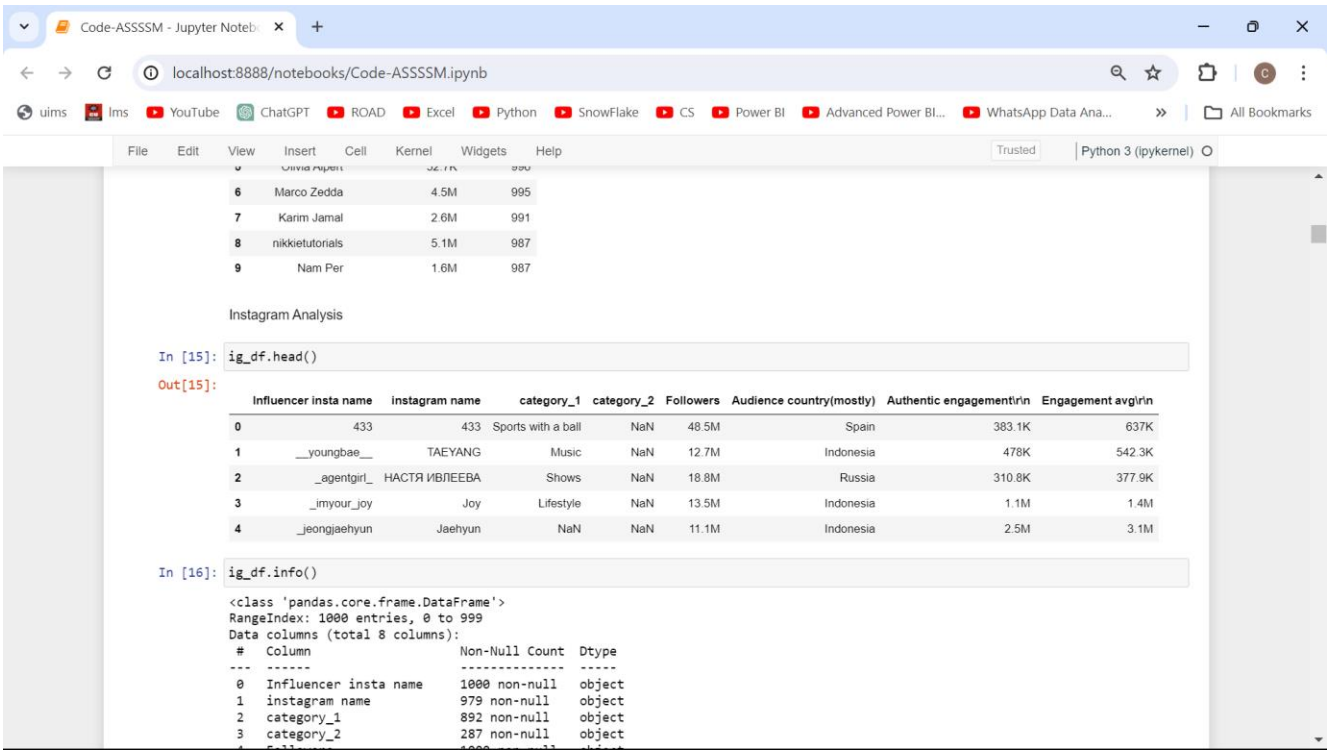
Reading te datasets

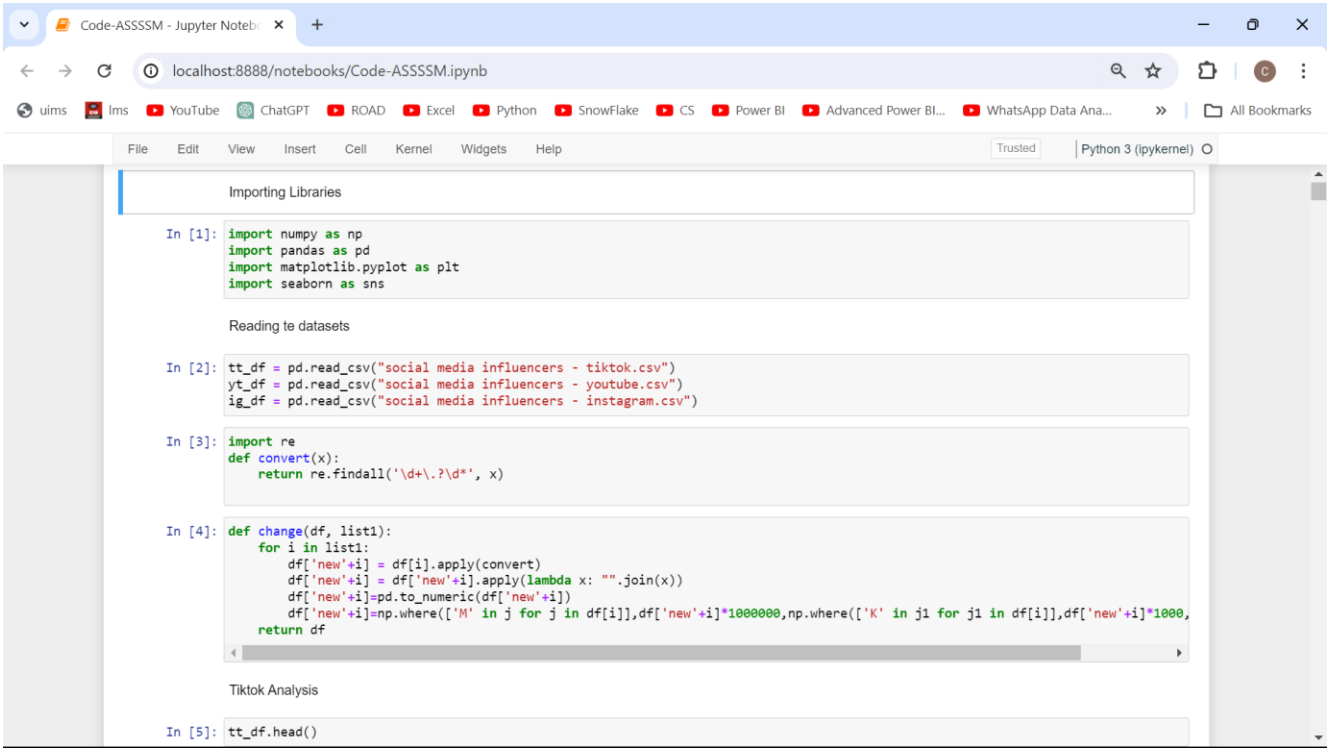
In [2]:
```
tt_df = pd.read_csv("social media influencers - tiktok.csv")
yt_df = pd.read_csv("social media influencers - youtube.csv")
ig_df = pd.read_csv("social media influencers - instagram.csv")
```

In [3]:
```
import re
def convert(x):
    return re.findall('\d+\.?\d*', x)
```

In [4]:
```
def change(df, list1):
    for i in list1:
        df['new'+i] = df[i].apply(convert)
        df['new'+i] = df['new'+i].apply(lambda x: "".join(x))
        df['new'+i]=pd.to_numeric(df['new'+i])
        df['new'+i]=np.where(['M' in j for j in df[i]],df['new'+i]*1000000,np.where(['K' in j1 for j1 in df[i]],df['new'+i]*1000,
    return df
```

Tiktok Analysis


Code-ASSSSM - Jupyter Noteb
localhost:8888/notebooks/Code-ASSSSM.ipynb
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
6 Marco Zedda 4.5M 995
7 Karim Jamal 2.6M 991
8 nikkietutorials 5.1M 987
9 Nam Per 1.6M 987
Instagram Analysis
In [15]: ig_df.head()
Out[15]:
Influencer insta name | instagram name | category_1 | category_2 | Followers | Audience country(mostly) | Authentic engagement\n | Engagement avg\n
0 433 433 Sports with a ball NaN 48.5M Spain 383.1K 637K
1 __youngbae__ TAEYANG Music NaN 12.7M Indonesia 478K 542.3K
2 _agentgirl_ НАСТЯ ИВЛЕЕВА Shows NaN 18.8M Russia 310.8K 377.9K
3 _imyour_joy Joy Lifestyle NaN 13.5M Indonesia 1.1M 1.4M
4 _jeongjaehyun Jaehyun NaN NaN 11.1M Indonesia 2.5M 3.1M
In [16]: ig_df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 8 columns):
Column Non-Null Count Dtype
0 Influencer insta name 1000 non-null object
1 instagram name 979 non-null object
2 category_1 892 non-null object
3 category_2 287 non-null object


