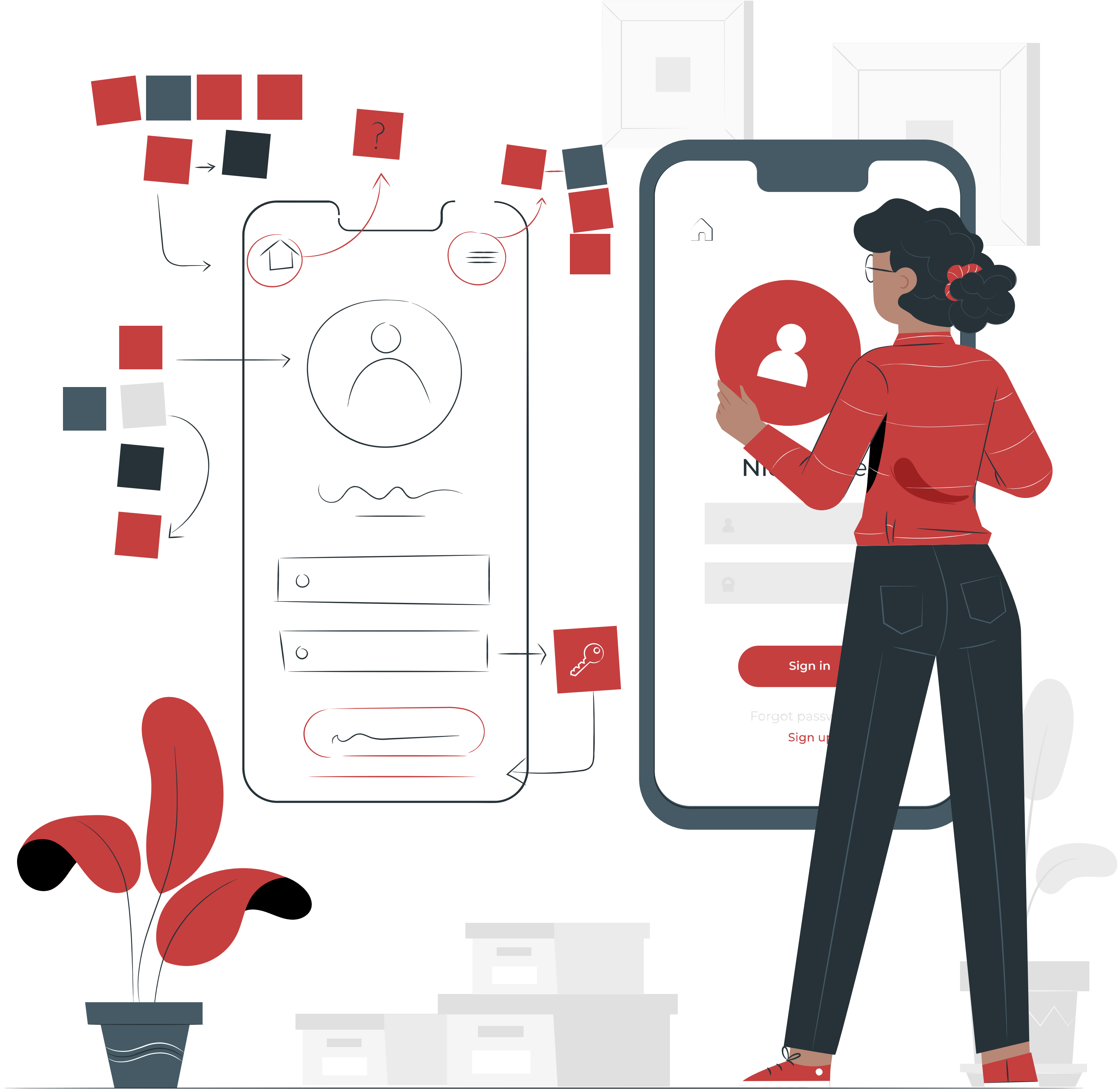
Social media effects on Mental Health

by Group 1 \

- Raghad fares
- Refal Alammari
- Alhanouf AldakelAllah

- Doaa Aldobai
- Razan Alkhaluqi
- Sereen Al-hmoud

Supervised by Dr. Khulood Alyahya



Content

1-overview

2-Algorithms Selected and Rationale

3-Model Evaluation Metrics

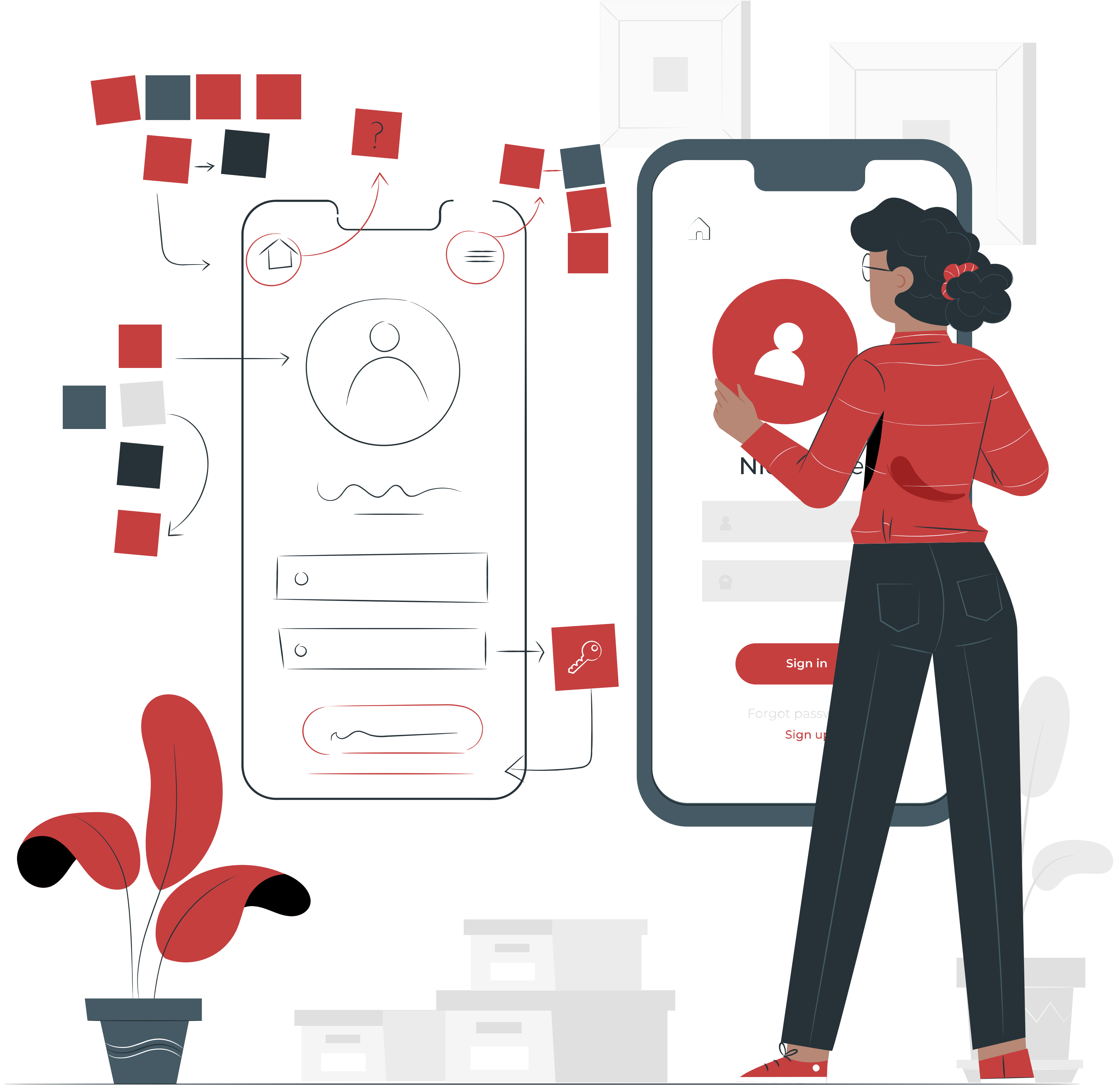
4-Technical Hurdles and Solutions

5-Modeling Efforts and Results

6-Findings on Large Language Models (LLMs)

7-Recommendations and Future Directions

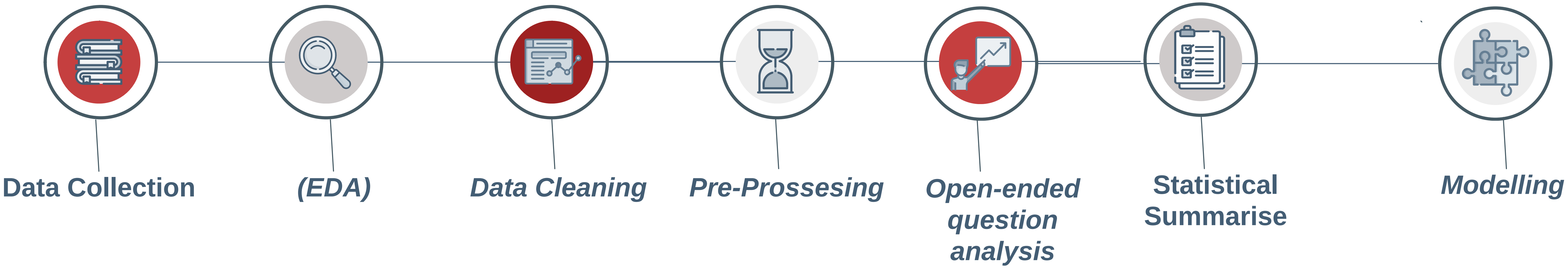
8-Conclusion



01 Overview of the Workflow

Overview of the Workflow

Steps Taken:



Data Collection

(EDA)

Data Cleaning

Pre-Prossesing

Open-ended question analysis

Statistical Summarise

Modelling

Overview of the Workflow: Data Collection

Primary Dataset:

1. The survey was adopted as a primary source, and its questions were developed with the help of a specialist to ensure that the responses align with our goal of analyzing this data
2. Time Frame: 5 September to 15 September 2024
3. Participants: 851 Saudis using social media.
4. Data Collection: Self-administered questionnaire.
5. Distributed via Twitter, Instagram, WhatsApp, and Telegram.
6. Sections: Demographics and 13 items on social media’s mental impact.
7. Included structured (quantitative) and(qualitative) questions and unstructured open-ended (qualitative) questions.

Overview of the Workflow: Data Collection

Secondary Dataset:

1. The Kaggle dataset investigated the potential correlation between the amount of time an individual spends on social media and its impact on their mental health.
2. Date of collection:4/18/2022
3. Data Collection: survey
4. Sections: Demographics and 14 items on social media's mental impact.
5. Included structured (quantitative) and (qualitative) questions.

Goal: Compare findings from the secondary dataset with the primary dataset to identify any differing results and contributing factors.

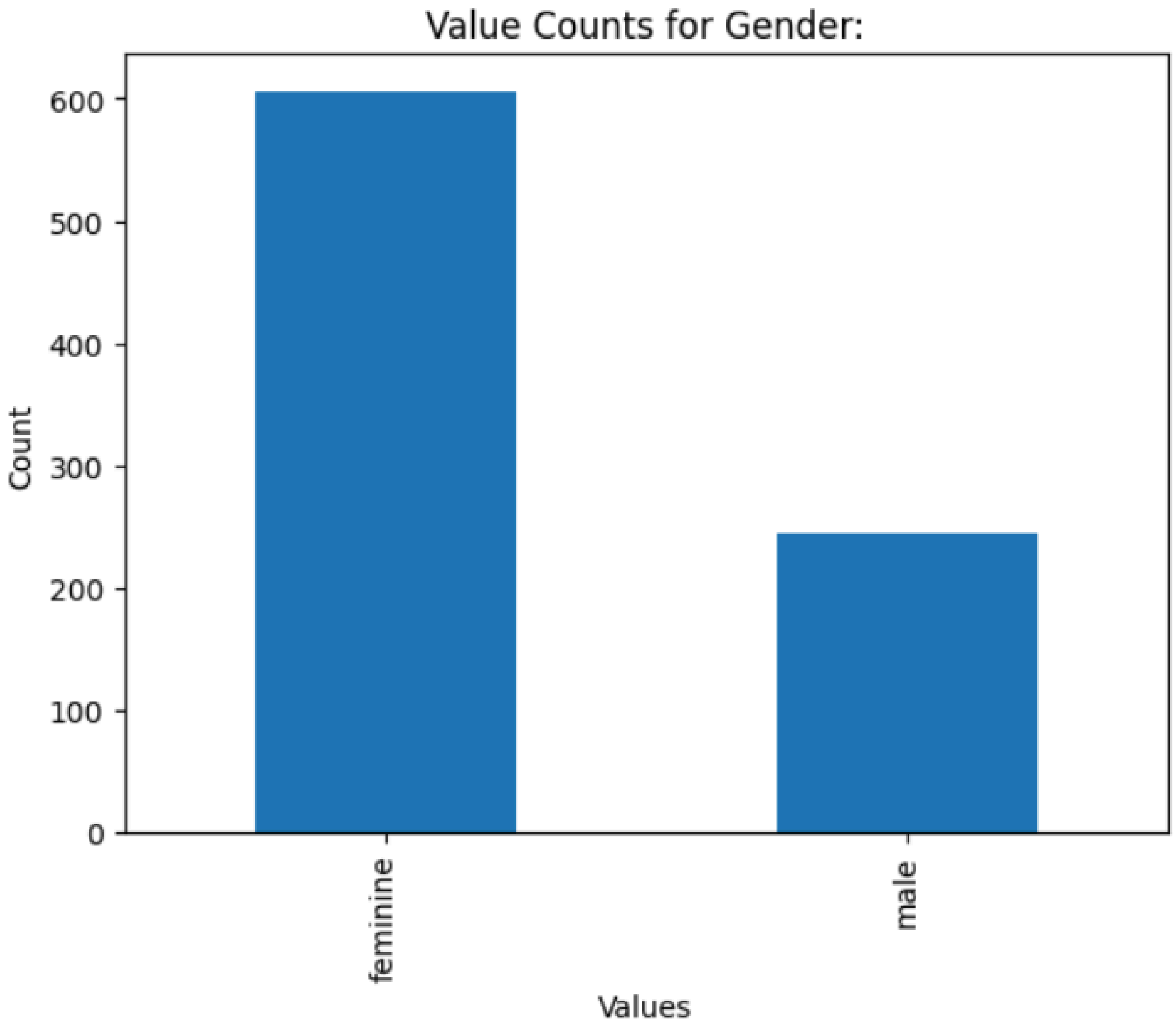
Overview of the Workflow: EDA (Primary)

