University of Niagara Falls Canada

Master's in Data Analytics
Winter 2025 Predictive Analytics (DAMO-510-3)

Predictive Statistical Problems Assignment Part 2

Professor: Zeeshan Ahmad

Project Group:09

Students:

Mounika Ravella (NF1003648)
Mahvish Kounain (NF1014902)
Mohammed Imran Khan (NF1001923)
Chittanoori Manish(NF1000588)

Title:

Prediction of Mental Health Status by Analyzing Social Media Activities.

1.Statement of Purpose

Introduction: The rise of social media platforms has provided a unique opportunity to gain insights into various aspects of individuals' lives, including their mental health status. By analyzing the vast amounts of data generated by users on these platforms, it is possible to identify patterns and trends that can help predict mental health conditions. This project uses data from social media activities (such as posts, comments, or interactions) to build a predictive model that can assess a person's mental health status. The goal is to identify potential mental health issues early, allowing for timely intervention and support for individuals who may need help.

Motivation: Mental health issues are a growing concern worldwide, with millions of people affected by conditions such as depression, anxiety, and stress. Early detection and intervention are crucial in addressing these issues and improving the quality of life for individuals. Traditional methods of mental health assessment often rely on self-reporting and clinical evaluations, which can be time-consuming and may not always capture the full extent of an individual's condition. Social media data offers an alternative and complementary approach to understanding mental health, as it reflects real-time, naturalistic interactions and behaviours.

Objectives:

1. To collect and preprocess social media data from various platforms (e.g., Twitter, Facebook, Instagram) for analysis.
2. To identify key features and indicators of mental health status based on users' social media activities, such as language patterns, sentiment, engagement levels, and posting frequency.
3. To develop and validate a predictive model for mental health status using machine learning techniques.
4. To evaluate the effectiveness and accuracy of the predictive model in identifying individuals at risk of mental health issues.
5. To explore the ethical implications and privacy concerns associated with using social media data for mental health prediction and ensure that the project adheres to all relevant guidelines and regulations.

Approach: The project will be conducted in several stages:

1. **Data Collection:** Gather social media data from Kaggle and ensure proper anonymization to protect user privacy.

2. **Feature Extraction:** Analyze the collected data to identify relevant features that may indicate mental health status.
3. **Model Development:** Employ machine learning algorithms to train and validate a predictive model based on the extracted features.
4. **Evaluation:** Assess the performance of the model using various metrics, such as accuracy, precision, recall, and F1 score.
5. Ethical Considerations: Address potential ethical issues and develop guidelines for responsible use of social media data in mental health prediction.

2.Scope of the Project:

This project is designed to assess how social media engagement—such as usage patterns, content interactions, and digital footprints—correlates with an individual's mental health status. The dataset comprises social media behavioral metrics, demographic details, and self-reported mental health indicators. The study will focus on:

- **Data Preprocessing:** Handling missing values, standardizing variables, and encoding categorical features.
- **Exploratory Data Analysis (EDA):** Identifying patterns and relationships using statistical and visualization techniques, including correlation analysis.
- **Predictive Modelling:** Developing and evaluating machine learning models to classify and predict mental health outcomes.
- **Model Validation:** Using statistical measures to assess model accuracy, reliability, and applicability.
- **Business Insights and Recommendations:** Providing actionable insights to improve digital well-being policies and mental health interventions.

  All the above mentioned as be briefly explained to get the exact idea about the project.

3.Background Research and Literature:

In this part we focus on atleast 2 references which includes 2 peer-reviewed articles that are supporting our project’s rationale. This can be explained as follows:

The role of social media in shaping mental health has been widely studied, with research indicating both positive and negative impacts. Studies suggest that prolonged exposure to certain types of content, engagement patterns,

and online interactions may contribute to increased stress, anxiety, and depression. Predictive modeling based on digital behaviors has been successfully used in mental health assessment, offering opportunities for early intervention. Relevant research includes:

1. **Keles, B., McCrae, N., & Grealish, A. (2020).** "The Effect of Social Media on Adolescent Mental Health: A Systematic Review." *Journal of Adolescence, 82*, 86-99. This study explores the direct correlation between social media engagement and mental health concerns among young individuals.
2. **Twenge, J. M., & Campbell, W. K. (2018).** "Associations Between Screen Time and Lower Psychological Well-Being Among Children and Adolescents." *Preventive Medicine Reports, 12*, 271-283. This research outlines the psychological effects of excessive screen time and its implications for mental health.

4.Design and Data Collection Methods:

1)Dataset Overview:

The dataset from Kaggle consists of 481 records and 21 features, focusing on demographics, social media usage, and mental health indicators. Below is a detailed breakdown of the dataset's structure and potential insights:

---

a. Demographic Information

These columns provide information about the respondent's background:

- **Age:** The age of the respondent.
- **Gender:** Categorical feature (Male, Female, Other).
- **Relationship Status:** Whether the person is Single, In a relationship, Married, etc.
- **Occupation Status:** Indicates whether the respondent is a Student, Employed, or Unemployed.
- **Affiliated Organizations:** The institution type the respondent is associated with (e.g., University, Workplace).

Potential Insight:

- Helps understand whether mental health status varies by age group, gender, or occupation.

---

b. Social Media Usage Patterns

These features capture how respondents use social media:

- **Use of Social Media:** Binary (Yes/No).

- **Common Platforms:** Lists the most frequently used platforms (Facebook, Twitter, Instagram, etc.).
- **Average Time Spent on Social Media Daily:** Ordinal feature ranging from “Less than 1 hour” to “More than 5 hours.”
- **Usage Patterns:** Includes behaviors such as using social media without purpose, distraction levels, and dependency.

Potential Insight:

- Can help analyze if excessive time on social media correlates with mental health issues like anxiety, stress, or depression.

---

c. Mental Health Indicators

These columns assess emotional well-being and mental health symptoms:

- **Restlessness without Social Media:** Measures withdrawal symptoms from social media.
- **Distraction Levels:** Self-reported score on a scale of 1-5.
- **Worries and Anxiety:** How much respondents are bothered by worries.
- **Difficulty in Concentration:** Self-reported levels of concentration difficulty.
- **Social Media Comparisons:** How often respondents compare themselves to successful individuals online.
- **Seeking Validation:** Measures dependency on likes, comments, and approval from others.
- **Depression Indicators:** Frequency of feeling down or depressed.
- **Fluctuations in Interest for Daily Activities:** Captures loss of interest in daily life.
- **Sleep Issues:** Reports on facing difficulty in sleep.

Potential Insight:

- These features help build a predictive model for mental health status based on social media behavior.
- Certain features like comparison on social media, sleep issues, and distraction may be strong predictors of depression or anxiety.

2) Data Preprocessing and Handling Missing Values

2.1 Checking for Missing Values

To ensure data integrity, we examined the dataset for missing values using the following Python code:

Checking for missing values:

print([f"NaN values in: {(element, value)}" for (element, value) in zip(df.isnull().sum().index, df.isnull().sum()) if value > 0])

This analysis revealed that there were 30 missing values in the dataset.

2.2 Understanding the Nature of Missing Data

To determine the type of missing data, we conducted a Missing Completely at Random (MCAR) test using a Chi-Square test. The following code was used:

 # Understanding the nature of missing data

Create a binary column: 1 if missing, 0 otherwise

df['missing'] = df['5. What type of organizations are you affiliated with?'].isnull().astype(int)

Run Chi-Square Test

cross_tab = pd.crosstab(df['5. What type of organizations are you affiliated with?'], df['missing'])

chi2, p, dof, expected = stats.chi2_contingency(cross_tab)

print(f"Chi-Square Statistic: {chi2}, P-value: {p}")

The test results indicated that the missing data was MCAR (Missing Completely at Random), meaning the missing values do not depend on other observed or unobserved data.