Code-ASSSSM - Jupyter Noteb
localhost:8888/notebooks/Code-ASSSSM.ipynb
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Importing Libraries
In [1]:
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
Reading te datasets
In [2]:
tt_df = pd.read_csv("social media influencers - tiktok.csv")
yt_df = pd.read_csv("social media influencers - youtube.csv")
ig_df = pd.read_csv("social media influencers - instagram.csv")
In [3]:
import re
def convert(x):
return re.findall('\d+\.?\d*', x)
In [4]:
def change(df, list1):
for i in list1:
df['new'+i] = df[i].apply(convert)
df['new'+i] = df['new'+i].apply(lambda x: "".join(x))
df['new'+i]=pd.to_numeric(df['new'+i])
df['new'+i]=np.where(['M' in j for j in df[i]],df['new'+i]*1000000,np.where(['K' in j1 for j1 in df[i]],df['new'+i]*1000,
return df
Tiktok Analysis
In [5]: tt_df.head()

DASHBOARD FOR COMPARATIVE ANALYSIS:

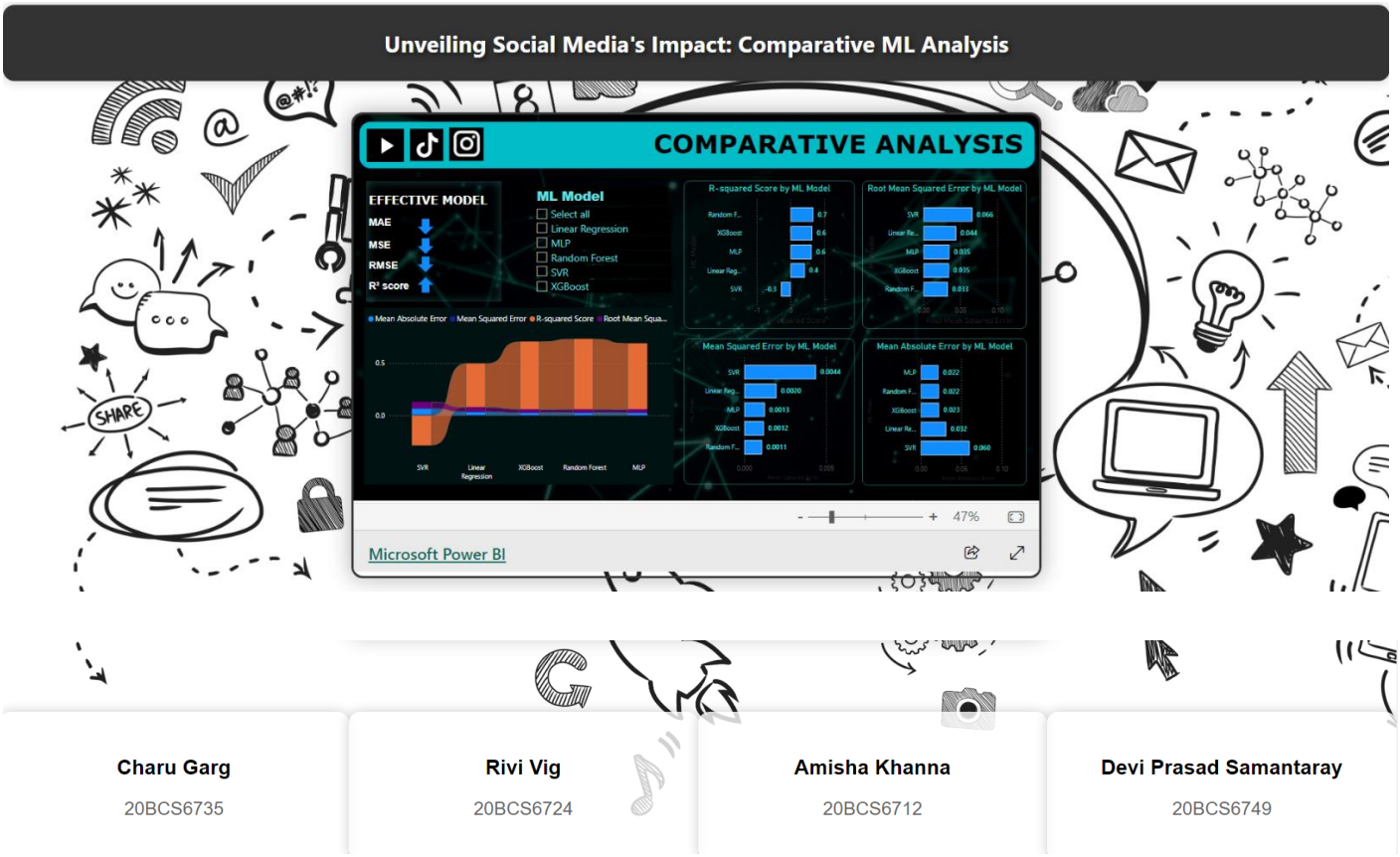


Figure 9 Dashboard for comparative Analysis

BEST MODEL: RANDOM FOREST	
Mean Absolute Error	0.02197470057510002
Mean Squared Error	0.0011007568227802986
Root Mean Squared Error	0.03317765547443488
R-squared Score	0.6763174357035447

Figure 9: Key performance Indicator of Random Forest

A comparative analysis was undertaken to evaluate the efficacy of five distinct regression algorithms: linear regression, XGBoost, random forest, multilayer perceptron (MLP), and support vector regression (SVR). The objective was to evaluate the predictive precision and resilience of each model in forecasting a target variable using a series of predictor variables.

Summary of Comparative Study:

Each model's effectiveness was calculated using various system of measurement, such as mean squared error (MSE), mean absolute error (MAE), and the R-squared (R^2) coefficient of determination.

The outcomes of the comparative analysis unveiled the following findings:

Linear Regression: Linear regression, while being a simple and interpretable model, exhibited limited predictive accuracy compared to more complex algorithms. It struggled to capture non-linear relationships between predictor and target variables.

XGBoost: XGBoost, a powerful gradient boosting algorithm, demonstrated superior predictive performance compared to linear regression. It effectively captured non-linear relationships and interactions between variables, resulting in lower MSE and MAE values.

Random Forest: Random forest, an ensemble learning technique, also performed well in predicting the target variable. It leveraged the collective wisdom of multiple decision trees to mitigate overfitting and improve generalization performance