Steps Taken:
1. Structure investigation: we have 22 columns and 851 Rows.
1.1. Structure of non-numerical features: All the features are nonnumerical- using select_dtypes(exclude="number")

2. Quality Investigation:
2.1. Duplicates: The number of duplicates was 0- using .duplicated().sum()

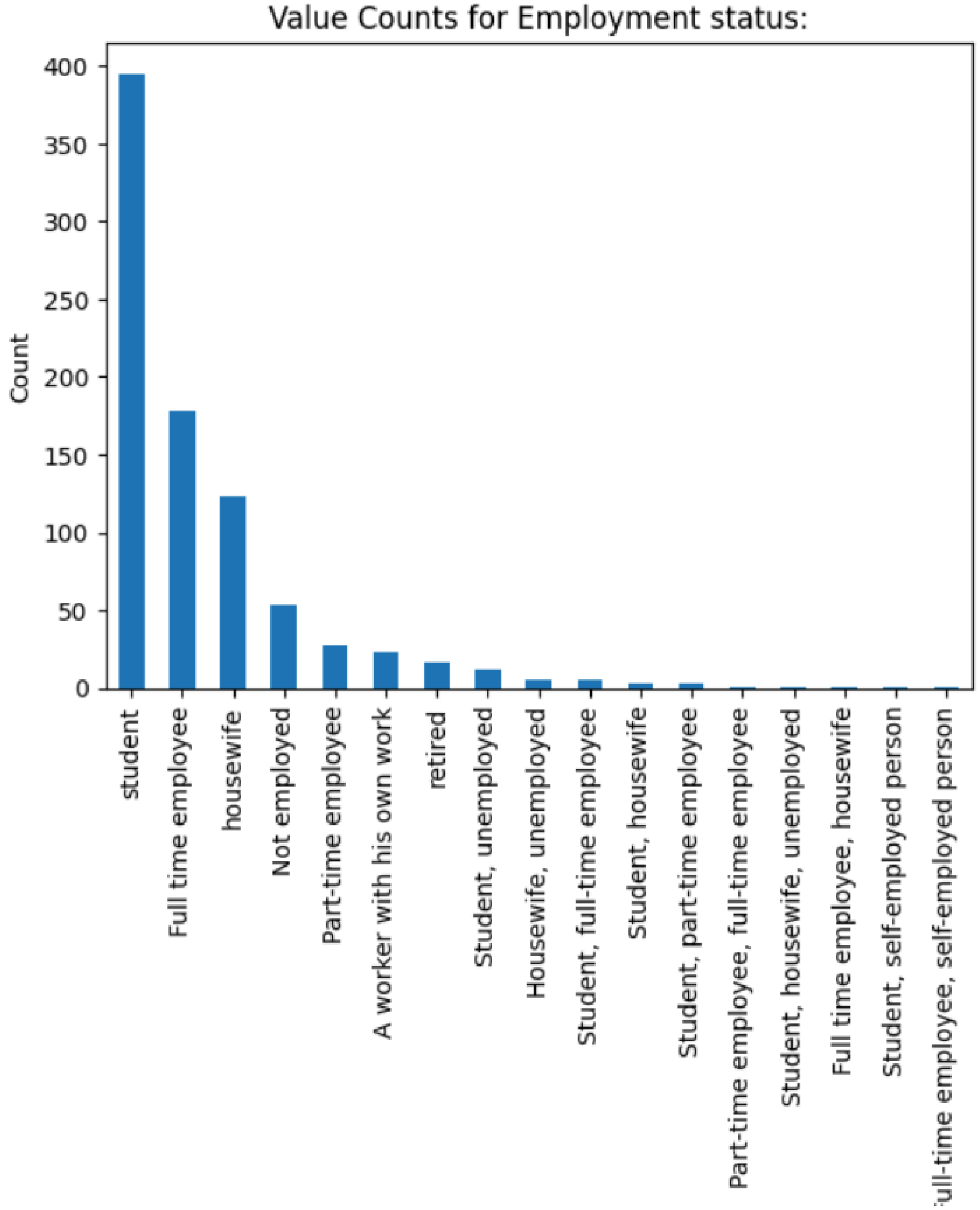
Overview of the Workflow: EDA (Primary)

Steps Taken:
2.3. Unwanted entries and recording errors
2.3.1. Non-Numerical features :loop and bar plot



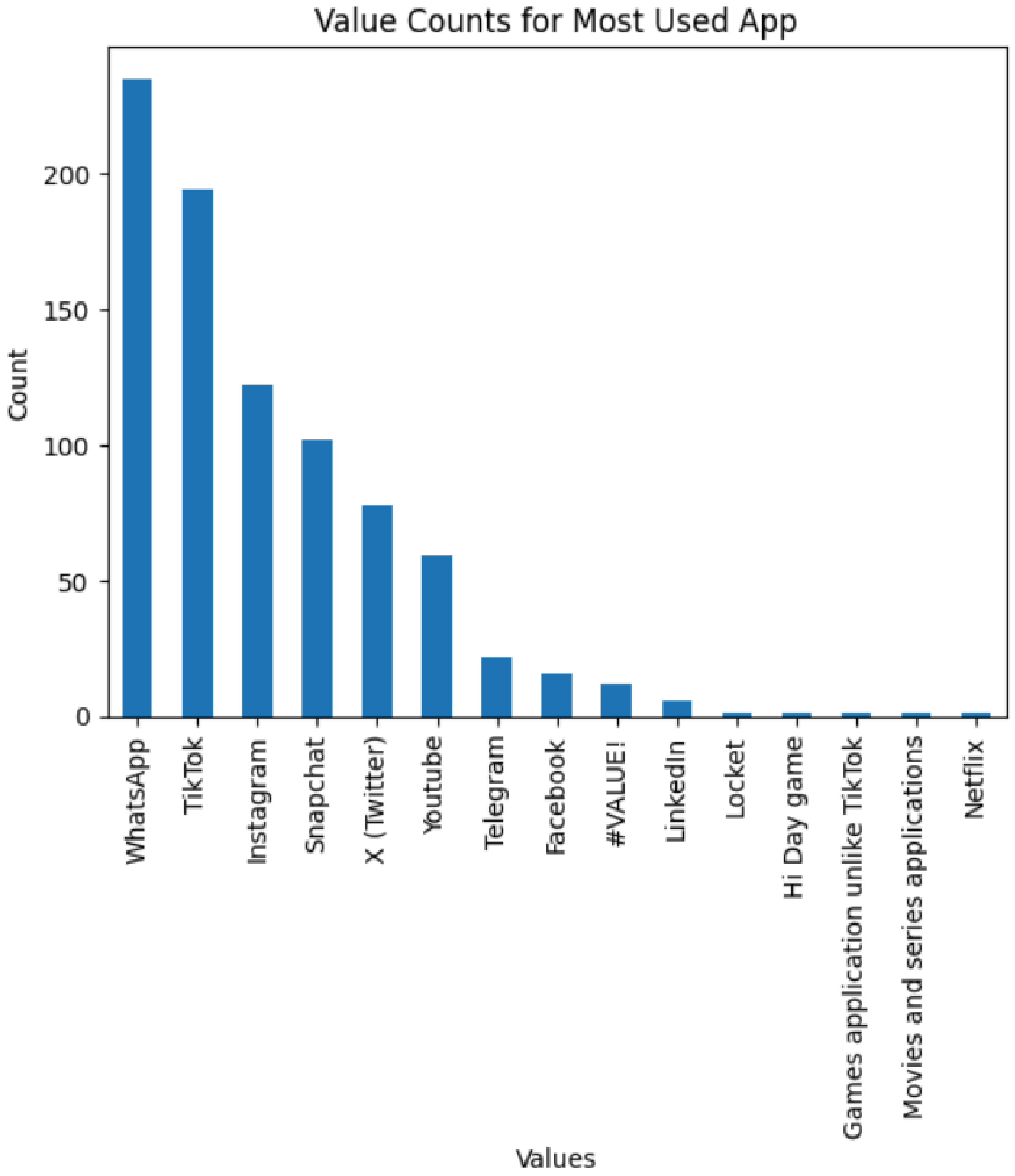
Overview of the Workflow: EDA (Primary)

Steps Taken:
2.3. Unwanted entries and recording errors
2.3.1. Non-Numerical features :loop and bar plot



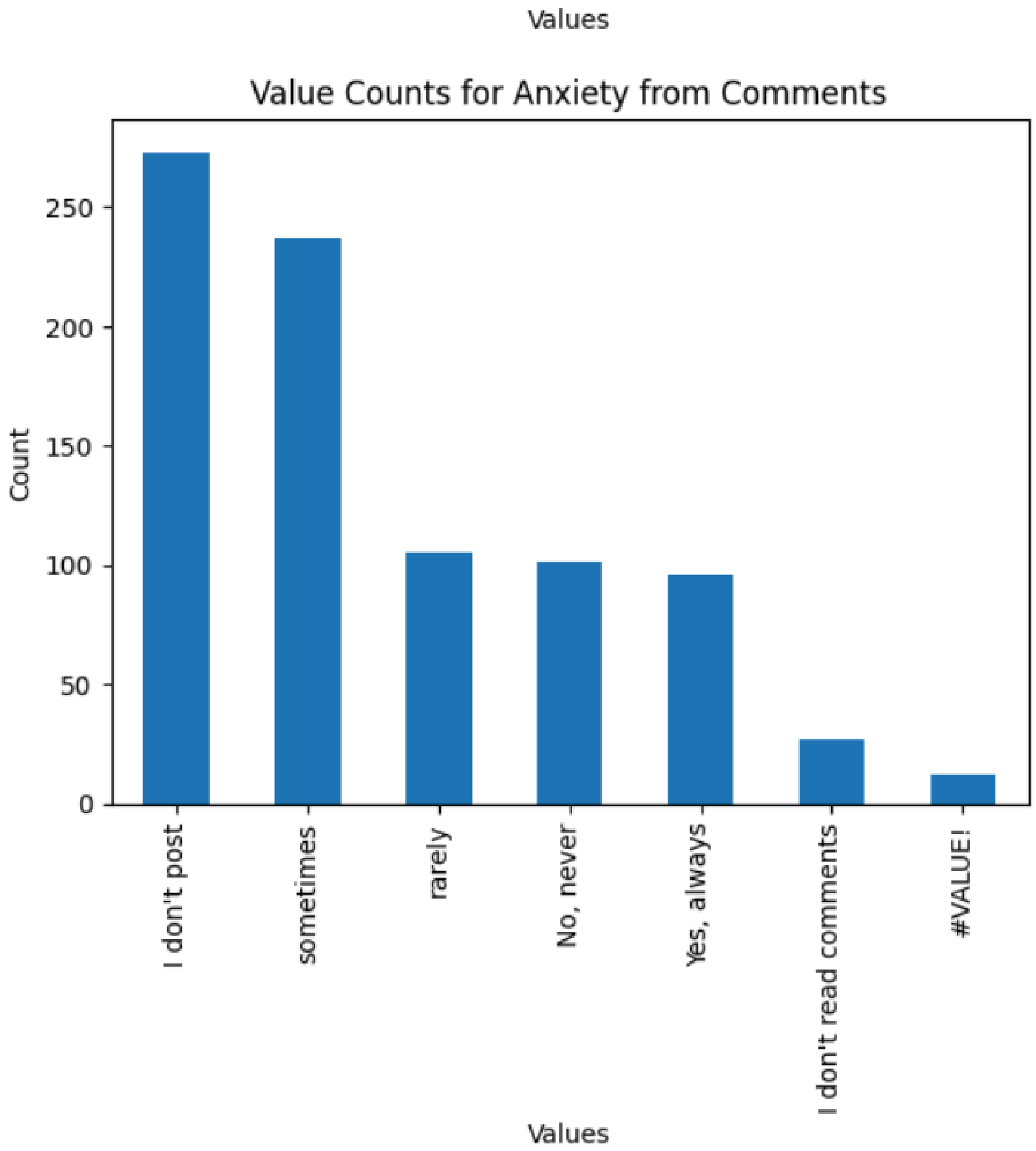
Overview of the Workflow: EDA (Primary)

Steps Taken:
2.3. Unwanted entries and recording errors
2.3.1. Non-Numerical features :loop and bar plot



