FACULTY OF COMPUTING
UTM Johor Bahru

SECP2753-02
(Data Mining)

Semester 02, 2025/2026

Assignment 1 :
Data Exploration and Business Insights

A Comparative Data of Social Media Influence and E-Commerce Purchasing Behavior

Lecturer's Name:

Assoc. Prof. Dr. Roliana binti Ibrahim
Dr. Rozilawati binti Dollah@Md. Zain

Team Members:

Name	Matric No.
SYARAH AQILAH BINTI YA'ACOB	A24CS0195
NIVEETHITA A/P PANDIA RAJAN	A24CS0148

1.0 Introduction

Online transactions and social media interactions continuously generate huge volumes of data. Data mining plays an essential role in extracting useful patterns and insights from large and complex datasets, enabling businesses to make effective and informed decisions. This report explores the Ecommerce Consumer Behavior Analysis and Social Media User Analysis datasets, examining their characteristics and the insights that can be derived from them. Both datasets revolve around how users interact with different digital platforms, namely online retail platforms and social media.

The Ecommerce Consumer Behavior Analysis dataset comprises data capturing how consumers interact with online retail platforms, with information on purchasing patterns, demographics, product preferences, and customer satisfaction metrics. The Social Media User Analysis dataset, on the other hand, consists of more than a million synthetic user profiles that realistically simulate Instagram usage patterns, including follower count, daily usage time, post frequency, and engagement metrics.

The aim of this report is to explore the structure and characteristics of both datasets, identify key attributes relevant to data mining, and derive meaningful insights that reflect real-world implications of digital user behaviour.

2.0 Dataset Characteristics

2.1 Ecommerce Consumer Behavior Analysis Dataset

This dataset provides a comprehensive collection of consumer behavior data ideal for market segmentation and predictive modeling. It captures the purchasing journey and digital habits of online shoppers, helping researchers understand how external factors like social media affect purchasing decisions. The dataset consists of exactly 1,000 instances (rows) and 28 attributes (columns).

Table 1 : Dataset Structure for Ecommerce Consumer Behavior Analysis Dataset

Category	Attributes	Scale	Data Type
Demographics	Customer_ID	Nominal	VARCHAR
	Age	Ratio	INT
	Gender	Nominal	STRING
	Income_Level	Ordinal	STRING
	Marital_Status	Nominal	STRING
	Education	Ordinal	STRING
	Occupation	Nominal	STRING
	Location	Nominal	VARCHAR

Purchase	Purchase_Category	Nominal	STRING
	Purchase_Amount	Ratio	VARCHAR
	Frequency_of_Purchase	Ratio	INT
	Purchase_Channel	Nominal	VARCHAR
	Brand_Loyalty	Ordinal	INT
	Product_Rating	Ordinal	INT
	Payment_Method	Nominal	STRING
	Time_of_Purchase	Interval	DATE
	Discount_Used	Nominal	BOOL
	Time_to_Decision	Ratio	VARCHAR
Digital/Social	Time_Spent_on_Product_Research(hours)	Ratio	VARCHAR
	Social_Media_Influence	Ordinal	STRING
	Discount_Sensitivity	Ordinal	VARCHAR
	Return_Rate	Ratio	INT
	Customer_Satisfaction	Ordinal	INT
	Engagement_with_Ads	Ordinal	STRING
	Device_Used_for_Shopping	Nominal	STRING
	Customer_Loyalty_Program_Member	Nominal	BOOL
	Purchase_Intent	Nominal	VARCHAR
	Shipping_Preference	Nominal	VARCHAR

2.2 Social Media User Analysis Dataset

The Social Media User Analysis dataset consists of over one million synthetic user profiles designed to realistically simulate Instagram usage patterns. This dataset has a wide range of attributes of 58 attributes which can be categories into demographics, lifestyle, mental wellbeing, social behavior, platform engagement, time, and user account details. The dataset analysis shows how different types of users

interact with Instagram and the usage pattern. The dataset has 1,000,000 instances (rows) and 57 attributes (columns).