Multilayer Perceptron (MLP): The multilayer perceptron, a type of artificial neural network, showed competitive performance, particularly in capturing complex non-linear patterns in the data. However, its training process could be sensitive to hyperparameter tuning and prone to overfitting with insufficient regularization.

Support Vector Regression (SVR): SVR, a regression variant of support vector machines, exhibited mixed performance depending on the dataset characteristics and kernel selection. While SVR could effectively model non-linear relationships, it required careful tuning of hyperparameters and kernel functions.

Overall, the comparative study demonstrated that XGBoost and random forest generally outperformed linear regression, MLP, and SVR in terms of predictive accuracy and robustness across a range of datasets.

Model	Best Score
Random Forest Regressor	-0.002470787692491047
Linear Regression	-13841356711.074116
XG Boost	-0.0024645940119547077
SVR	-0.0054478303369689
MLP Regressor	-0.0027734980453929903

Figure 10: Best scored of different algorithms

4.3 Report Preparation

Our signifies the critical phase where the culmination of research, design, implementation, and analysis is transformed into a comprehensive and structured document. This report serves as the primary medium for communicating our findings, methodologies, and outcomes related to understanding the influence of social media on consumer behaviour. It encapsulates the entire journey, from the inception of the project to its execution and results.

The report is meticulously structured to provide readers with a clear understanding of the project's objectives, the methodologies employed, the challenges encountered, and, most importantly, the insights gained. It incorporates a wealth of data, including consumer behaviour analysis results, social media engagement metrics, and practical applications of our findings. The document also highlights the ethical considerations, privacy implications, and societal impact of our project.

Our report preparation process emphasizes clarity, accuracy, and accessibility, ensuring that the knowledge and advancements made in the realm of social media influence on consumer behaviour are shared effectively with businesses, marketers, policymakers, and researchers in the field of digital marketing and consumer psychology.

CHAPTER-5

CONCLUSION AND FUTURE WORK

Through a comprehensive exploration of social media platforms and their impact on consumer behaviour, this project has illuminated the profound influence of digital connectivity on the modern consumer journey. The findings of this project underscore the pivotal role of social media in shaping consumer preferences, attitudes, and purchasing decisions.