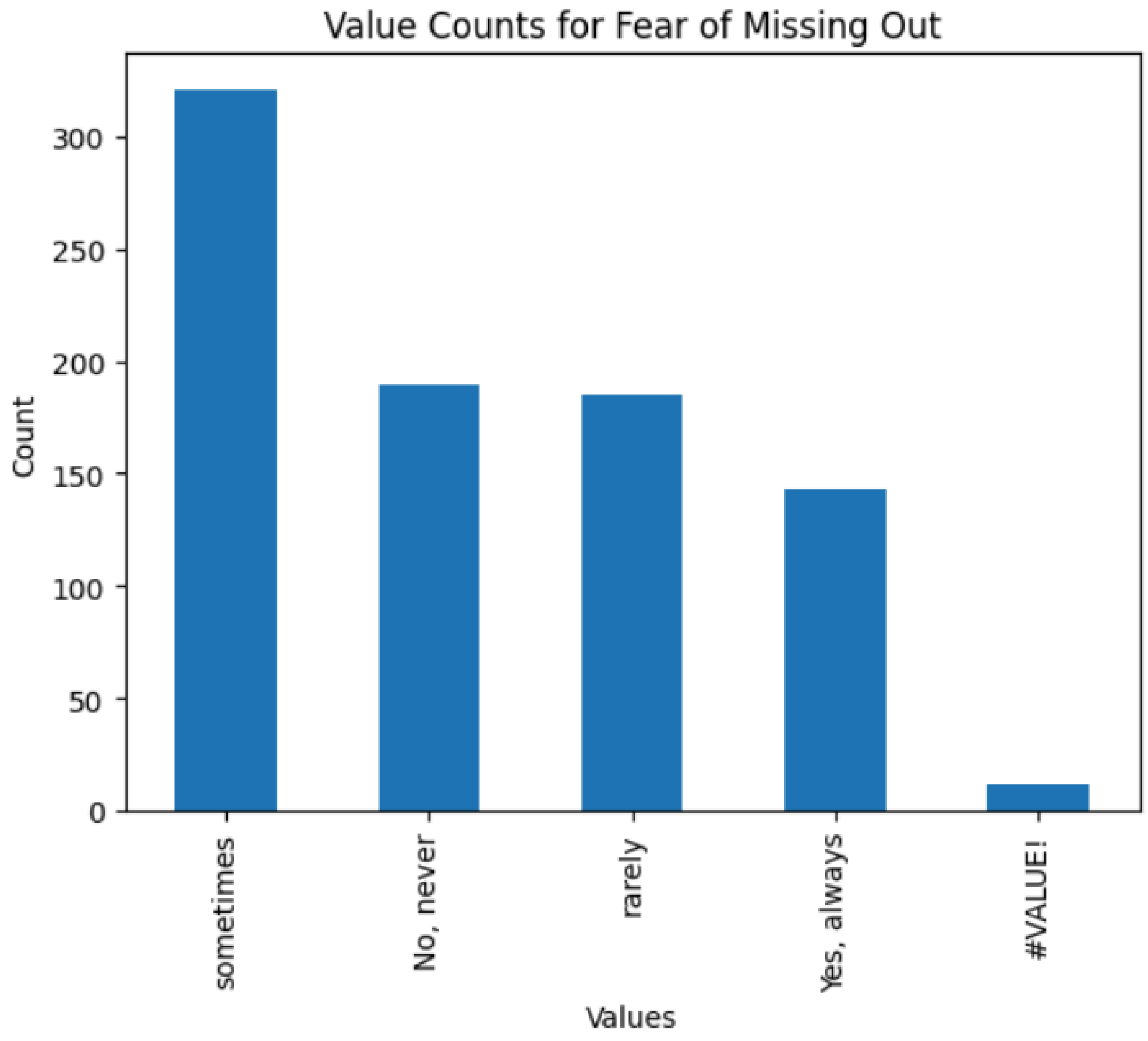
Overview of the Workflow: EDA (Primary)

Steps Taken:
2.3. Unwanted entries and recording errors
2.3.1. Non-Numerical features :loop and bar plot



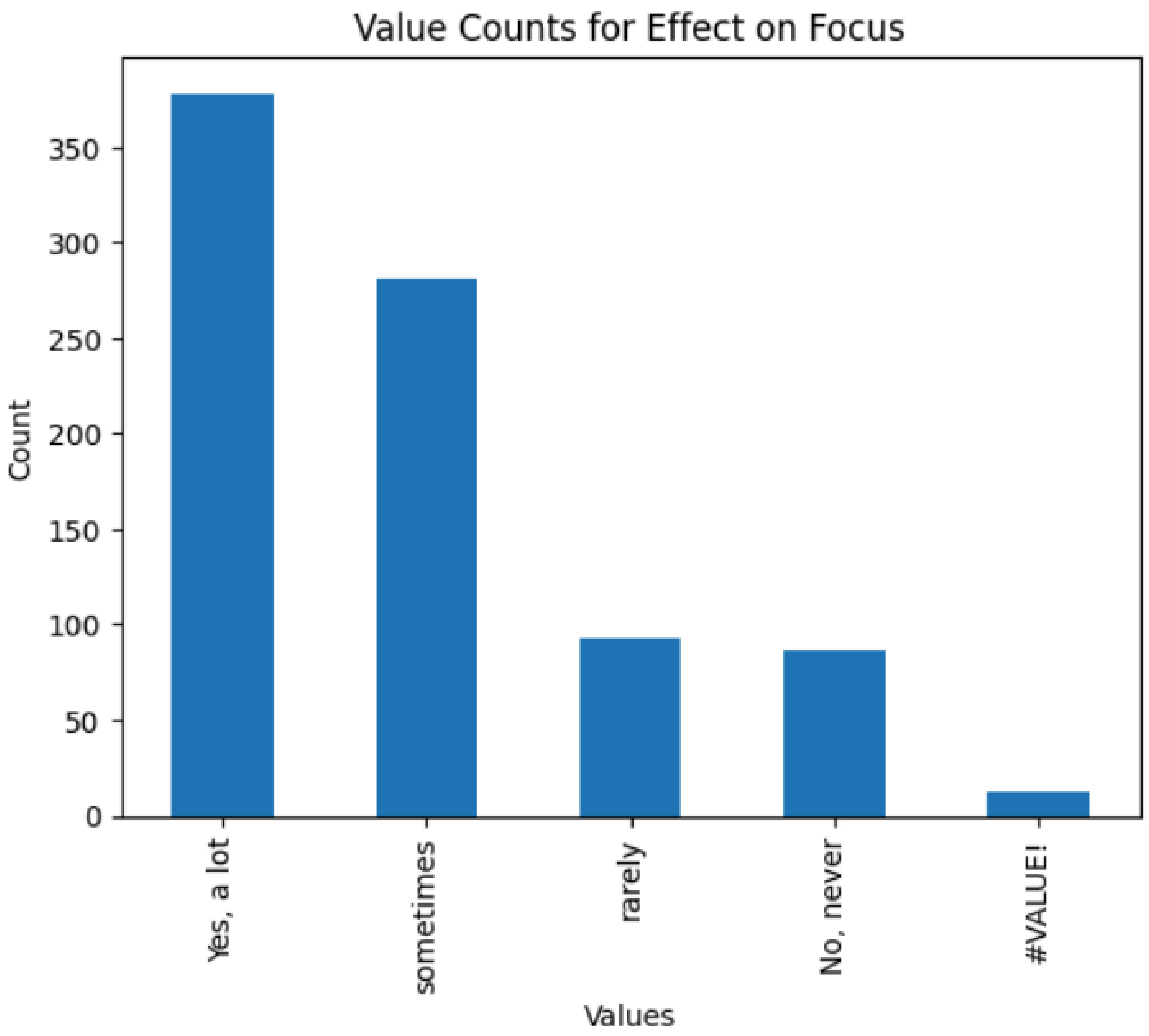
Overview of the Workflow: EDA (Primary)

Steps Taken:
2.3. Unwanted entries and recording errors
2.3.1. Non-Numerical features :loop and bar plot



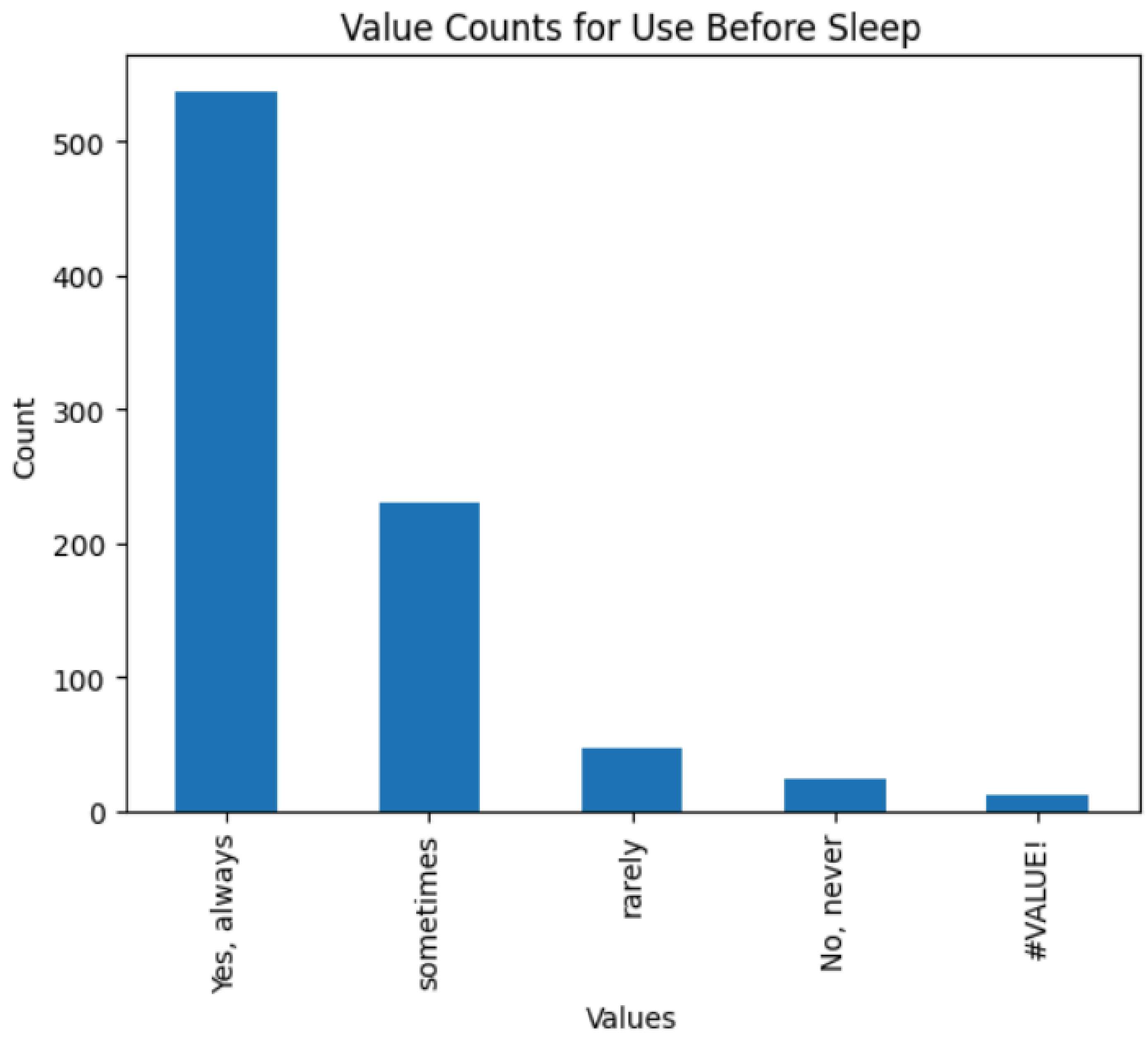
Overview of the Workflow: EDA (Primary)

Steps Taken:
2.3. Unwanted entries and recording errors
2.3.1. Non-Numerical features :loop and bar plot



Overview of the Workflow: EDA (Primary)

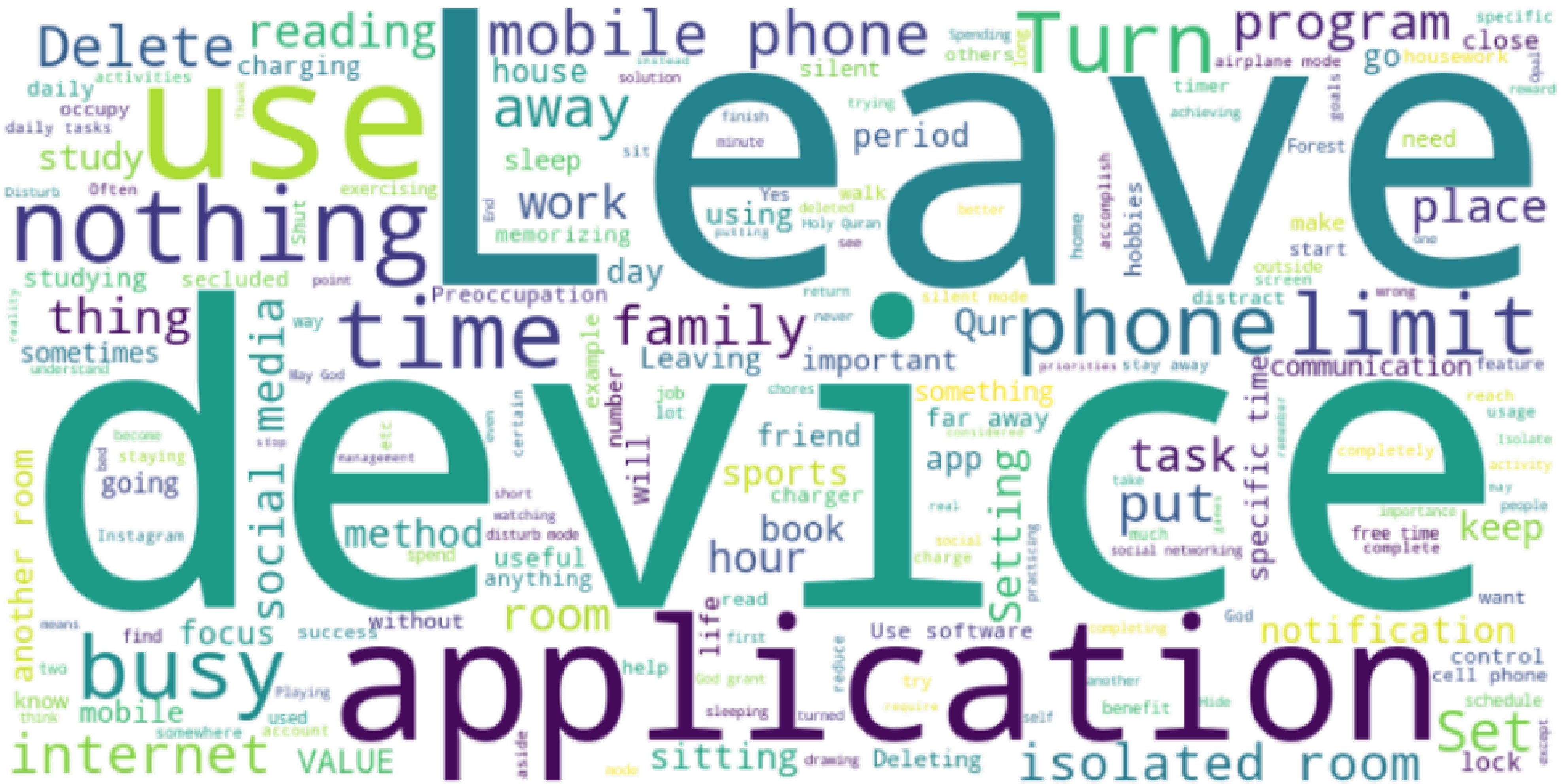
Steps Taken:
2.3. Unwanted entries and recording errors
2.3.1. Non-Numerical features :loop and bar plot