2.3 Handling Missing Values

Since the missing data was MCAR, the best approach was to replace the missing values with the mode of the respective column. The following code was used to handle missing values:

```
# MCAR, so replacing with mode
df["5. What type of organizations are you affiliated with?"] = df["5. What type of organizations are you affiliated with?"].fillna(df["5. What type of organizations are you affiliated with?"].mode()[0])
```

This ensures data consistency and prevents data loss.

Checking for Outliers

Detecting Outliers

To identify potential outliers in the dataset, a boxplot was used:

```
# Checking for outliers
# Boxplot to visualize outliers
plt.figure(figsize=(8, 5))
sns.boxplot(data=df, palette="coolwarm")

# Customize the plot
plt.title("Box Plot for Outlier Detection", fontsize=14)
plt.xlabel("Variables", fontsize=12)
plt.ylabel("Values", fontsize=12)
plt.show()
```

This visualization revealed that the Age variable contained outliers.

2.4 Handling Outliers

Outliers in the dataset were observed in the Age variable, ranging from 13 to 91 years. Typically, outliers can indicate data errors or extreme values that may distort the analysis. However, in this case, all detected outliers were within a reasonable range of human ages and were not considered anomalies. Since the dataset focuses on analyzing mental health across different age groups, removing these outliers would result in loss of valuable information about older and younger individuals.

Additionally, a few negligible outliers were observed in other fields, but all remained within an acceptable range. As these values did not significantly impact the analysis, the dataset was kept as it is to preserve data integrity and inclusivity. No modifications were made to these values.

2.5 Correlation Analysis: Understanding Variable Relationships and Feature Selection:

We run correlation analysis to understand the relationship between different variables in the dataset. Here's why it's important:

1. **Identifying Relationships –** Correlation helps determine if two variables move together (positive correlation) or in opposite directions (negative correlation). For example, in your mental health prediction project, you might check if age is correlated with stress levels or if social media activity correlates with anxiety scores.
2. **Feature Selection –** Highly correlated independent variables (multicollinearity) can lead to redundancy in predictive models. Correlation analysis helps in selecting the most relevant features and removing unnecessary ones.
3. **Understanding Data Patterns –** It helps uncover trends in the dataset. For instance, if time spent on social media has a strong correlation with depression scores, this insight can guide further analysis**.**
4. **Model Improvement –** Some machine learning models, like linear regression, assume minimal multicollinearity. Identifying strong correlations early can help in modifying the dataset for better model performance.
5. **Detecting Unexpected Insights –** Sometimes, correlation analysis reveals surprising relationships that were not initially considered, helping refine hypotheses and research direction.

2.6 Correlation Matrix & Feature Selection

a. Overview

Here, correlation matrix examines how various questionnaire items—focusing on social media usage, social comparison, and mental health—relate to each other. The heatmap ranges from negative correlations (blue) through near-zero (white) to

positive correlations (red). Diagonal values are 1.0, reflecting perfect correlation of each variable with itself.

---

b. Key Correlation Clusters

1. **Compulsive Social Media Usage (Q2–Q7)**
    - Items such as using social media without purpose (Q2), getting distracted (Q3), feeling restless without it (Q4), difficulty reducing usage (Q5), easy distractibility (Q6), and difficulty concentrating (Q7) all show moderate-to-strong positive correlations with each other.
    - Indicates a common factor of problematic or excessive social media use.
2. **Social Comparison (Q8, Q9)**
    - Comparing oneself to others on social media (Q8) and the emotional impact of that comparison (Q9) cluster together.
    - These correlate moderately with mental health indicators, suggesting that negative feelings from comparisons may predict lower mood or well-being.
3. **Mental Health Indicators (Q11–Q13)**
    - Depression/loneliness (Q11), fluctuating interest in daily activities (Q12), and sleep issues (Q13) correlate with one another, forming a mental health cluster.
    - These also show moderate correlations with compulsive usage and social comparison, suggesting overlap between heavy social media engagement and poorer mental health.

---

c. Notable Cross-Correlations

- **Social Media Usage & Mental Health**: Higher compulsive usage (Q2–Q7) often pairs with greater depression, lower interest in daily life, and more sleep problems.
- **Social Comparison & Mental Health**: Those who feel worse after comparing themselves (Q9) often report more depressive or sleep-related issues.

---

d. Best Features for Predicting Mental Health

- **Target Variables**:
    - Depression/Loneliness (Q11)
    - Interest in Activities (Q12)

- Sleep Issues (Q13)

- **Predictors**:
  - **Compulsive Usage**: Combine or individually use Q2–Q7.
  - **Social Comparison**: Q8 (comparison level) and Q9 (emotional impact).
  - **Optional**: Q10 (multitasking) if it shows meaningful correlations.
- **Age (Q1)**: Shows weak correlations; can be included as a control if desired.

5.Methodology/Strategies:

In this step, we are explaining about the plan which we had implemented for our project proposal, which includes statistical tests and model development in includes visualization.

2.1 Data Preprocessing and Feature Engineering

The preprocessing phase involved transforming raw data into more usable forms and creating new composite features to capture social media usage patterns and their psychological impacts. This section outlines the key steps:

a.Categorical to Numerical Transformation (Social Media Usage Time): The survey asked participants about their average daily social media usage in categorical terms. To convert this into a usable format, the data was mapped to numerical values based on the time intervals:

1. :
   - "Less than an Hour" → 0.5
   - "Between 1 and 2 hours" → 1.5
   - "Between 2 and 3 hours" → 2.5
   - "Between 3 and 4 hours" → 3.5
   - "Between 4 and 5 hours" → 4.5
   - "More than 5 hours" → 6

   The following code converts the categorical responses to numerical values:

   column_name = "8. What is the average time you spend on social media every day?"

   # Strip spaces and standardize case

```
df[column_name] = df[column_name].str.strip()


# Define mapping for ordinal categories to numerical values
time_mapping = {
    "Less than an Hour": 0.5,
    "Between 1 and 2 hours": 1.5,
    "Between 2 and 3 hours": 2.5,
    "Between 3 and 4 hours": 3.5,
    "Between 4 and 5 hours": 4.5,
    "More than 5 hours": 6
}


# Convert the column
df[column_name] = df[column_name].map(time_mapping)


# Print the transformed column
print(df[[column_name]].head())
```

b.One-Hot Encoding of Social Media Platforms: To handle the data about the platforms participants use, the responses were converted into binary variables through one-hot encoding. Each platform had its own column, where a "1" indicated usage and "0" indicated non-usage. A new "Platform Score" column was created by summing the binary values across the platform columns.

```
platforms = ["Facebook", "Twitter", "Instagram", "YouTube", "Snapchat", "Discord", "Reddit", "Pinterest", "TikTok"]


# Create binary columns for each platform
for platform in platforms:
    df[platform] = df[platforms_column].apply(lambda x: 1 if isinstance(x, str) and platform in x else 0)


# Create a "Platform Score" by summing all platform columns
df["Platform Score"] = df[platforms].sum(axis=1)
```

```
# Print the first few rows of the updated dataset
print(df[platforms + ["Platform Score"]].head())
```

c.Handling Continuous Data (Age): The continuous age data was converted into categorical bins (age groups) to better analyze social media usage by different age segments.