Key Findings:

1. Decryption of Key Factors Driving Consumer Behaviour:

- The project has identified and decrypted key factors driving consumer behaviour on social media platforms. This includes analyzing user engagement metrics, sentiment analysis, and behavioural patterns to gain insights into consumer preferences and decision-making processes.

2. Importance of Mixed-Methods Approach:

- By employing a mixed-methods approach integrating qualitative and quantitative analyses, the project has facilitated a nuanced exploration of social media influence.

Implications for Businesses:

1. Tailoring Marketing Strategies:

- Businesses can leverage insights from the project to tailor their marketing strategies according to the preferences and behaviours of their target audience on social media platforms. Understanding user engagement metrics, sentiment dynamics, and behavioural patterns can inform the development of more effective and targeted marketing campaigns.

2. Strategic Decision-Making:

- The insights garnered from the project can inform strategic decision-making processes for businesses across various industries. By understanding consumer trends and preferences on social media, organizations can make informed decisions regarding product development, brand positioning, and market expansion strategies.

Future Directions:

1. Leveraging Advanced Analytics Techniques:

- Moving forward, organizations can leverage advanced analytics techniques and predictive modelling to anticipate consumer trends and optimize their online presence.

2. Fostering Meaningful Engagement:

- The project highlights the importance of fostering meaningful engagement with the target audience on social media platforms. By creating content that resonates with consumers and actively engaging with them online, businesses can build stronger relationships and enhance brand loyalty.

In summary, the insights gleaned from this project have significant implications for businesses seeking to navigate the digital landscape effectively. By understanding the intricate dynamics of social media influence on consumer behaviour, organizations can develop more targeted and impactful strategies to engage with their audience and drive business growth.

REFERENCES

[1]. Prakash Singh Measuring social media impact on Impulse Buying Behaviour Journal: Cogent Business & Management Year: 2023

[2]. P. Grover, A.K. Kar and Y. Dwivedi International Journal of Information Management Data Insights 2 (2022) 100116

[3]. N K Hoi European Journal of Business and Management Research. (2023). Short Videos, Big Decisions: A Preliminary Study of the Impact of Short Videos on Consumers' Purchase Intention. Vol 8, Issue 3, Pages 73-81

[4]. Sachin Gupta International Journal of Creative Research Thoughts (IJCRT), Volume 8, Issue 6, June 2020, ISSN: 2320-2882, IJCRT2006265.

[5]. J.-F.F.-G., P.M.-B., M.P.-L., J.R.-R. (2020). Title of the Paper. Sustainability, 12(15), 1506

[6]. Sykora, M., Elayan, S., Hodgkinson, I. R., Jackson, T. W., & West, A. (2022). The power of emotions: Leveraging user generated content for customer experience management. Journal of Business Research, 144, 997–1006

[7]. Hasan Mahmud: Factors influencing algorithm aversion: A systematic literature review Technological Forecasting and Social Change Volume 175, February 2022, 121390

[8]. Becker, L., & Jaakkola, E. (2020). Customer experience: Fundamental premises and implications for research. Journal of the Academy of Marketing Science, 48(4), 630–648

[9]. Borg, A., & Boldt, M. (2020). Using VADER sentiment and SVM for predicting customer response sentiment. Expert Systems with Applications, 162, Article 113746

[10]. Canhoto, A. I., & Clark, M. (2013). Customer service 140 characters at a time: The users' perspective. Journal of Marketing Management, 29(5–6), 522–544

[11]. .Ebrahimi, P., Basirat, M., Yousefi, A., Nekmahmud, M., Gholampour, A., & Fekete-Farkas, M. (2022). Social Networks Marketing and Consumer Purchase Behavior: The Combination of SEM and

Unsupervised Machine Learning Approaches. Big Data Cogn. Comput., 6, 35. https://doi.org/10.3390/bdcc6020035

[12]. .Bui Thanh Khoa, Tran Trong Huynh. (2023). "Enhancing Electronic Consumer Loyalty and Online Trust through Social Media Marketing." International Journal of Digital Marketing and E-Commerce, 9(2), 1-10

[13]. Meier, Raphael. "Social Media Influence Operations." arXiv:2309.03670v1 [cs.CY] 7 Sep 2023