Overview of the Workflow: EDA (Primary)

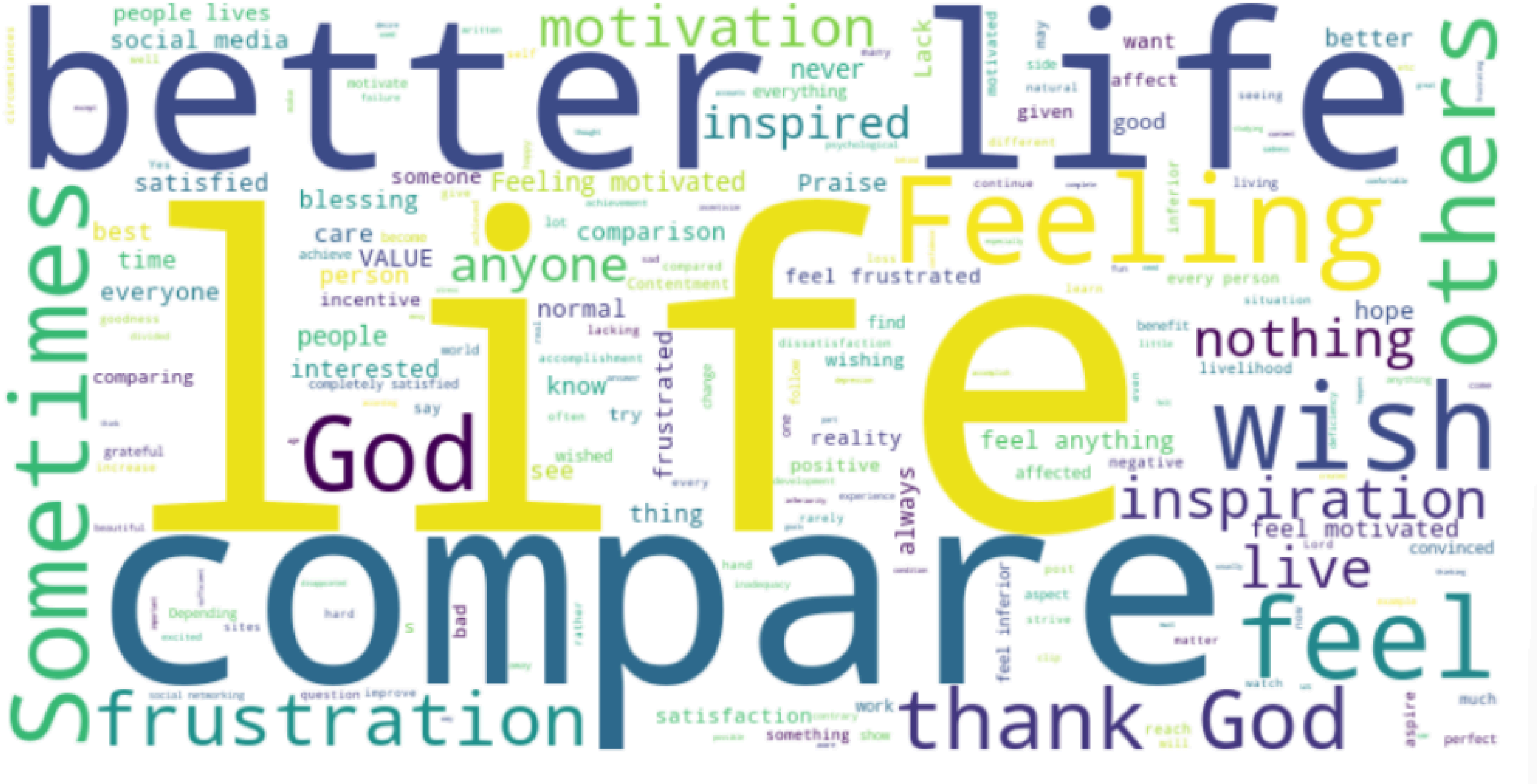
Steps Taken:
2.3. Unwanted entries and recording errors
2.3.1. Non-Numerical features : Values for Methods to Limit your social media access -using : WordCloud

Word Cloud of Social Media Platforms Used



Overview of the Workflow: EDA (Primary)

Steps Taken:
2.3. Unwanted entries and recording errors
2.3.1. Non-Numerical features : Values for Comparing with others- using :WordCloud



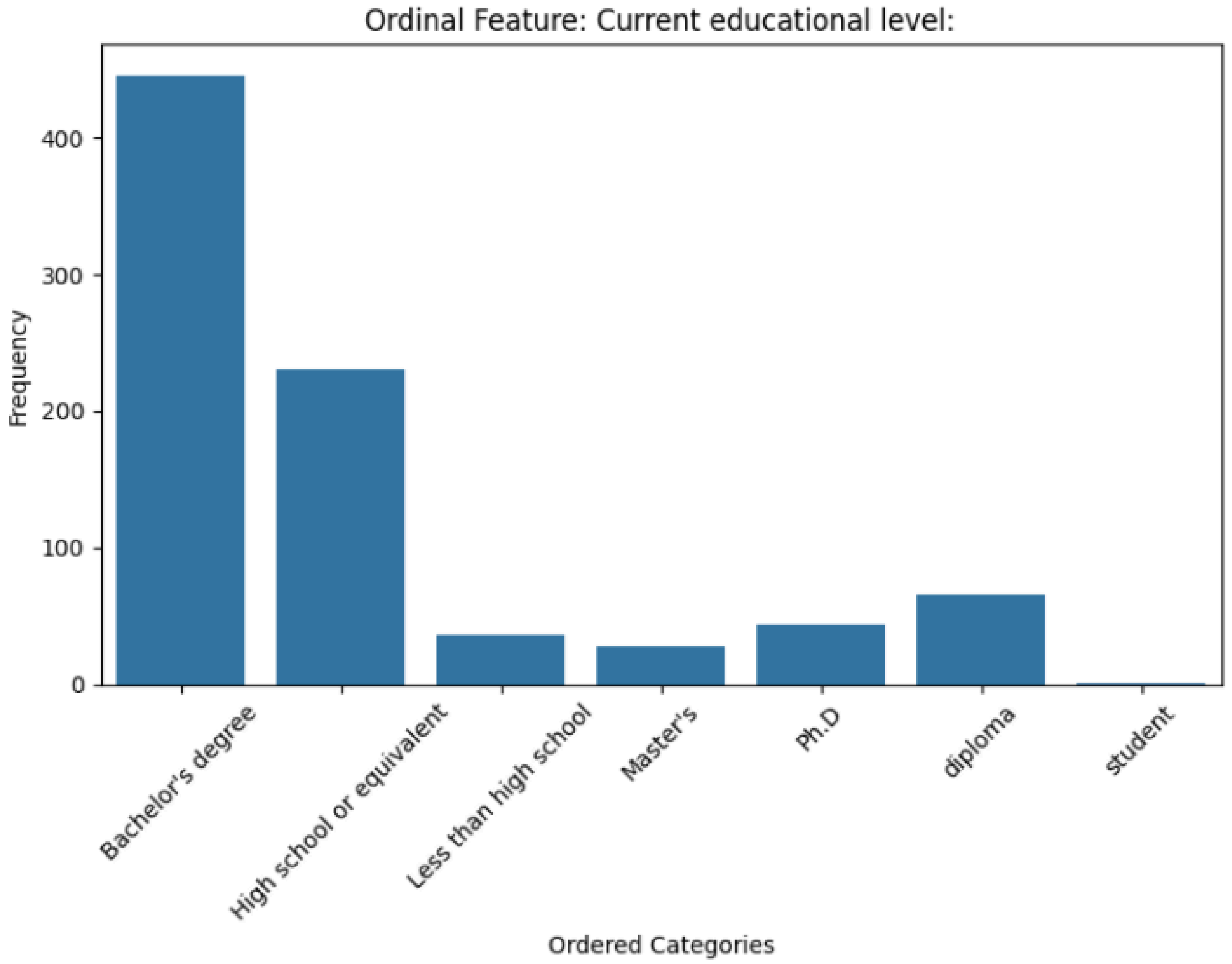
Overview of the Workflow: EDA (Primary)

Steps Taken:

3. Content Investigation:

3.2.1. Continuous features:There's no numerical or datetime columns, nothing to plot.

3.2.2. Discrete and ordinal features



Overview of the Workflow: EDA (Primary)

Steps Taken:
3.4. Conclusion of Content Investigation:

- Social Media Usage Patterns: Most respondents spend low to moderate time on social media; 3-5 hours daily is common.
- Anxiety and FOMO: Moderate levels of anxiety and FOMO are prevalent, with some users experiencing high distress related to likes and comments.
- Impact of Engagement: Strong emotional impact from likes and comments is observed, with moderate-to-high effects.
- Education: Social media use is highest among individuals with a Bachelor’s or high school education.

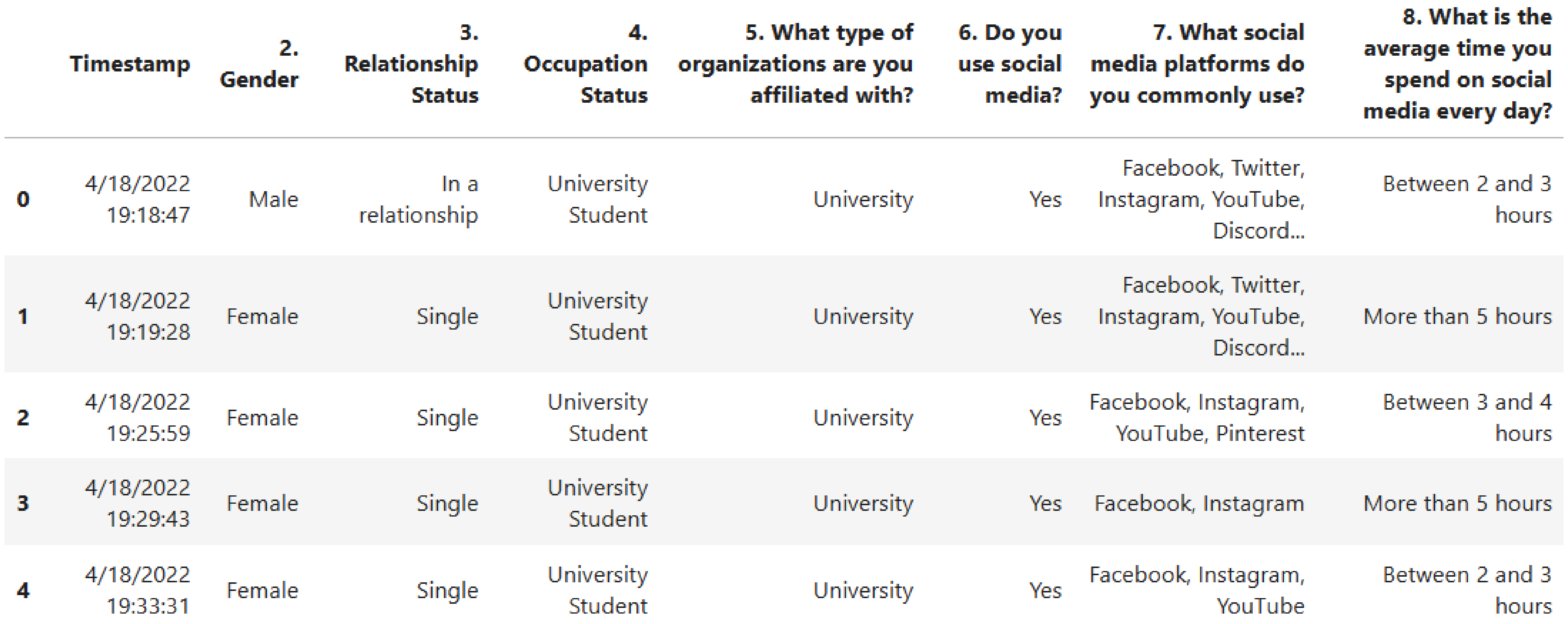
Mental Health Trends: Moderate to high effects on some mental health aspects, particularly anxiety and FOMO, are evident.