Python code:

```
age_column = "1. What is your age?"
age_bins = [10, 18, 25, 35, 50, 65, 100]
age_labels = ["10-17", "18-24", "25-34", "35-49", "50-64", "65+"]
df["Age Group"] = pd.cut(df[age_column], bins=age_bins, labels=age_labels)


# Print the transformed age group data
print(df[["Age Group"]].head())
```

2.2 Feature Engineering

Feature engineering was used to combine multiple columns into two composite features representing social media usage and its psychological impact:

a.Social Media Score: The "social_media" score was created by summing several questionnaire responses related to purposeless social media use, daily usage time, frequency of self-comparison with successful individuals, and distraction while engaged in other activities.

```
social_media_columns = [
    "9. How often do you find yourself using Social media without a specific purpose?",
    "8. What is the average time you spend on social media every day?",
    "15. On a scale of 1-5, how often do you compare yourself to other successful people through the use of social media?",
    "10. How often do you get distracted by Social media when you are busy doing something?"
]
df["social_media"] = df[social_media_columns].sum(axis=1)


# Display the first few rows to verify
```

```
print(df[["social_media"]].head())
```

b.Impact Score: The "impact" score was created by summing responses related to the psychological and behavioral consequences of social media use, such as distractibility, worry, sleep issues, emotional responses to comparisons, and feelings of restlessness.

```
impact_columns = [
    "12. On a scale of 1 to 5, how easily distracted are you?",
    "13. On a scale of 1 to 5, how much are you bothered by worries?",
    "14. Do you find it difficult to concentrate on things?",
    "18. How often do you feel depressed or down?",
    "16. Following the previous question, how do you feel about these comparisons, generally speaking?",
    "19. On a scale of 1 to 5, how frequently does your interest in daily activities fluctuate?",
    "20. On a scale of 1 to 5, how often do you face issues regarding sleep?",
    "17. How often do you look to seek validation from features of social media?",
    "11. Do you feel restless if you haven't used Social media in a while?"
]
df["impact"] = df[impact_columns].sum(axis=1)


# Display the first few rows to verify
print(df[["social_media", "impact"]].head())
```

c.Data Visualization

The next phase of the analysis involved visualizing the relationship between age groups and social media usage patterns across various platforms. Several types of visualizations were employed to uncover trends and provide intuitive insights into the data:

a.Heatmap (Platform Usage by Age Group): The heatmap visualizes platform usage across age groups, with color intensity indicating the number of users for each platform. It reveals the relative popularity of each platform in different age segments, providing insights into which platforms are preferred by different generations.

b.Grouped Bar Chart (Platform Usage by Age Group): This bar chart compares the number of users for each platform across different age groups. It allows for direct

comparison of how various age groups engage with social media platforms.

c.Line Chart (Platform Trends Across Age Groups): The line chart tracks the usage trends for each platform across different age groups. It helps identify whether certain platforms are gaining or losing popularity as users age.

d.Stacked Bar Chart (Cumulative Platform Usage by Age Group): The stacked bar chart provides a cumulative view of platform usage by age group, allowing a clear understanding of how each platform contributes to the overall usage within each group.

3)Comprehensive Linear Regression Model Analysis

1. Model Performance Metrics

These metrics provide insight into how well the linear regression model fits the data.

1.1 Mean Squared Error (MSE)

MSE: 0.4621

MSE represents the average squared difference between actual and predicted values.

A lower MSE indicates better model performance.

Here, 0.4621 is relatively low, suggesting that the model does not have large errors in predictions.

1.2 R^2 Score (Coefficient of Determination)

R^2 Score: 0.5559 (55.59%)

It measures how well the independent variable(s) explain the variance in the dependent variable.

A value of 0.5559 means that 55.59% of the variability in the dependent variable is explained by the model.

The remaining 44.41% is due to other factors not included in the model.

1.3 Adjusted R^2 Score

Adjusted R^2 Score: 0.5512

Adjusted R^2 penalizes the addition of unnecessary predictors that do not improve the model significantly.

Since Adjusted R^2 is slightly lower than R^2, it suggests that the model does not suffer from excessive overfitting.

1.4 Intercept and Coefficient (Slope)

Intercept: -0.0118

This represents the expected value of the dependent variable when all independent variables are zero.

Coefficient (Slope): 0.6777

This means that for every unit increase in the independent variable, the dependent variable increases by 0.6777 units on average.

2. Checking Assumptions of Linear Regression

For a linear regression model to be valid and reliable, certain assumptions must hold. We analyze them below:

2.1 Independence of Residuals (Durbin-Watson Test)

Durbin-Watson Statistic: 2.0772

This test detects autocorrelation in residuals (errors).

The ideal value is close to 2:

If it's close to 0 → Strong positive autocorrelation (residuals are correlated, bad for the model).

If it's close to 4 → Strong negative autocorrelation.

If it's near 2 → No significant autocorrelation, assumption is satisfied.

Since 2.0772 is very close to 2, we do not have autocorrelation issues.

Assumption of independent residuals is satisfied.

2.2 Homoscedasticity (Constant Variance of Errors)

The Residual Plot (first image) helps in checking homoscedasticity.

In the plot, residuals are randomly scattered around zero without forming any pattern.

No visible "funnel shape" or increasing spread suggests no heteroscedasticity (unequal variance).

Assumption of homoscedasticity is satisfied.

2.3 Normality of Residuals

Histogram of Residuals (both images):

The histogram shows a roughly bell-shaped distribution.

There is some slight skewness, but it is not severe.

Q-Q Plot (second image):

This plot compares the distribution of residuals against a normal distribution.

If residuals lie close to the red diagonal line, it confirms normality.

The plot shows most points aligned along the red line, supporting the normality assumption.

Assumption of normal residuals is reasonably satisfied.

3. Conclusion & Model Interpretation

Based on the above analysis:

The model is statistically valid for prediction as it satisfies key assumptions.

55.59% of the variance in the dependent variable is explained by the model.

The independent variable positively impacts the dependent variable (coefficient = 0.6777).

There is no serious violation of assumptions, meaning the model is well-fitted.

Here’s a detailed analysis of the OLS regression model summary:

Key Insights from the Model:

Model Fit (R-squared and Adj. R-squared):

R-squared (0.467) shows that approximately 46.7% of the variance in the dependent variable ("impact") is explained by the independent variable ("social_media").

Adjusted R-squared (0.465) accounts for model complexity and confirms a good fit without overfitting.

Interpretation: This is a moderately strong relationship, suggesting "social_media" is a significant factor influencing "impact."

Significance of Variables (P-value and T-statistics):

The P-value for "social_media" (0.000) indicates it's highly statistically significant (less than 0.05).

T-statistic for "social_media" (18.286) shows a strong relationship. The larger the T-statistic, the stronger the evidence against the null hypothesis.

Interpretation: "social_media" significantly affects "impact," and this relationship is unlikely due to chance.

Coefficients (Effect Size):

The coefficient for "social_media" (0.6777) suggests that for each unit increase in "social_media," the "impact" increases by approximately 0.678 units.

The constant (-0.0118) is not statistically significant (P = 0.751), indicating that the baseline impact, when "social_media" is zero, isn’t a reliable prediction.

Interpretation: Positive and significant effect of "social_media" on "impact."

F-statistic and Prob (F-statistic):

F-statistic (334.4) with a P-value of 4.14e-54 shows the overall model is statistically significant.

Interpretation: The regression model is a good fit and not driven by random noise.

Diagnostic Statistics (Durbin-Watson, Skewness, and Kurtosis):

Durbin-Watson (1.975) indicates no significant autocorrelation.

Skewness (-0.011) and Kurtosis (2.707) suggest the residuals are approximately normally distributed.

One-Variate Summary Statistics

One-variate summary statistics provide an overview of key variables and their distribution. These statistics help in understanding data trends, detecting outliers, and assessing central tendencies.