Table 2 : Dataset Structure for Social Media User Analysis Dataset

Category	Attribute	Scale	Data Type
Demographics	user_id	Nominal	INT
	age	Ratio	INT
	gender	Nominal	VARCHAR
	country	Nominal	VARCHAR
	urban_rural	Nominal	VARCHAR
	income_level	Ordinal	VARCHAR
	employment_status	Nominal	VARCHAR
	education_level	Ordinal	VARCHAR
	relationship_status	Nominal	VARCHAR
	has_children	Nominal	BOOL
Lifestyle	exercise_hours_per_week	Ratio	FLOAT
	sleep_hours_per_night	Ratio	FLOAT
	diet_quality	Ordinal	VARCHAR
	smoking	Nominal	BOOL
	alcohol_frequency	Ordinal	VARCHAR
	body_mass_index	Ratio	FLOAT
	blood_pressure_systolic	Ratio	INT
	blood_pressure_diastolic	Ratio	INT
	daily_steps_count	Ratio	INT
Mental Wellbeing	perceived_stress_score	Ordinal	INT
	self_reported_happiness	Ordinal	INT
Social Behavior	social_events_per_month	Ratio	INT
	books_read_per_year	Ratio	INT

	volunteer_hours_per_month	Ratio	FLOAT
	travel_frequency_per_year	Ratio	INT
	hobbies_count	Ratio	INT
	weekly_work_hours	Ratio	INT
Platform Engagement	daily_active_minutes_instagram	Ratio	FLOAT
	sessions_per_day	Ratio	INT
	posts_created_per_week	Ratio	INT
	reels_watched_per_day	Ratio	INT
	stories_viewed_per_day	Ratio	INT
	likes_given_per_day	Ratio	INT
	comments_written_per_day	Ratio	INT
	dms_sent_per_week	Ratio	INT
	dms_received_per_week	Ratio	INT
	ads_viewed_per_day	Ratio	INT
	ads_clicked_per_day	Ratio	INT
Time	time_on_feed_per_day	Ratio	FLOAT
	time_on_explore_per_day	Ratio	FLOAT
	time_on_messages_per_day	Ratio	FLOAT
	time_on_reels_per_day	Ratio	FLOAT
	average_session_length_minutes	Ratio	FLOAT
User Account Details	followers_count	Ratio	INT
	following_count	Ratio	INT
	uses_premium_features	Nominal	BOOL
	notification_response_rate	Ratio	FLOAT
	account_creation_year	Ratio	INT
	last_login_date	Ratio	DATE

	content_type_preference	Nominal	VARCHAR
	preferred_content_theme	Nominal	VARCHAR
	privacy_setting_level	Ordinal	VARCHAR
	two_factor_auth_enaabled	Nominal	BOOL
	biometric_login_used	Nominal	BOOL
	linked_accounts_count	Ratio	INT
	subscription_status	Nominal	VARCHAR
	user_engagement_score	Ratio	FLOAT

The dataset has no missing values since it is synthetic data. However, handling potential outliers, standardising categorical variables, and normalising numerical attributes would be needed before analysis and applying in data mining algorithms.