Overview of the Workflow: EDA (Secondary)

Steps Taken:

1. Structure investigation: we have 21 columns and 481 Rows.

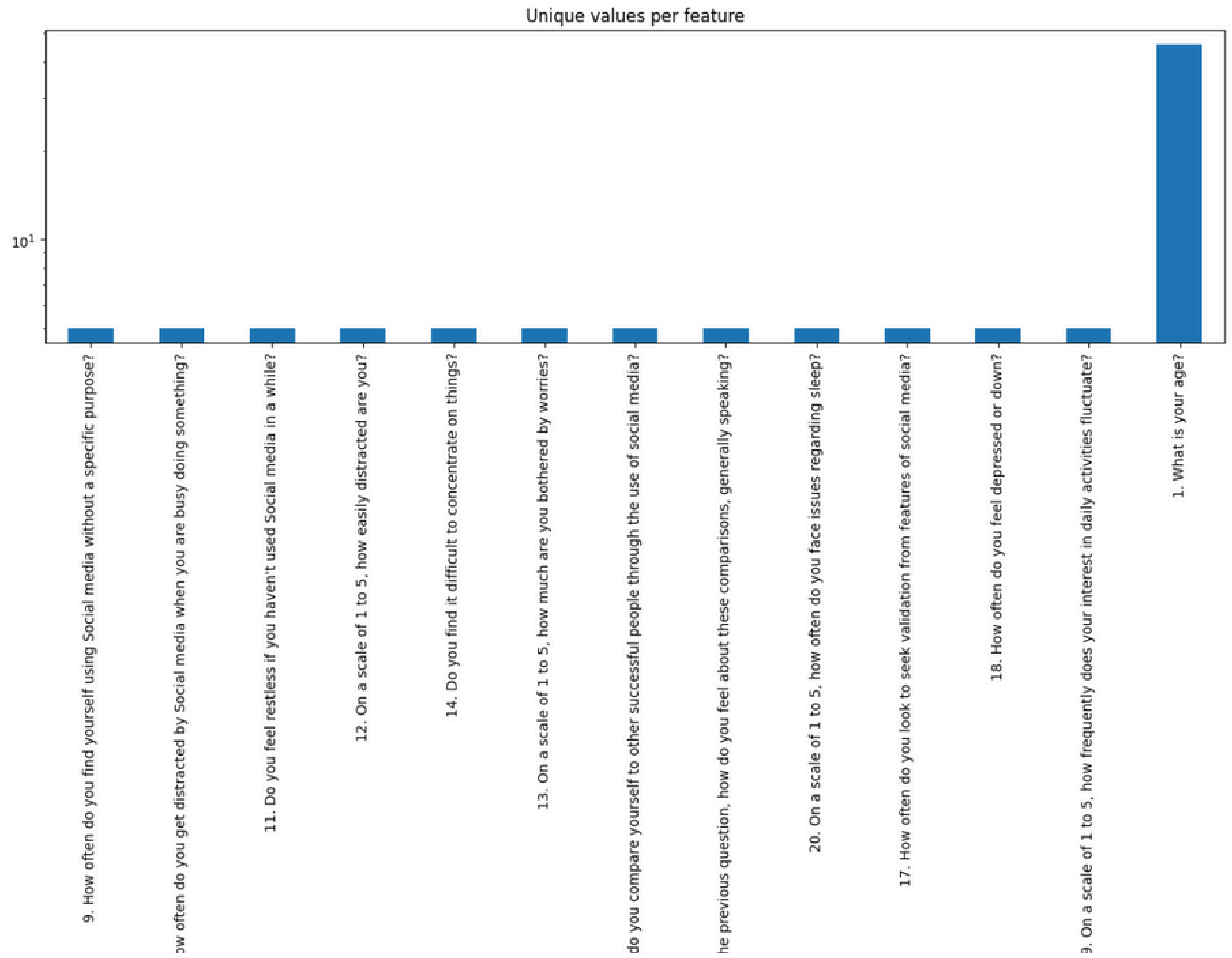
1.1. Structure of non-numerical features:



	Timestamp	2. Gender	3. Relationship Status	4. Occupation Status	5. What type of organizations are you affiliated with?	6. Do you use social media?	7. What social media platforms do you commonly use?	8. What is the average time you spend on social media every day?
0	4/18/2022 19:18:47	Male	In a relationship	University Student	University	Yes	Facebook, Twitter, Instagram, YouTube, Discord...	Between 2 and 3 hours
1	4/18/2022 19:19:28	Female	Single	University Student	University	Yes	Facebook, Twitter, Instagram, YouTube, Discord...	More than 5 hours
2	4/18/2022 19:25:59	Female	Single	University Student	University	Yes	Facebook, Instagram, YouTube, Pinterest	Between 3 and 4 hours
3	4/18/2022 19:29:43	Female	Single	University Student	University	Yes	Facebook, Instagram	More than 5 hours
4	4/18/2022 19:33:31	Female	Single	University Student	University	Yes	Facebook, Instagram, YouTube	Between 2 and 3 hours

Overview of the Workflow: EDA (Secondary)

Steps Taken:
1.2. Structure of numerical features
we determined the unique values in these numerical features- using.nunique().sort_values(): in Age

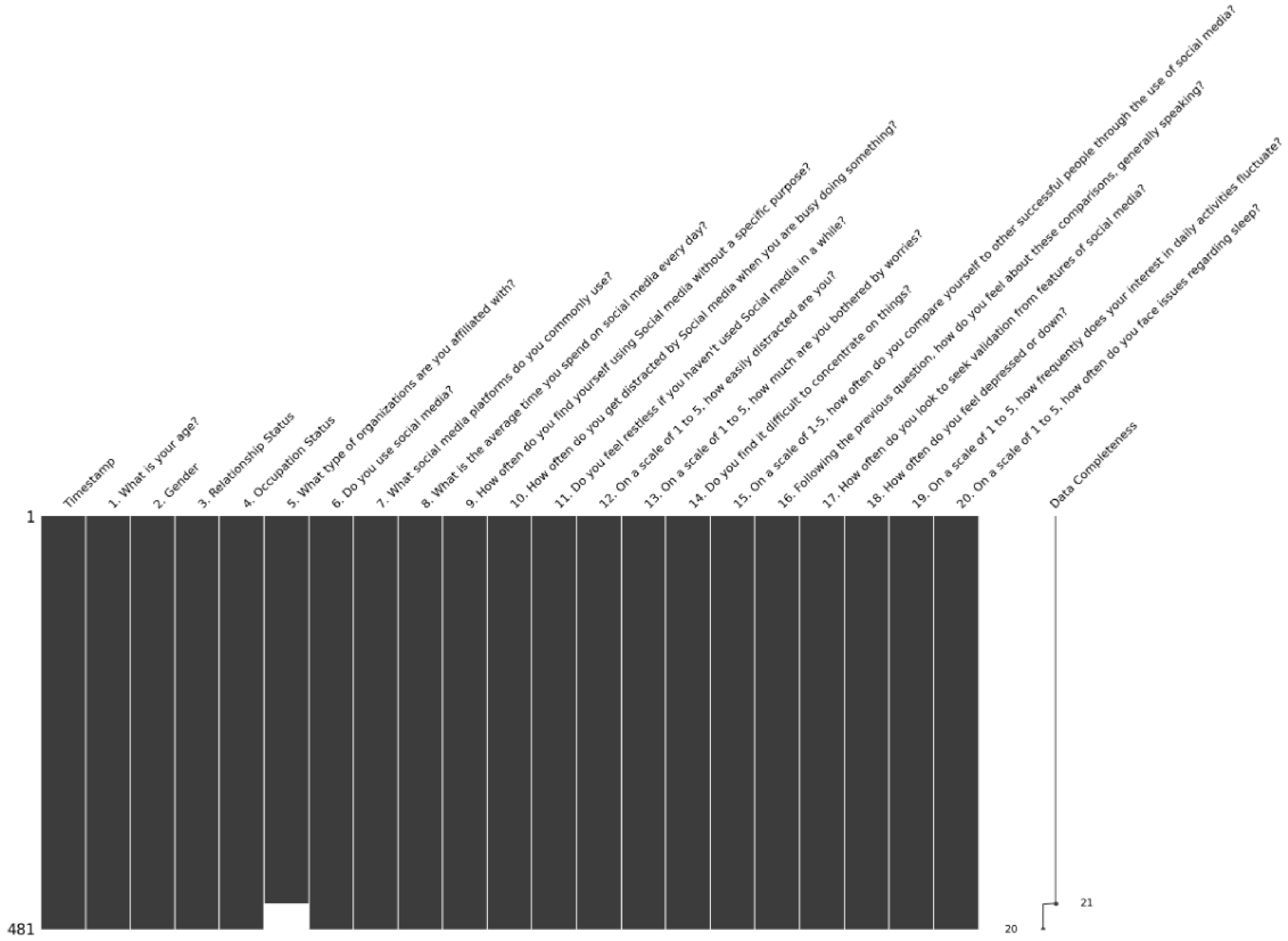


Overview of the Workflow: EDA (Secondary)

Steps Taken:

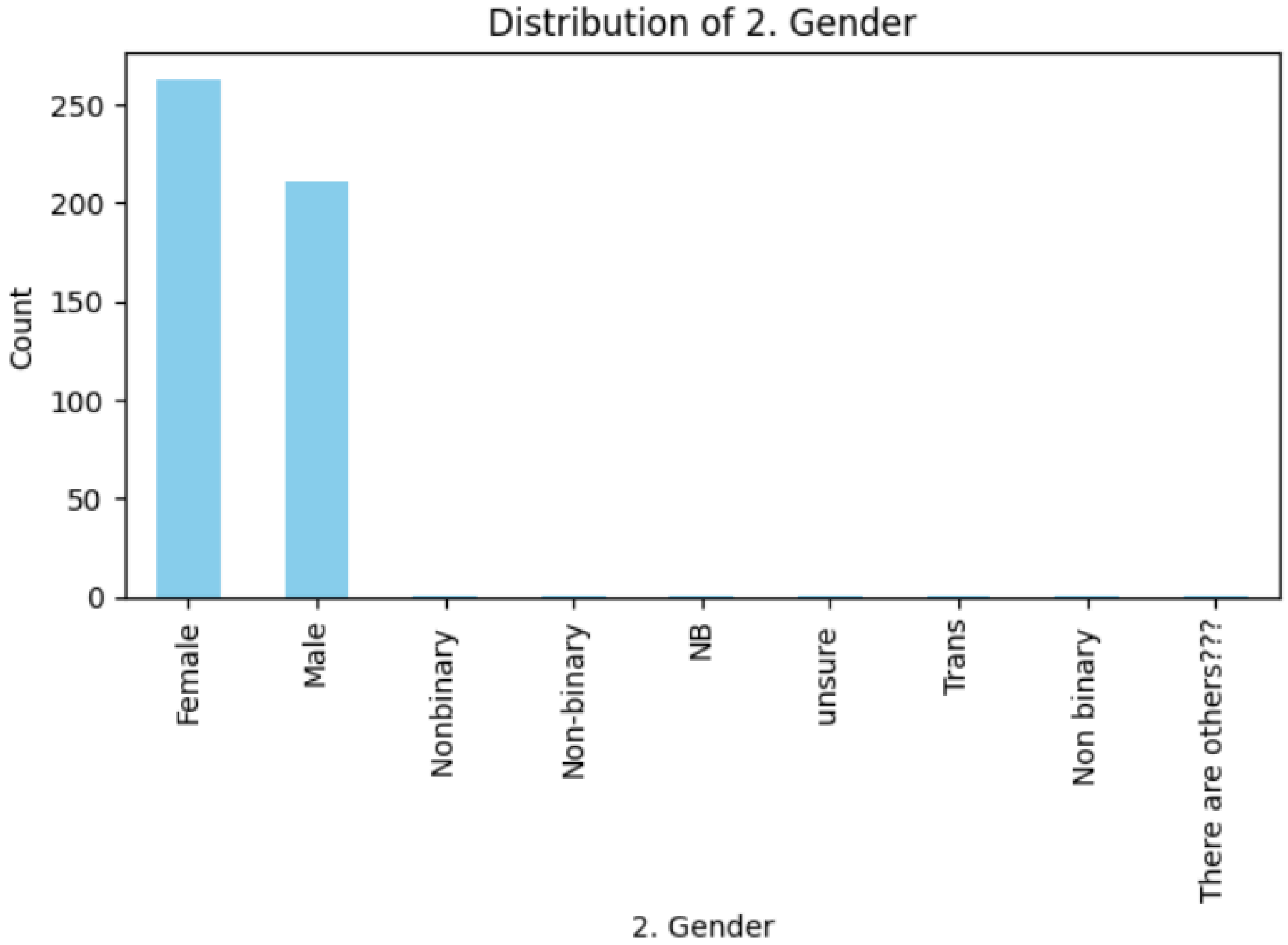
2. Quality Investigation:
2.1. Duplicates: .duplicated().sum()= 0
2.2. Missing values:
 2.2.1. Per sample: missingno

we found some in Q: 5. What type of organizations are you affiliated with? and because it small percentage we dropped the rows that has it



Overview of the Workflow: EDA (Secondary)

Steps Taken:
2.3. Unwanted entries and recording errors
2.3.2. Non-numerical features: different genders