Summary Statistics:

Summary Statistics:

1. What is your age?

count	481.0000
Mean	26.13659
std	9.91511
min	13.00000
25%	21.00000
50%	22.00000
75%	26.00000
max	91.00000

8. What is the average time you spend on social media every day?

Count	481.00000
Mean	3.529106
Std	1.755107
Min	0.500000
25%	2.500000
50%	3.500000
75%	4.500000
max	6.00000

A histogram was plotted to visualize the distribution of the impact score, which helps in identifying the frequency distribution of mental health impacts across users.

Code:

```
import seaborn as sns

import matplotlib.pyplot as plt


sns.histplot(df["impact"], bins=20, kde=True)

plt.title("Distribution of Impact Score")

plt.xlabel("Impact Score")

plt.ylabel("Frequency")

plt.show()
```

6.Business Impact and Conclusion

Business Impact

The analysis of social media usage and its psychological effects offers key business benefits, including:

1. **Targeted Marketing:** By understanding platform popularity among different age groups, businesses can refine marketing efforts and tailor campaigns to resonate with users' psychological needs.
2. **Product Development:** Insights into social media behaviors can guide the development of new features or apps that promote healthier usage or enhance engagement.
3. **Mental Health Initiatives:** Social media platforms can introduce features supporting mental well-being, while mental health companies can create products addressing social media-induced stress.
4. **User Engagement:** Data on platform trends and behaviors can help improve user interaction, retention, and satisfaction by optimizing features that reduce negative behaviors.

Recommendations

1. **Platform-Specific Strategies:** Focus on high-engagement platforms for targeted marketing.
2. **Promote Mental Well-Being:** Design features that encourage healthy social media use and reduce negative behaviors.
3. **Personalized User Experiences:** Use social media impact scores to offer tailored content.
4. **Corporate Social Responsibility (CSR):** Launch initiatives to support mental health and manage social media impact.
5. **Collaborate with Mental Health Organizations:** Develop tools to help users manage social media use and its psychological effects.

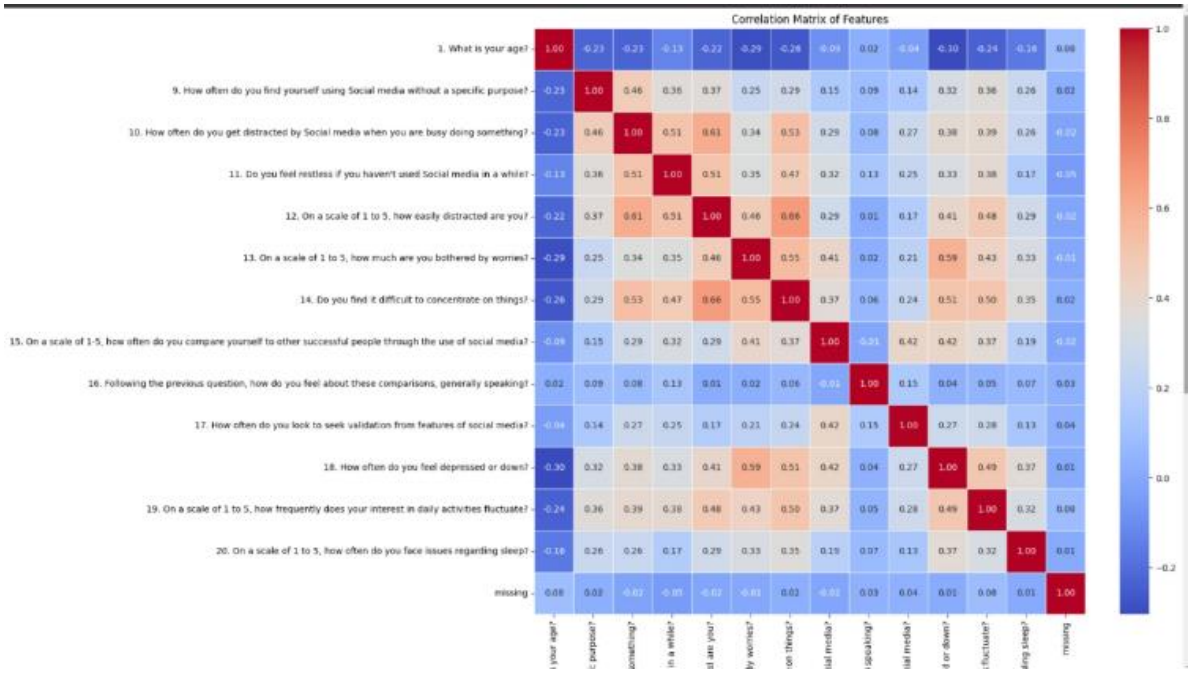
Conclusion:
The analysis of social media behaviors provides a new dimension for assessing mental health trends. With the help of machine learning models, organizations can gain valuable insights into user behaviors and implement proactive interventions. By responsibly leveraging data analytics, businesses can enhance user engagement, develop wellness-driven policies, and contribute to better mental health outcomes in digital spaces.

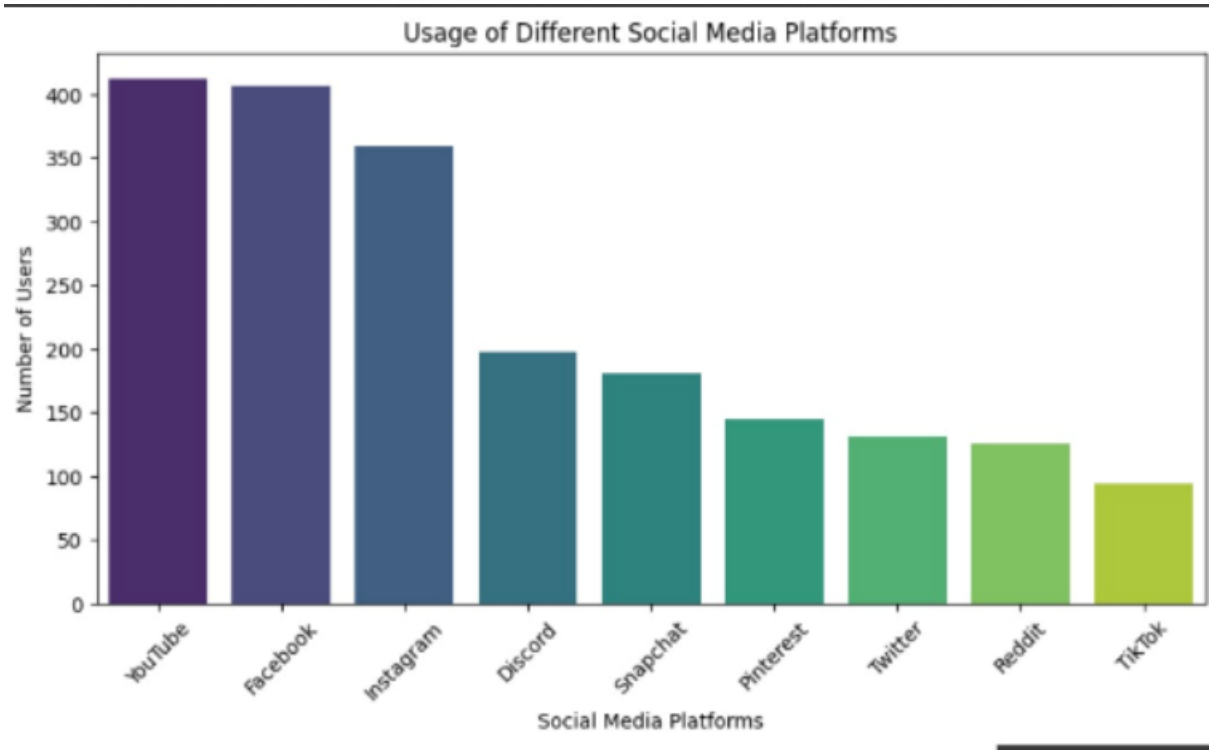
References:

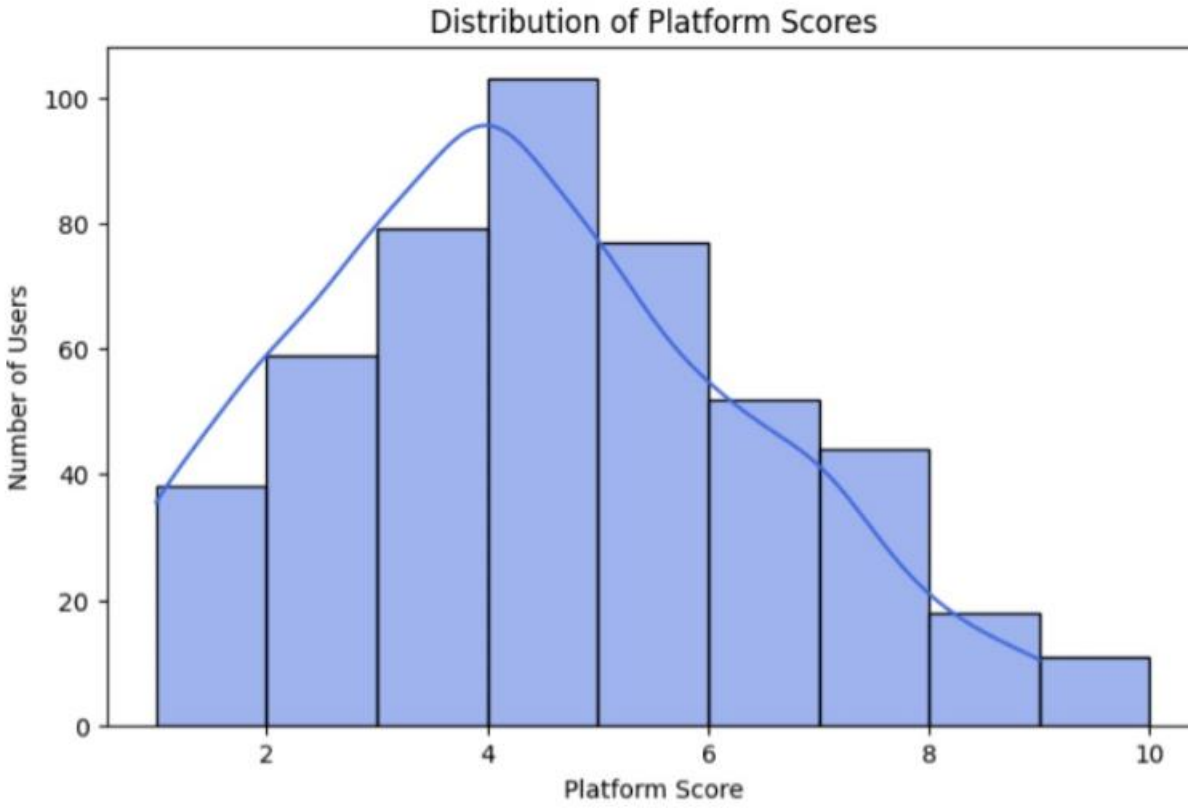
To properly reference the given Kaggle dataset, you can use these related references that help support your work:

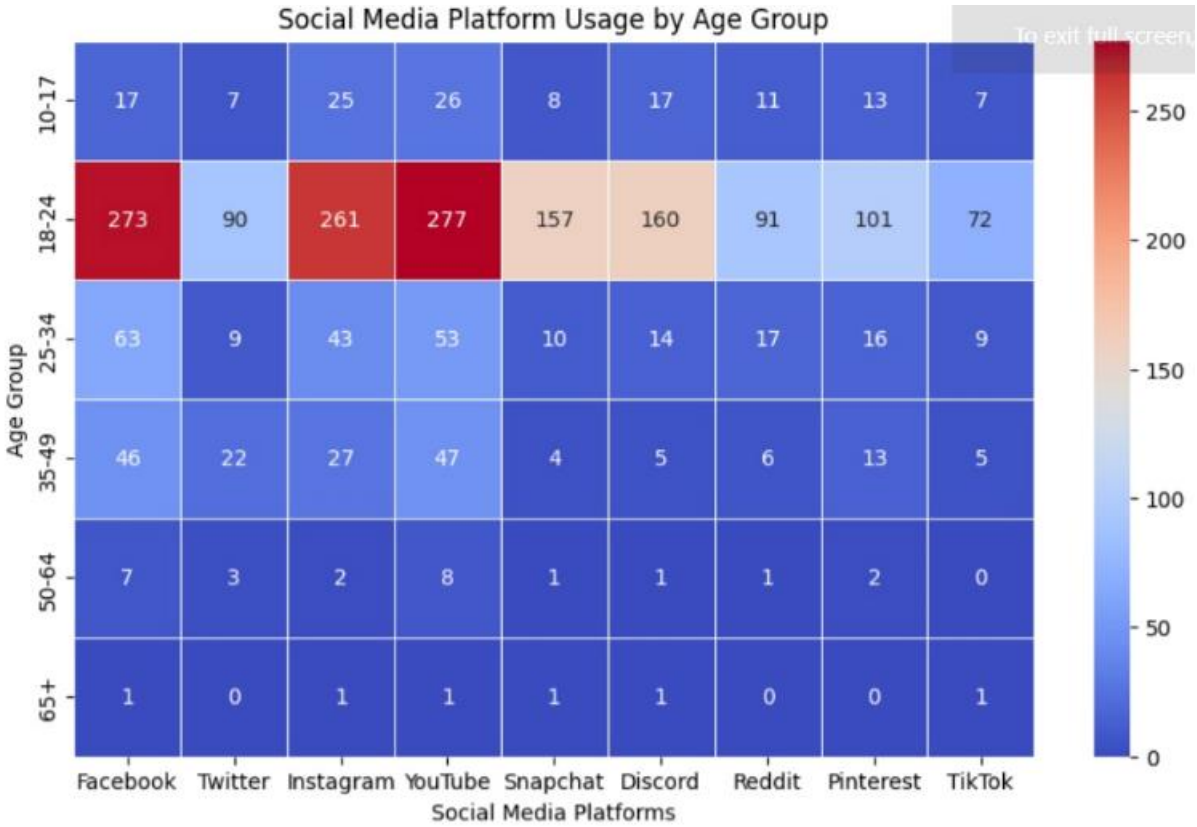
1. **Souvik Ahmed.** (2021). *Social Media and Mental Health*. Kaggle Dataset. Available: https://www.kaggle.com/datasets/souvikahmed071/social-media-and-mental-health
2. **Pew Research Center.** (2021). *Social Media Use in 2021*. Available: https://www.pewresearch.org/fact-tank/2021/04/07/social-media-use-in-2021/
3. **Smith, A., & Duggan, M.** (2013). *Online Dating & Relationships*. Pew Research Center. Available: https://www.pewresearch.org/fact-tank/2013/02/14/online-dating-relationships/
4. **Kuss, D. J., & Griffiths, M. D.** (2017). Social networking sites and addiction: Ten lessons learned. *International Journal of Environmental Research and Public Health*, 14(3), 311. DOI: 10.3390/ijerph14030311
5. **Twenge, J. M., Joiner, T. E., Rogers, M. L., & Martin, G. N.** (2018). Increases in depressive symptoms, suicide-related outcomes, and suicide rates among US adolescents after 2010 and links to increased smartphone use. *Clinical Psychological Science*, 6(1), 3-17. DOI: 10.1177/2167702617723376