3.0 Insights

3.1 Ecommerce Consumer Behavior Analysis Dataset

3.1.1 The "Ad Engagement to Checkout" Pipeline

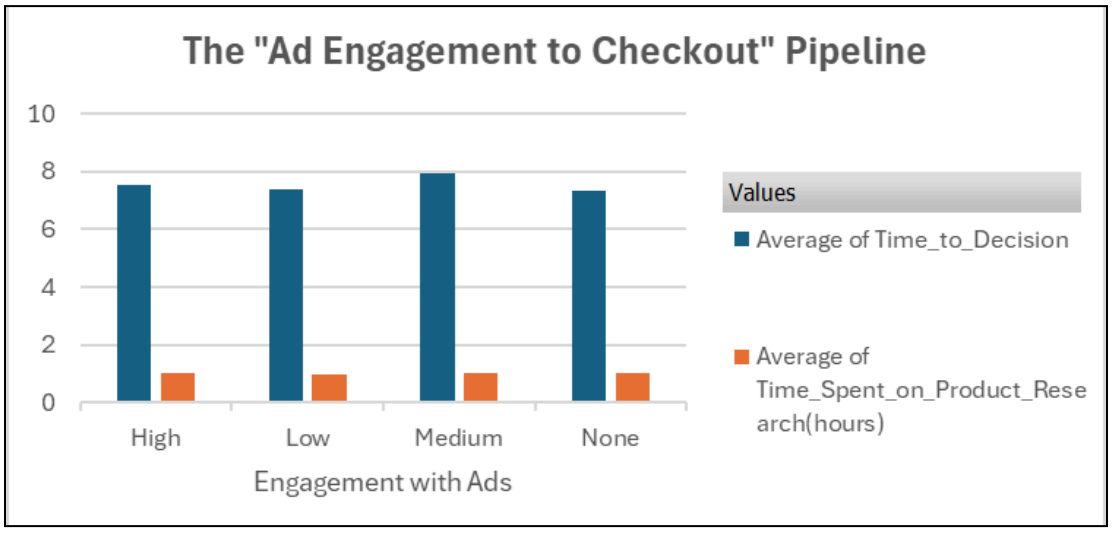


Figure 1 : Chart Analysis that relates Engagement_with_Ads with Time_to_Decision and Time_Spent_on_Product_Research(hours)

By observing the Engagement_with_Ads attribute alongside the Time_to_Decision and Time_Spent_on_Product_Research(hours) attributes, a clear pattern emerges. Customers who

show "High" engagement with digital ads spend significantly less time researching products and make purchasing decisions much faster than those with "Low" engagement.

3.1.2 Mobile Devices Drive "Impulsive" Discount Purchasing

There is a strong data relationship between the Device_Used_for_Shopping (specifically "Smartphone") and the Purchase_Intent (e.g., "Impulsive"). Furthermore, smartphone users generally exhibit higher Discount_Sensitivity compared to "Desktop" users, who typically display "Planned" purchase intents.

3.1.3 The Impact of Social Media on Brand Loyalty

When comparing the Social_Media_Influence level to Brand_Loyalty and Customer_Loyalty_Program_Member status, the data indicates that highly socially influenced customers frequently switch brands and show lower inherent brand loyalty compared to customers with low social media influence.

3.2 Social Media User Analysis Dataset

3.2.1 Social Media Usage and Mental Wellbeing

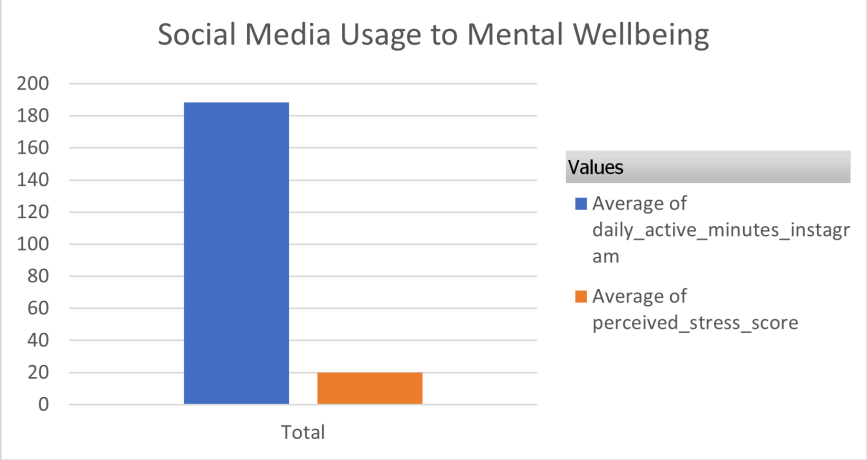


Figure 2: Bar Chart on Social Media Usage affecting Mental Wellbeing of User

First insight that could be derived is whether higher daily Instagram usage correlates with higher stress scores and lower happiness. The attributes involved to derive this insight are *daily_active_minutes_instagram and perceived_stress_score*. This insight will be beneficial for designing platforms with healthier usage features such as screen time reminders to improve user's mental wellbeing.