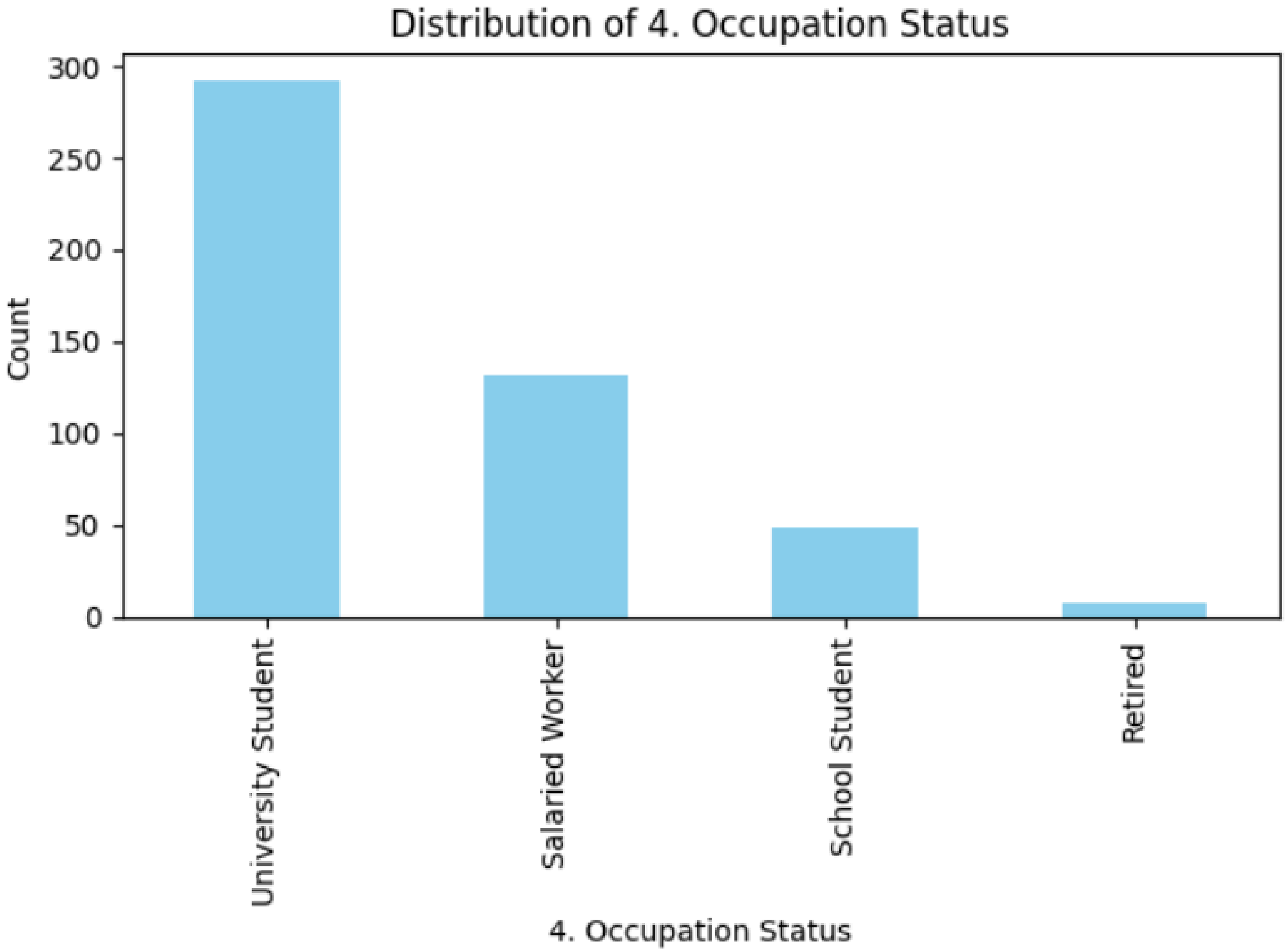


Overview of the Workflow: EDA (Secondary)

Steps Taken:

2.3. Unwanted entries and recording errors

2.3.2. Non-numerical features

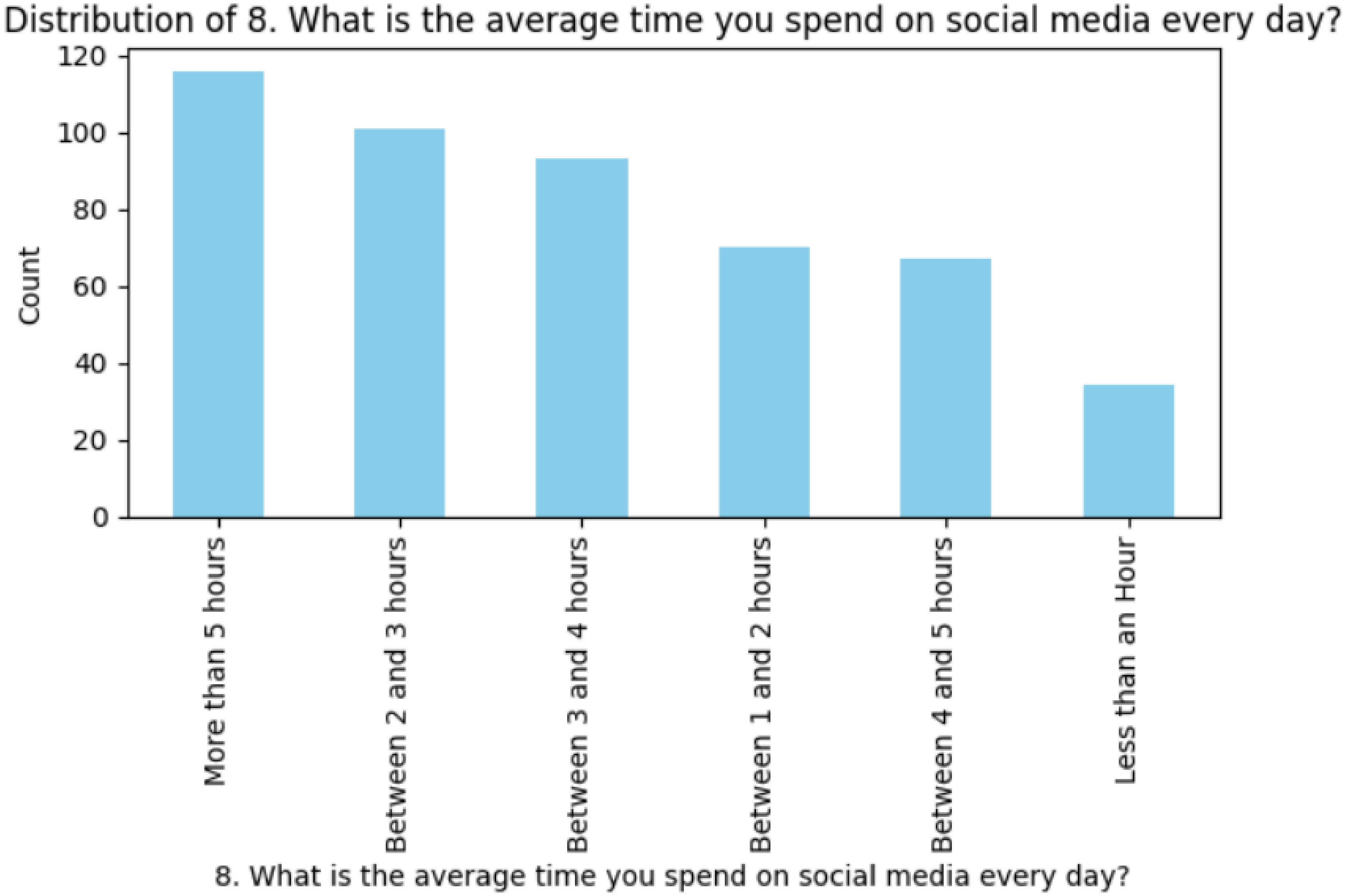


Overview of the Workflow: EDA (Secondary)

Steps Taken:

2.3. Unwanted entries and recording errors

2.3.2. Non-numerical features

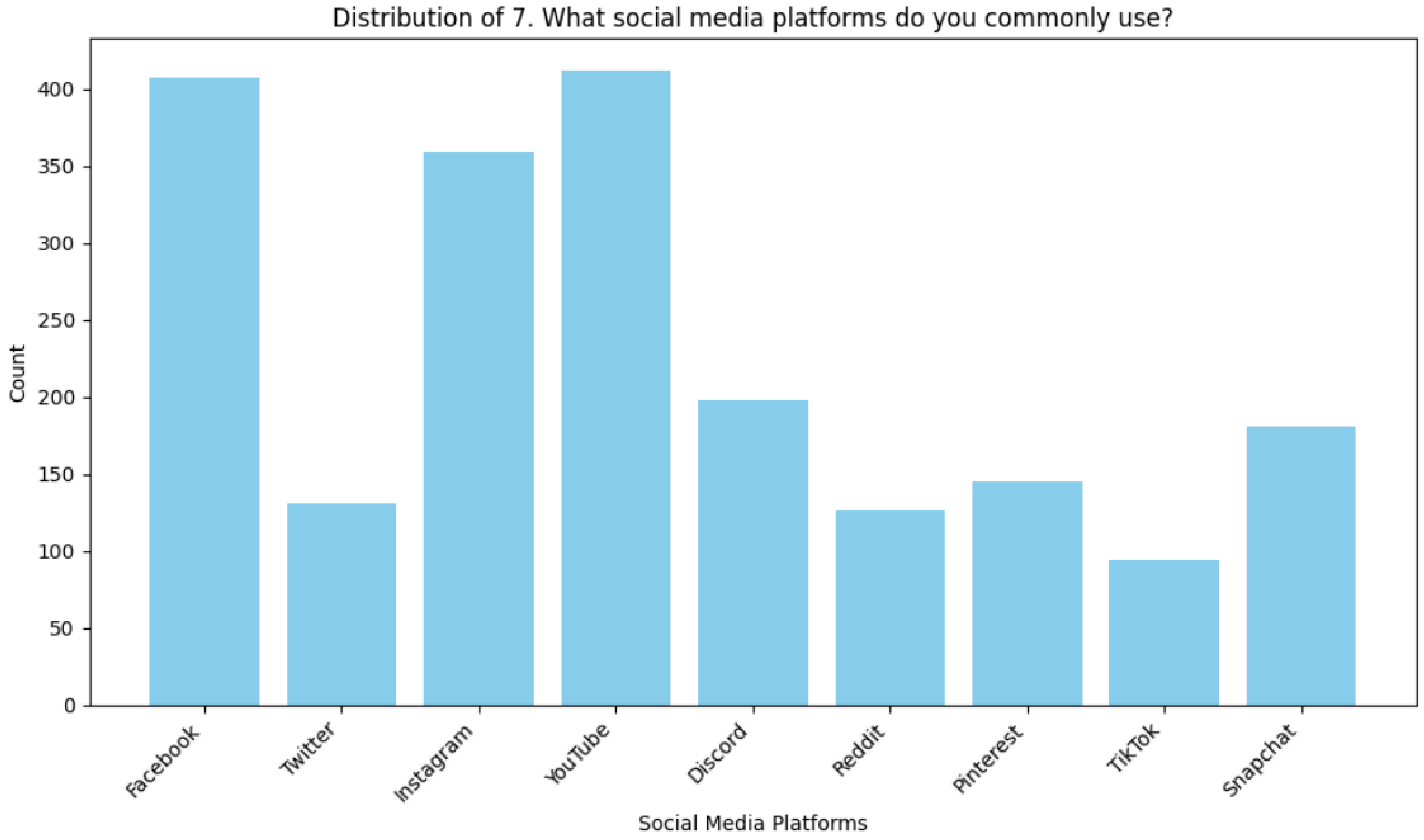


Overview of the Workflow: EDA (Secondary)

Steps Taken:

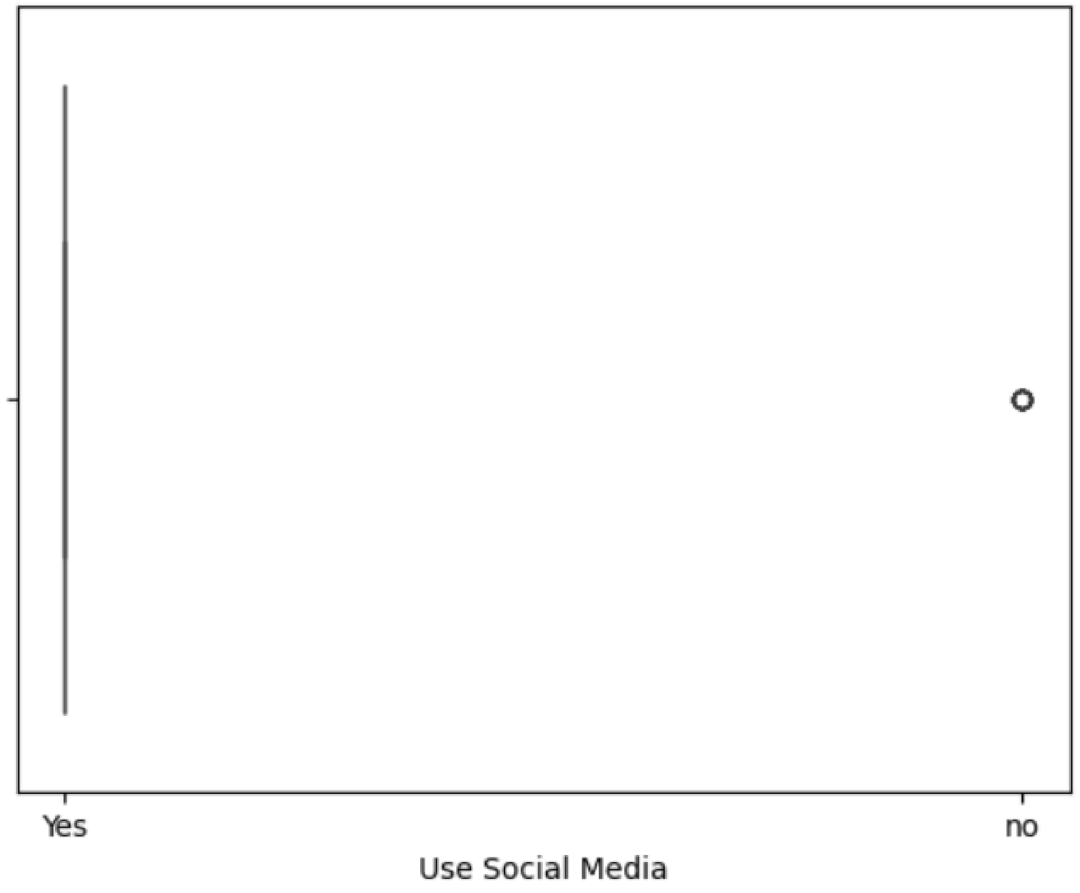
2.3. Unwanted entries and recording errors

2.3.2. Non-numerical features



Overview of the Workflow:Data Cleaning(primary)

Steps Taken:
4.1 Handling Missing Values: no missing values
4.2 Handling Non-Numeric Data
4.3 Handling Outliers: filtered “ no ”then we removed the Q:Use Social Media' column entirely.