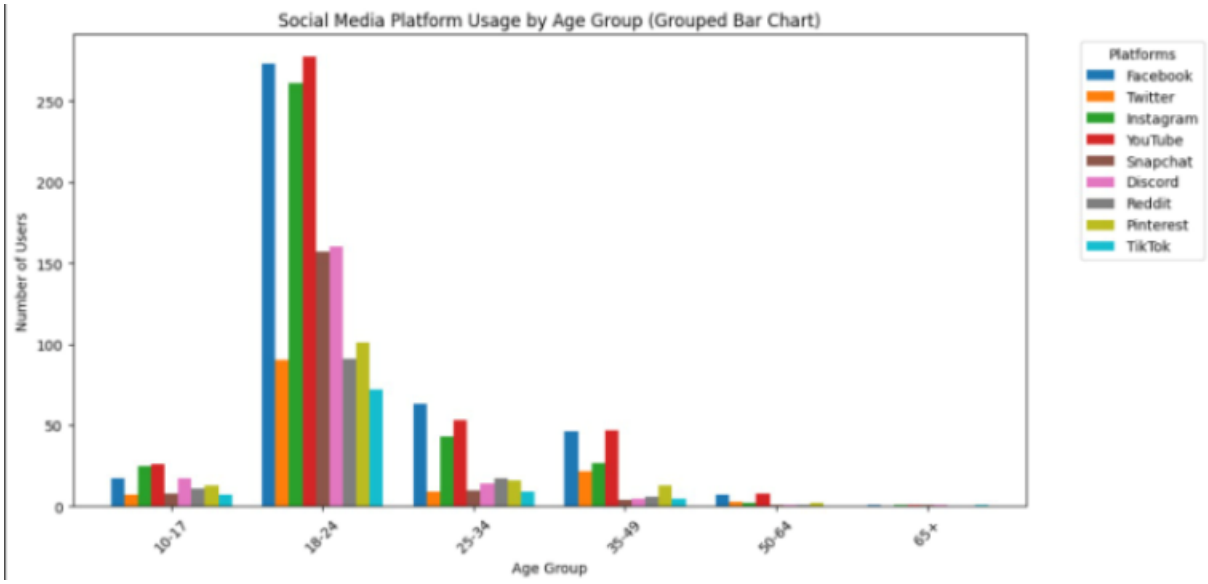
APPENDIX:

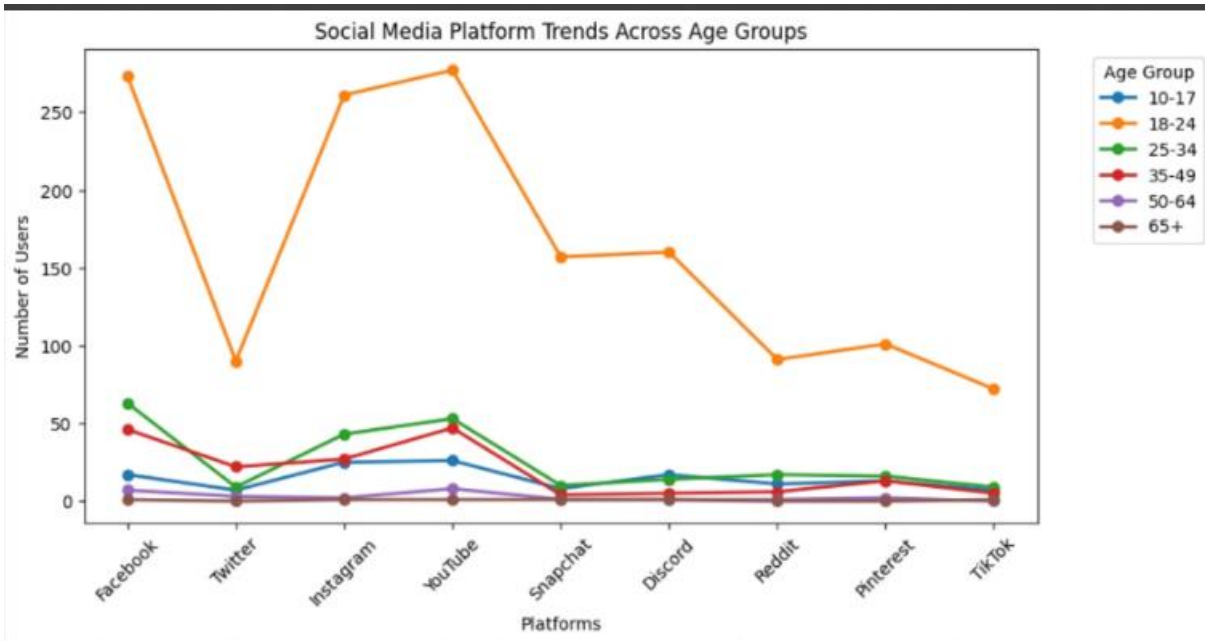

Correlation Matrix of Features
1. What is your age?
9. How often do you find yourself using Social media without a specific purpose?
10. How often do you get distracted by Social media when you are busy doing something?
11. Do you feel restless if you haven't used Social media in a while?
12. On a scale of 1 to 5, how easily distracted are you?
13. On a scale of 1 to 5, how much are you bothered by worries?
14. Do you find it difficult to concentrate on things?
15. On a scale of 1-5, how often do you compare yourself to other successful people through the use of social media?
16. Following the previous question, how do you feel about these comparisons, generally speaking?
17. How often do you look to seek validation from features of social media?
18. How often do you feel depressed or down?
19. On a scale of 1 to 5, how frequently does your interest in daily activities fluctuate?
20. On a scale of 1 to 5, how often do you face issues regarding sleep?
missing

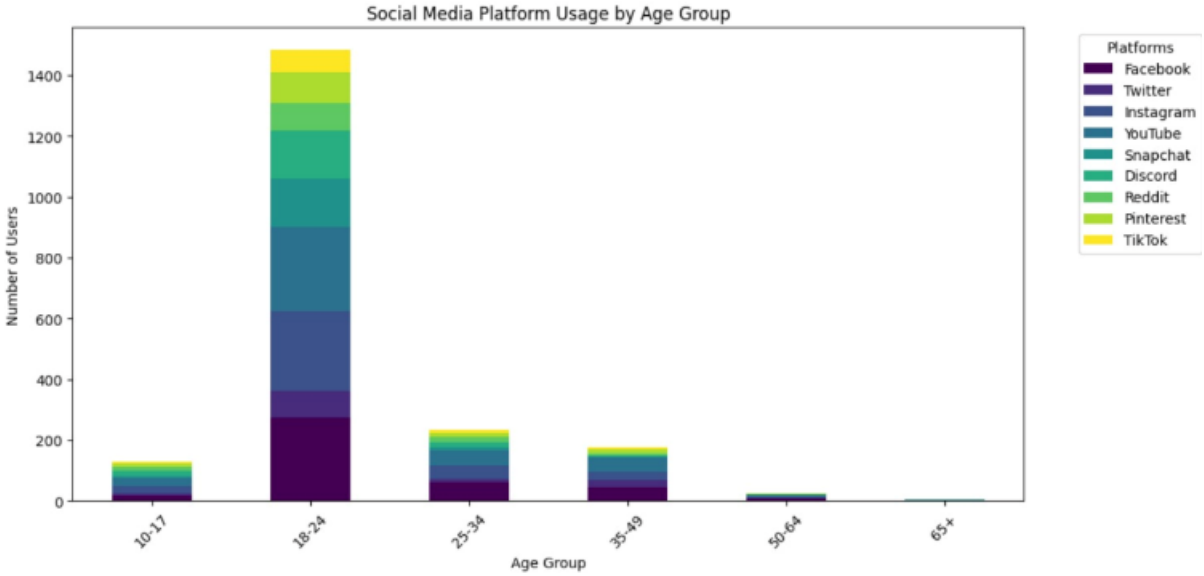

Usage of Different Social Media Platforms
Number of Users
0
50
100
150
200
250
300
350
400
YouTube
Facebook
Instagram
Discord
Snapchat
Pinterest
Twitter
Reddit
TikTok
Social Media Platforms