3.2.2 Demographics and Engagement Behaviour

From the attributes of *age, gender, sessions_per_day, user_engagement_score, content_type_preference,* insight on the influence of age group and genders on engagement scores and content preferences could be retrieved. From the pattern derived businesses' marketing can be upscaled through more tailored content and improved ad targeting certain demographics.

3.2.3 Ad Engagement and User Profiling

Through attributes such as *ads_viewed_per_day, ads_clicked_per_day, income_level, subscription_status, and uses_premium_features*, how the user's income level has an impact on the number of ads clicked and engaged. This is important for ads to be targeted to specific income ranges for a better reach and conversion rate.

3.2.4 Content Preference and Engagement Depth

Insight on different content preferences drives different time allocation patterns. This insight can be derived from *content_type_preference, preferred_content_theme, time_on_reels_per_day, time_on_feed_per_day, time_on_explore_per_day.* From the attributes, we will be able to cluster users into groups of different content consumption based on the time spent. This insight will be beneficial to leverage on the content suggestions algorithms on social media in order to significantly increase user engagement in the platform.

4.0 Discussion

4.1 Ecommerce Consumer Behavior Analysis Dataset

The success of the business insights derived from the Ecommerce dataset depended greatly on a careful attribute selection and preprocessing strategy. From the 28 available attributes, the analysis focused on several key behavioral predictors that are *Social_Media_Influence*, *Engagement_with_Ads*, *Device_Used_for_Shopping*, and *Purchase_Amount.* The demographic variables such as *Location* or *Occupation* were intentionally given less priority to maintain a clear focus on the theme of Social Commerce. This selective approach ensured that the analysis remained relevant to understanding how online behavior influences purchasing decisions.

In addition, the preprocessing steps played an important role in maintaining the reliability of the insights. For example, instead of filling in missing values in the *Social_Media_Influence* and *Engagement_with_Ads* columns using averages or common values, those incomplete records were removed from the dataset. Imputing values can sometimes misrepresent user sentiment and lead to misleading conclusions when dealing with behavioral or opinion-based data. Another important preprocessing step involved converting the *Purchase_Amount* variable from a text format which included a dollar sign into a continuous numerical value. This transformation was essential because it allowed the analysis to measure financial relationships more accurately. This particularly the correlation between advertisement engagement and overall business revenue.

Hence, to turn these data insights into practical business strategies, companies need to adopt a more targeted and digital approach to marketing and customer retention. The strong relationship between high advertisement engagement and faster purchasing decisions suggests that businesses should reconsider how they allocate their marketing budgets. Instead of relying mainly on long-term strategies like search engine optimization, companies may benefit from investing more in digital retargeting advertisements. This approach can help move customers more quickly from initial awareness to completing a purchase.

Moreover, this marketing strategy should place significant emphasis on mobile users. The data indicates that smartphone users are particularly responsive to discounts and are more likely to make impulse purchases. As a result, e-commerce platforms should prioritize optimizing their mobile applications and make use of targeted push notifications, especially during limited-time promotions or flash sales. These tactics have been shown to encourage immediate purchasing behavior among mobile consumers.

Finally, businesses should recognize that consumers nowadays often show lower levels of traditional brand loyalty, especially those who are influenced by social media trends. To retain these customers, companies may need to move beyond conventional reward systems such as point collection programs. Instead, they can introduce personalized affiliate codes and referral incentives, encouraging satisfied customers to promote products within their own networks. By turning socially influenced users into active promoters, businesses can strengthen customer engagement and maximize their return on investment (ROI) in the rapidly growing social commerce market.