Overview of the Workflow:Data Cleaning(primary)

Steps Taken:
4.4 Removing '#VALUE!' value: using replace with the null with NAN.

Overview of the Workflow:Data Cleaning(secondary)

Steps Taken:
Step:1 Handling Irrelevant Features
Are there irrelevant features?

- Low-variance features might be considered irrelevant if they provide little differentiation. For example, "Feelings About Comparisons" has a variance of 1.12, but this does align with the analysis goals, so it couldn't potentially be dropped.

- Step 2: Handling Outlier:
- Number of outliers in age: 1
- Outliers:Find the outliers (ages less than 10 or greater than 75

- 1. What is your age?
- 256 91.0

- Dealing with the rest of the outliers :

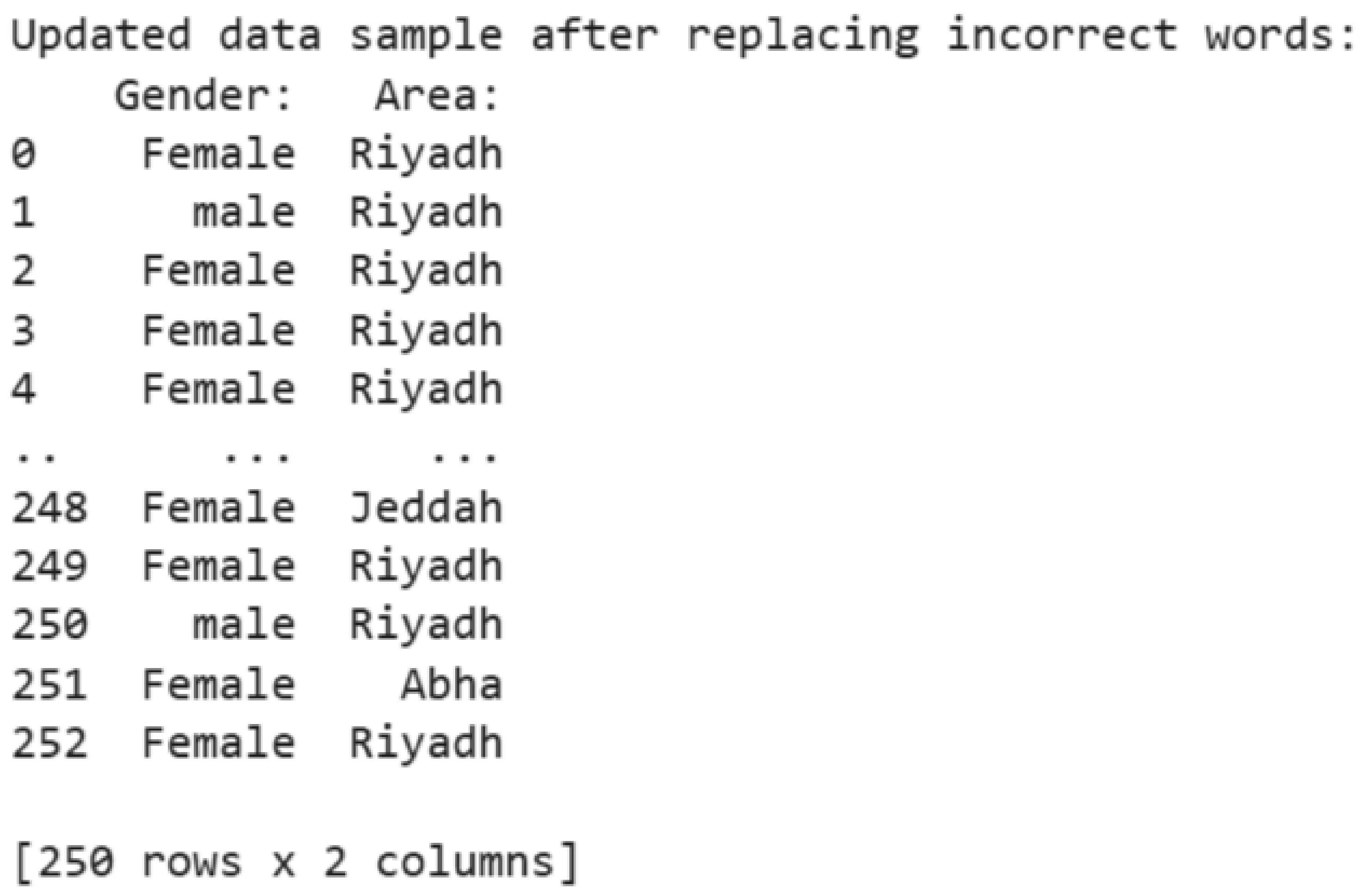
Likert scale outliers like 1 or 5 are typically valid responses and shouldn't be removed, so i decide leave these values as they are and save dataset without removing them

Overview of the Workflow:Pre-Prossesing (primary)

Steps Taken:

- Inconsistent Gender Values: Replaced "feminine" with "Female" in the Gender column.
- City Name Corrections:
- "grandmother" corrected to "Jeddah."
- "the news" corrected to "Khobar."
- "City" corrected to "Madinah."

All using replace()



```
Updated data sample after replacing incorrect words:
    Gender:   Area:
0     Female  Riyadh
1       male  Riyadh
2     Female  Riyadh
3     Female  Riyadh
4     Female  Riyadh
..       ...     ...
248   Female  Jeddah
249   Female  Riyadh
250     male  Riyadh
251   Female    Abha
252   Female  Riyadh

[250 rows x 2 columns]
```

Overview of the Workflow:Pre-Prossesing (primary)

Steps Taken:

Range(1-5):

- columns_to_convert =[Q]
- response_mapping =[numbers]
- converting using replace()

Overview of the Workflow:Pre-Prossesing (Secondary)

Steps Taken:

- Encoding:
- replace({'Male': 0, 'Female': 1})

Overview of the Workflow:Open-ended question analysis(primary)

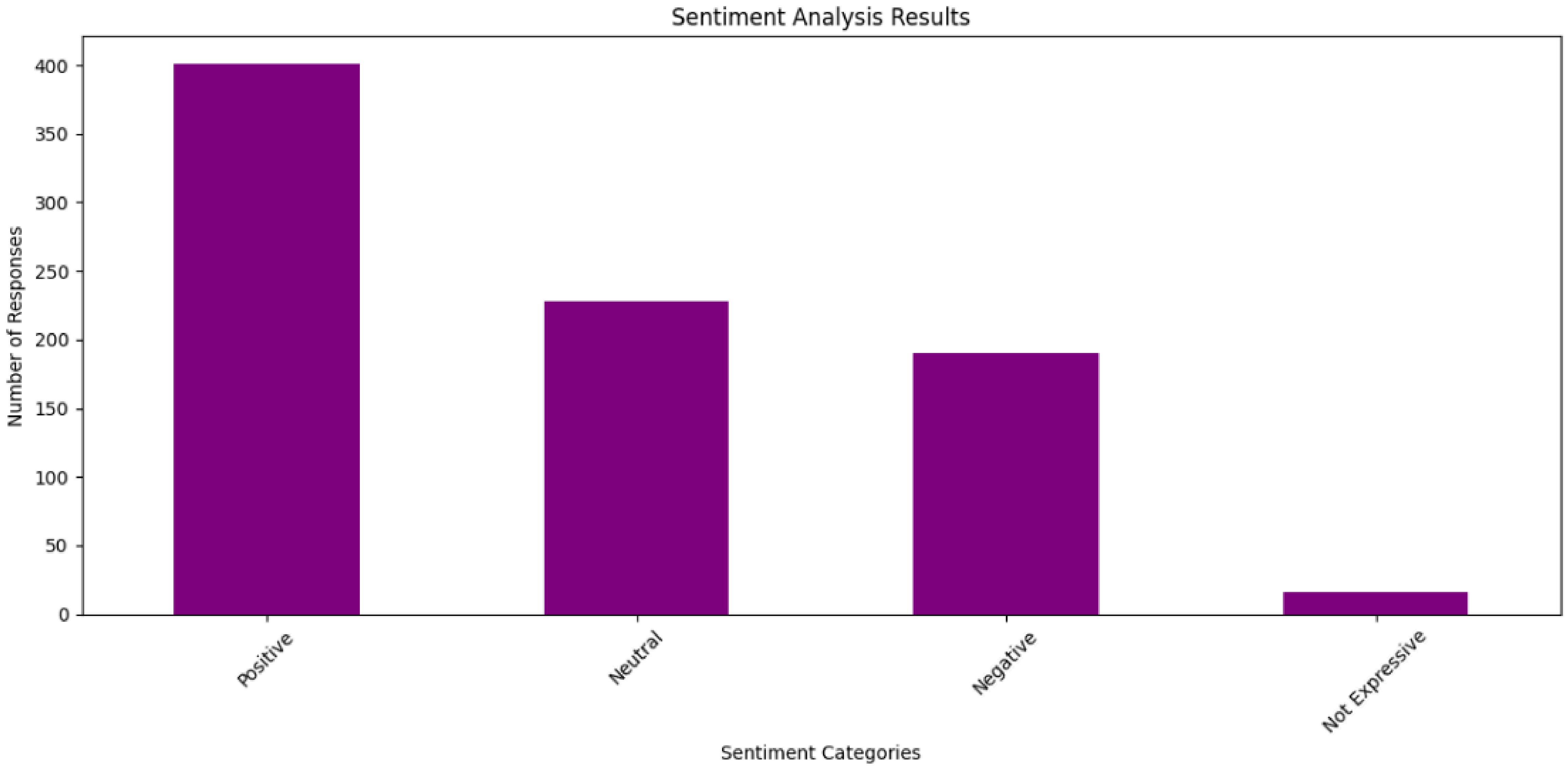
Q1: How do you feel when you compare your life to the lives of others on social media?

Steps Taken:

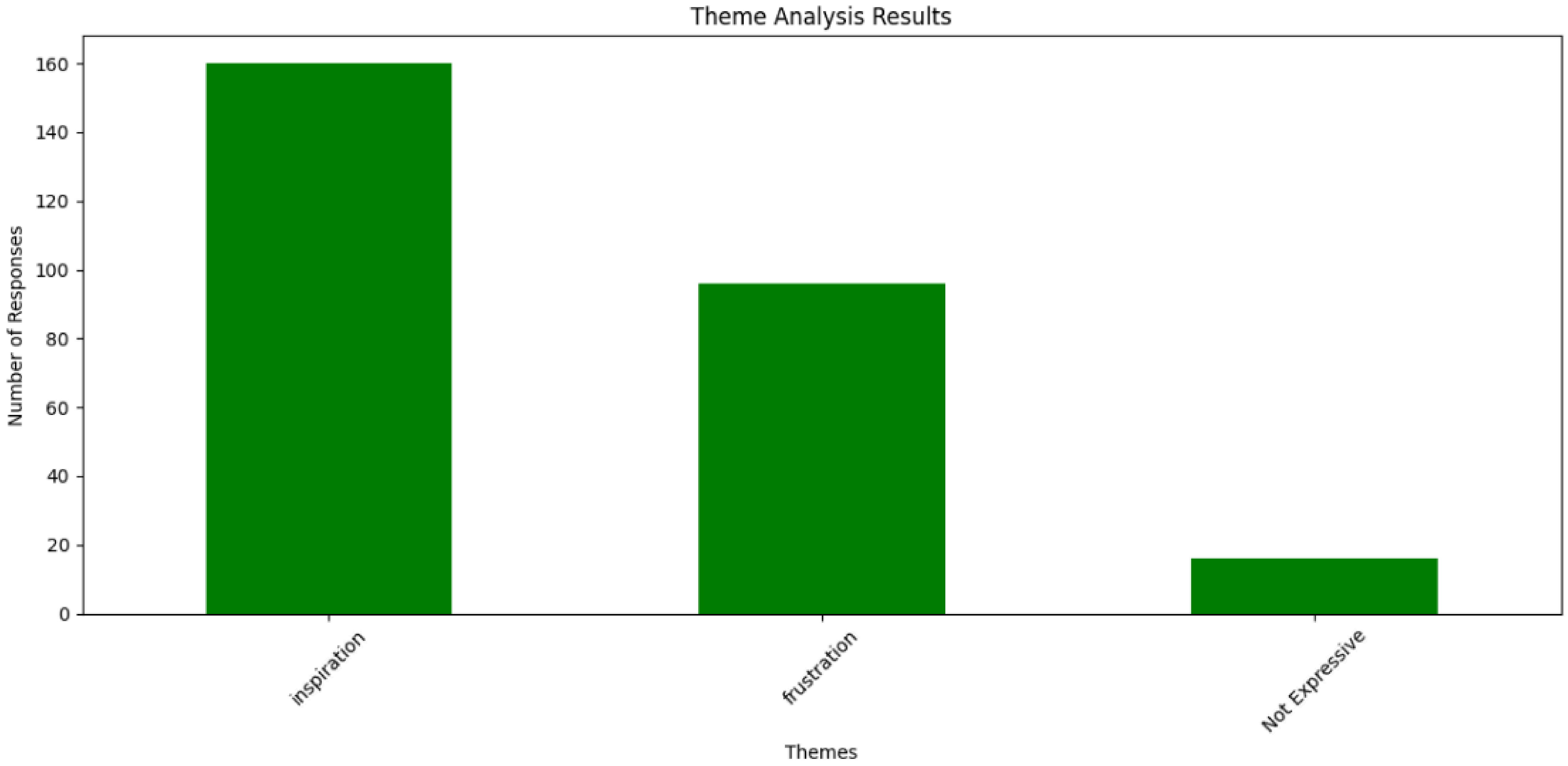
Framework for Analysis:

1. Sentiment Analysis:
   - Used VADER (Valence Aware Dictionary and sEntiment Reasoner) for sentiment scoring.
   - Categorized sentiments into Positive, Neutral, and Negative based on predefined score thresholds.
2. Thematic Categorization:
   - Created themes (e.g., Inspiration, Frustration) using predefined keyword dictionaries.
3. Filtering Non-Expressive Text:
   - Removed meaningless or irrelevant responses to focus on substantial insights.
4. Visualization:
   - Used Matplotlib to visualize the distribution of themes and sentiments.
   - Bar plots provided clear representation of the frequency of themes and sentiments.