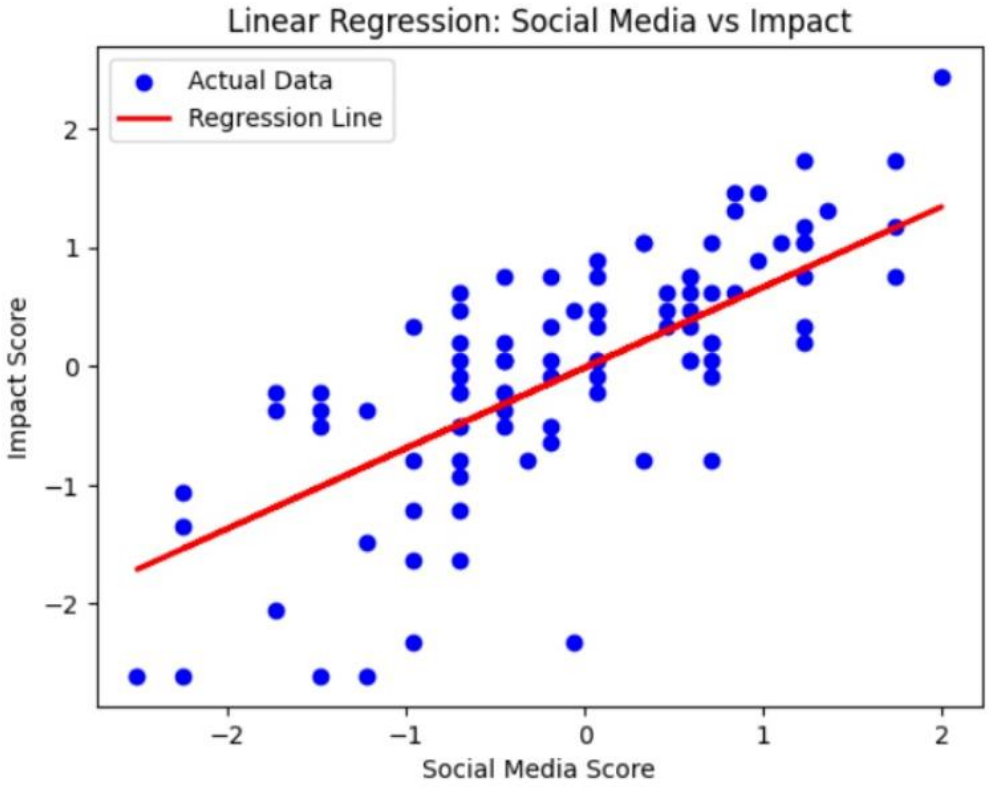

Distribution of Platform Scores
Number of Users
Platform Score
0
20
40
60
80
100
2
4
6
8
10

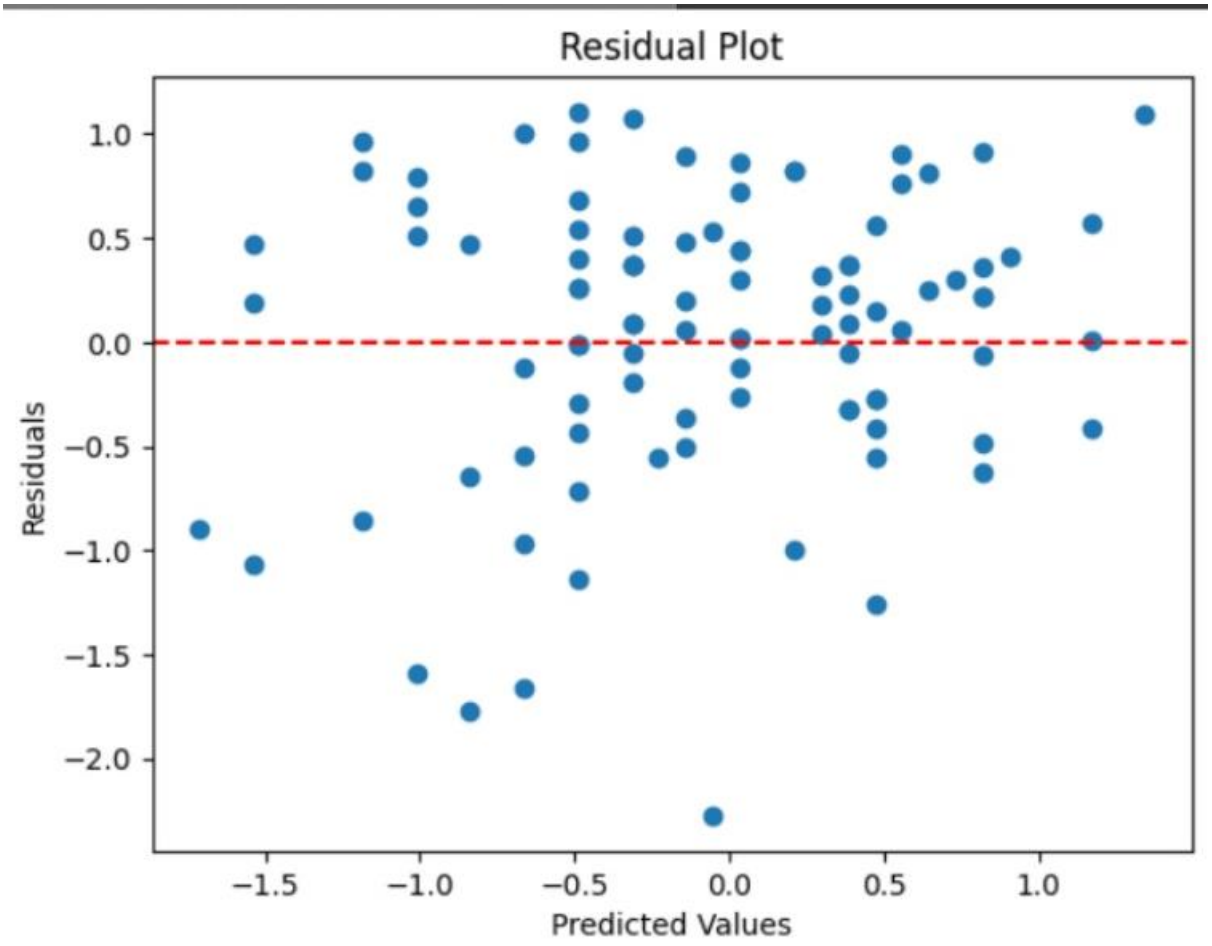

Social Media Platform Usage by Age Group
Age Group
10-17
18-24
25-34
35-49
50-64
65+
17 7 25 26 8 17 11 13 7
273 90 261 277 157 160 91 101 72
63 9 43 53 10 14 17 16 9
46 22 27 47 4 5 6 13 5
7 3 2 8 1 1 1 2 0
1 0 1 1 1 1 0 0 1
Facebook
Twitter
Instagram
YouTube
Snapchat
Discord
Reddit
Pinterest
TikTok
Social Media Platforms
250
200
150
100
50
0