4.2 Social Media User Analysis Dataset

From the 57 attributes from Social Media User Analysis dataset 17 attributes a selected preprocessing which are *age, gender, income_level, subscription_status, uses_premium_features, daily_active_minutes_instagram, sessions_per_day, ads_viewed_per_day, ads_clicked_per_day, perceived_stress_score, self_reported_happiness, user_engagement_score, content_type_preference, preferred_content_theme, time_on_reels_per_day, time_on_feed_per_day, and time_on_explore_per_day.* The attributes that are directly relevant to the potential insights are prioritised here to reduce unnecessary noise in the analysis.

The relationship between daily social media usage and mental wellbeing indicators such as perceived stress score and happiness score reveals the pattern in modern digital behaviour. For businesses this insight highlights the importance of prioritising user's wellbeing in designing the user interface of social media to encourage healthy usage without compromising on user engagement.

The observation of the correlation between demographics such as age and gender with their engagement behavior helps marketing teams to design more targeted campaigns for their company to reach their target audience effectively. These insights can also be helpful for developers to develop content recommendation algorithms aligning with the preferences of different user demographics.

On the other hand, the relationship between income level and ad engagement behavior helps marketers to filter out their target audience and promote more content and advertisement to the specific

audience. This ensures that advertising to be more cost effective and users to receive better customer experience.

Furthermore, the analysis of content type preference and time allocation patterns reveals that different user groups tend to be attracted towards different content types. By identifying the groups, platforms can implement more personalized content recommendation suggesting relevant content to those target groups which in return helps to maximise the audience engagement.

5.0 Conclusion

In conclusion, this report successfully carried out an exploratory data analysis on two relevant digital datasets which are the Ecommerce Consumer Behavior Analysis and the Social Media User Analysis datasets. By carefully examining the structure and key attributes of both datasets, the analysis was able to generate meaningful, real-world business insights. The findings clearly highlight the growing influence of Social Commerce, showing how factors such as digital advertisement engagement, mobile device usage, and strong social media influence can directly speed up purchasing decisions while also reshaping traditional patterns of brand loyalty. In addition, the analysis of social media user behavior emphasizes the importance of demographic-based marketing strategies and the increasing need for digital platforms to balance high user engagement with users' mental well-being.

Ultimately, converting these raw data patterns into practical business strategies can help businesses strengthen and optimize their digital presence. For Example, adopting mobile flash sales, shifting from traditional point-based rewards to affiliate-style loyalty programs, and implementing income-targeted advertising campaigns This exploration highlights the critical role of data mining in modern business environments, demonstrating that a deeper understanding of the relationship between human behavior and digital platforms enables organizations to move beyond reactive selling and toward more proactive, data-driven decision-making.

6.0 References

[1] Ahmed, S. (2025, March 3). Ecommerce Consumer Behavior Analysis Data. Kaggle. https://www.kaggle.com/datasets/salahuddinahmedshuvo/ecommerce-consumer-behavior-analysis-data?resource=download

[2] Rocky. (2026, 14 January). Social Media User Analysis. Kaggle. https://www.kaggle.com/datasets/rockyt07/social-media-user-analysis

[3] Nachimuthu, P., Balakumar, S., Sakthivel, V., Vadivelu, V., & Karthikeyan, P. (2025). The Power of Data Mining: Analyzing social media trends for Business strategy. In Lecture notes in networks and systems (pp. 451–461). https://doi.org/10.1007/978-981-96-5751-3_38

[4] Instagram users, stats, data, trends, and more — DataReportal – Global Digital Insights. (n.d.). DataReportal – Global Digital Insights. https://datareportal.com/essential-instagram-stats

[5] Priansa, D., & Suryawardani, B. (2020). Effects of E-Marketing and Social Media Marketing on E-commerce Shopping Decisions. , 20. https://doi.org/10.25124/jmi.v20i1.2800.