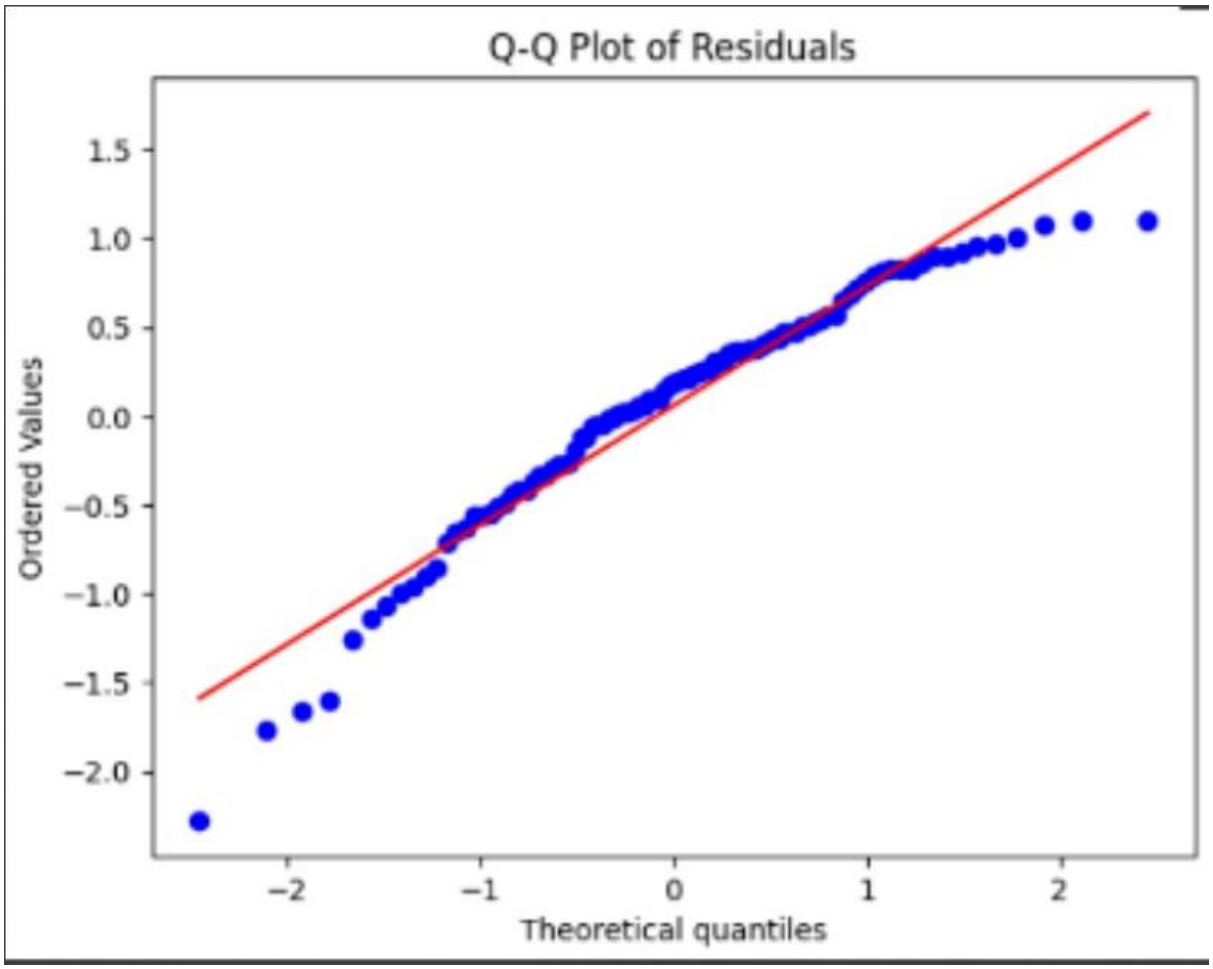

Social Media Platform Usage by Age Group (Grouped Bar Chart)
Number of Users
0
50
100
150
200
250
10-17
18-24
25-34
35-49
50-64
65+
Age Group
Platforms
Facebook
Twitter
Instagram
YouTube
Snapchat
Discord
Reddit
Pinterest
TikTok

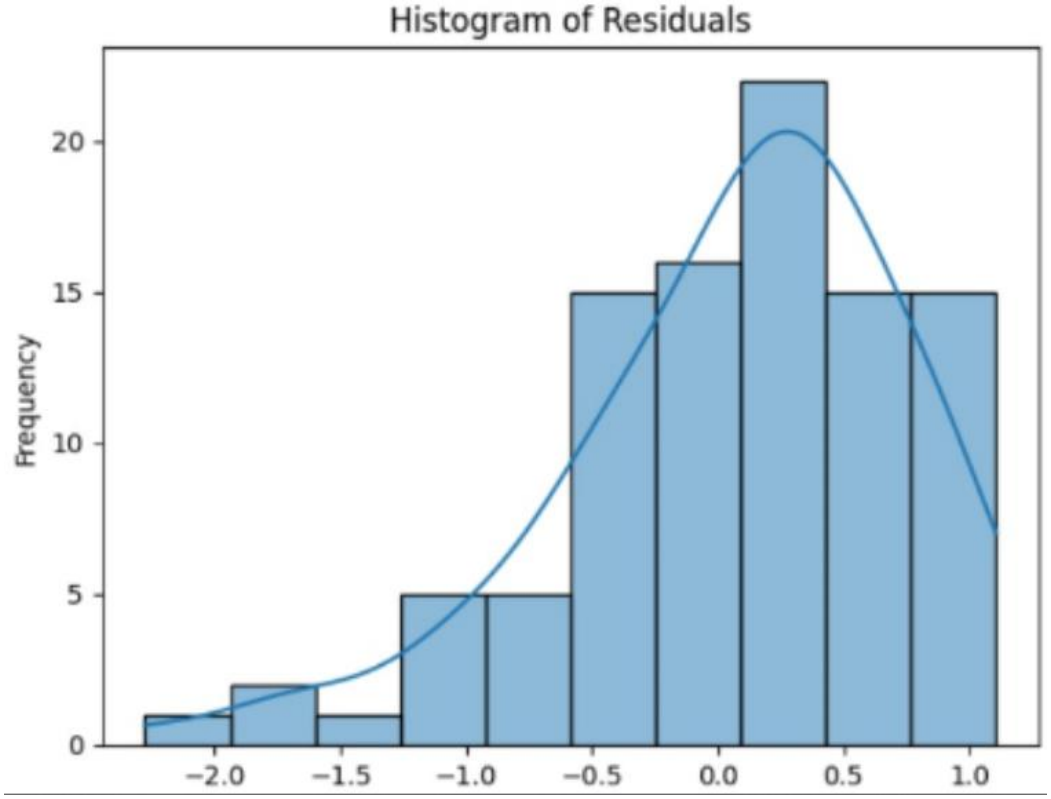

Social Media Platform Trends Across Age Groups
Number of Users
0
50
100
150
200
250
Facebook
Twitter
Instagram
YouTube
Snapchat
Discord
Reddit
Pinterest
TikTok
Platforms
Age Group
10-17
18-24
25-34
35-49
50-64
65+


Social Media Platform Usage by Age Group
Number of Users
Age Group
10-17
18-24
25-34
35-49
50-64
65+
Platforms
Facebook
Twitter
Instagram
YouTube
Snapchat
Discord
Reddit
Pinterest
TikTok


Linear Regression: Social Media vs Impact
Actual Data
Regression Line
Impact Score
Social Media Score


Residual Plot
Residuals
Predicted Values
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0


Q-Q Plot of Residuals
Ordered Values
Theoretical quantiles
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2
−1
0
1
2


Histogram of Residuals
Frequency
20
15
10
5
0
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0