Overview of the Workflow:Open-ended question analysis(primary)



Overview of the Workflow:Open-ended question analysis(primary)



Overview of the Workflow:Open-ended question analysis(primary)

Q2: What methods, if any, do you use to limit your social media access?

Steps Taken:1. Data Cleaning

- Removed irrelevant or meaningless responses (e.g., '.', '??', '...').
- Applied a cleaning function to standardize and filter out invalid text entries.

2. Tokenization

- Tokenized cleaned responses into individual words for analysis.

3. Initial Coding

- Created initial codes based on keywords related to strategies (e.g., "limit," "apps," "detox").
- Mapped tokens in responses to corresponding codes.

4. Thematic Assignment

- Grouped initial codes into broader themes such as:
  - Device Management
  - Time Management
  - Mindfulness and Self-Reflection
  - Social Interaction
  - Distraction Techniques
- Assigned themes to responses based on the identified codes.

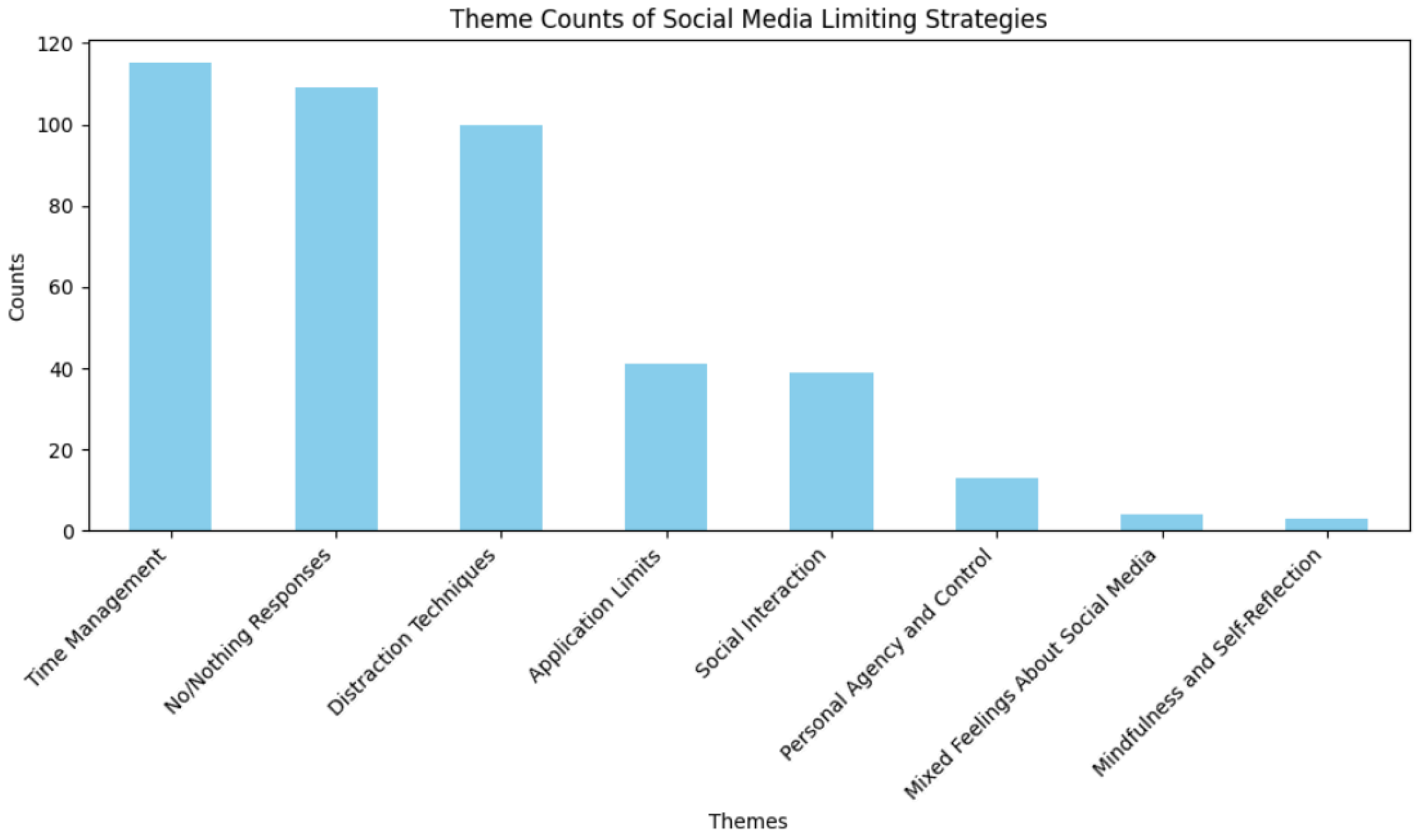
5. Frequency Analysis

- Counted the occurrence of each theme to determine which strategies were most common.

6. Visualization

- Used a bar chart to display the frequency of themes, with clear labels and formatting for easy interpretation.

Overview of the Workflow:Open-ended question analysis(primary)

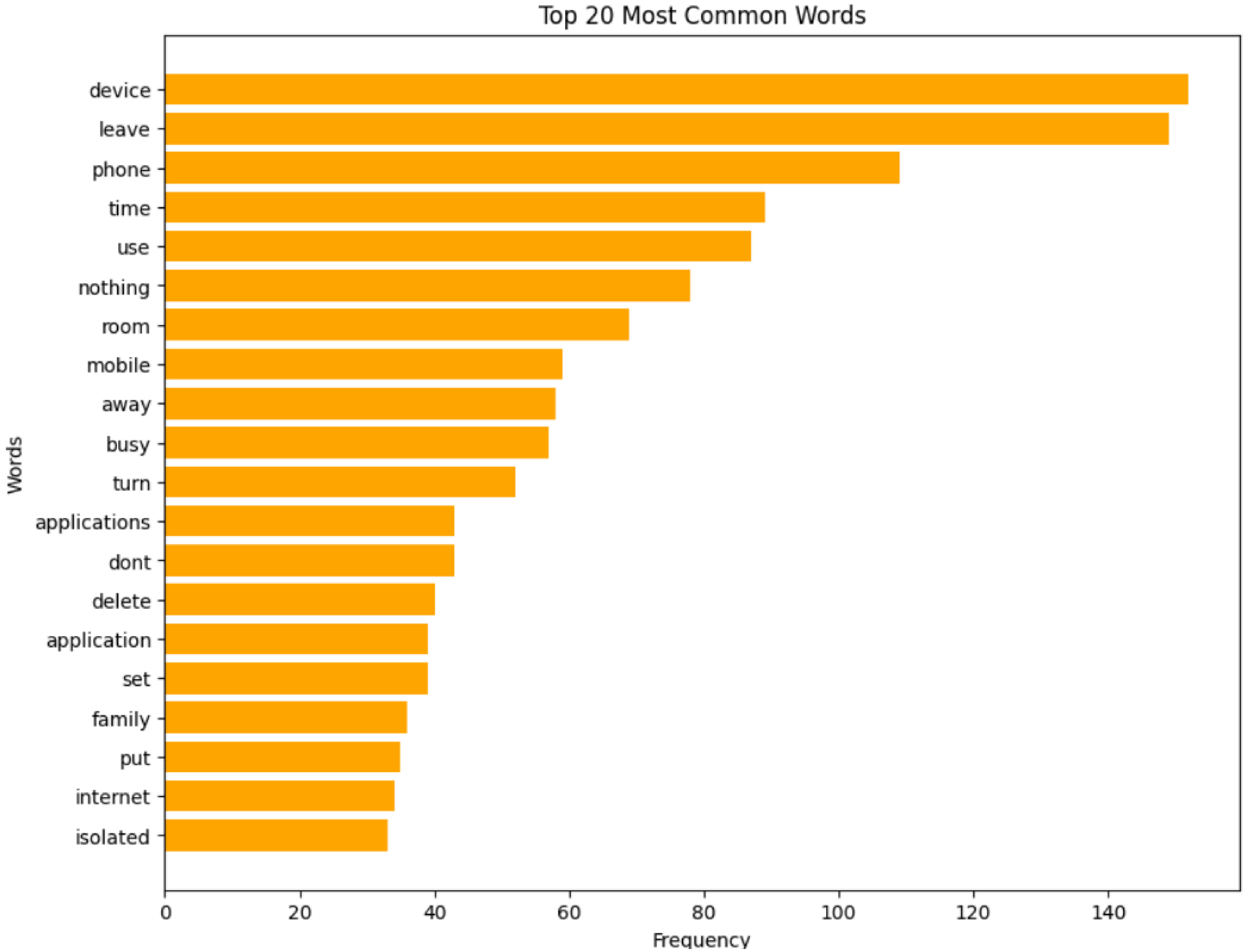


Overview of the Workflow:Open-ended question analysis(primary)

Q2: What methods, if any, do you use to limit your social media access?

- Text Cleaning and Tokenization
- Word Frequency Calculation
- Identified the 20 most common words and their frequencies.

Overview of the Workflow:Open-ended question analysis(primary)



Overview of the Workflow: statical summary(primary)

- **Time Spent on Social Media:**

Most people spend 3–5 hours daily, with a few reporting up to 12 hours.

- **Anxiety from Comments:**

Moderate levels of anxiety are common, with scores mostly between 3 and 4.

- **Fear of Missing Out (FOMO):**

Many experience moderate FOMO, with scores concentrated around 3–4.

- **Effect on Focus:**

Mild impacts on focus are reported, with most scores ranging from 2 to 3.

- **Use Before Sleep:**

Most users report little to no social media use before bed, with low average scores.

Overview of the Workflow: statical summary(Secondary)

- **FOMO**: Mild restlessness is common, with scores mostly between 2 and 3.
- **Ease of Distraction**: Many individuals find themselves moderately to highly distracted, with most responses between 3 and 4.
- **Anxiety**: Moderate to high levels of worry are common, with most scores between 4 and 5.
- **Effect on Focus**: Moderate difficulty concentrating is reported, with most scores between 3 and 4.

Overview of the Workflow:Model Building(primary and secondary)

Implemented models:

- Linear Regression
- Random Forest
- Gradient Boosting (e.g., XGBoost)
- KNN
- Support Vector Regressor (SVR).

Mental Health Index (MHI) served as the target variable, calculated from weighted contributions of factors like anxiety, FOMO, etc.

Overview of the Workflow:model evaluation(primary and secondary)

Model Evaluation:

- Mean Squared Error (MSE)
- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- R-squared (R^2)

Algorithms Selected and Rationale

Algorithms Selected and Rationale

a) Linear Regression:

- Rationale: Benchmark model to establish baseline performance and relationships between predictors and outcomes.

b) Random Forest:

- Rationale: Effective for capturing non-linear relationships and feature interactions. Provides insights through feature importance.

c) Gradient Boosting (e.g., XGBoost):

- Rationale: Reduces residual errors iteratively for better performance on noisy and imbalanced datasets.

d) Support Vector Regressor (SVR):

- Rationale: Suitable for handling high-dimensional data and non-linear relationships.

e) K-Nearest Neighbors (KNN)

- Rationale: is effective for capturing local patterns and relationships in the data based on proximity in feature space.

Model Evaluation Metrics

Metrics Used:

1. Mean Absolute Error (MAE)
2. Mean Squared Error (MSE)
3. Root Mean Squared Error (RMSE)
4. R² (R-squared)

What It Is:

1

MAE
measure the squared mean difference between predicted and actual values.

2

MSE
measures the average squared difference between the predicted values and the actual values. It calculates how far off predictions are from the true values.

3

RMSE
measures the square root of the average squared differences between the predicted values and the actual values. (more interpretable than MSE)

4

R²
measures the variance in the dependent variable that is predictable from the independent variables. It evaluates how well the model fits the data.

Why It Was Chosen:

MAE offers a simplified understanding of predictive accuracy in specific contexts. MSE is useful for optimizing models during training, such as minimizing loss. RMSE provides a clear sense of prediction accuracy in real-world terms by using the same units as the target variable. R² indicates the proportion of variance captured by the model, making it helpful for comparing models.

What It Reveals:

1

MAE

Higher MSA indicates better prediction accuracy.
lower MSA reflects less accuracy.

2

MSE

Lower MSE shows better performance.
higher MSE indicates larger prediction errors.

3

RMSE

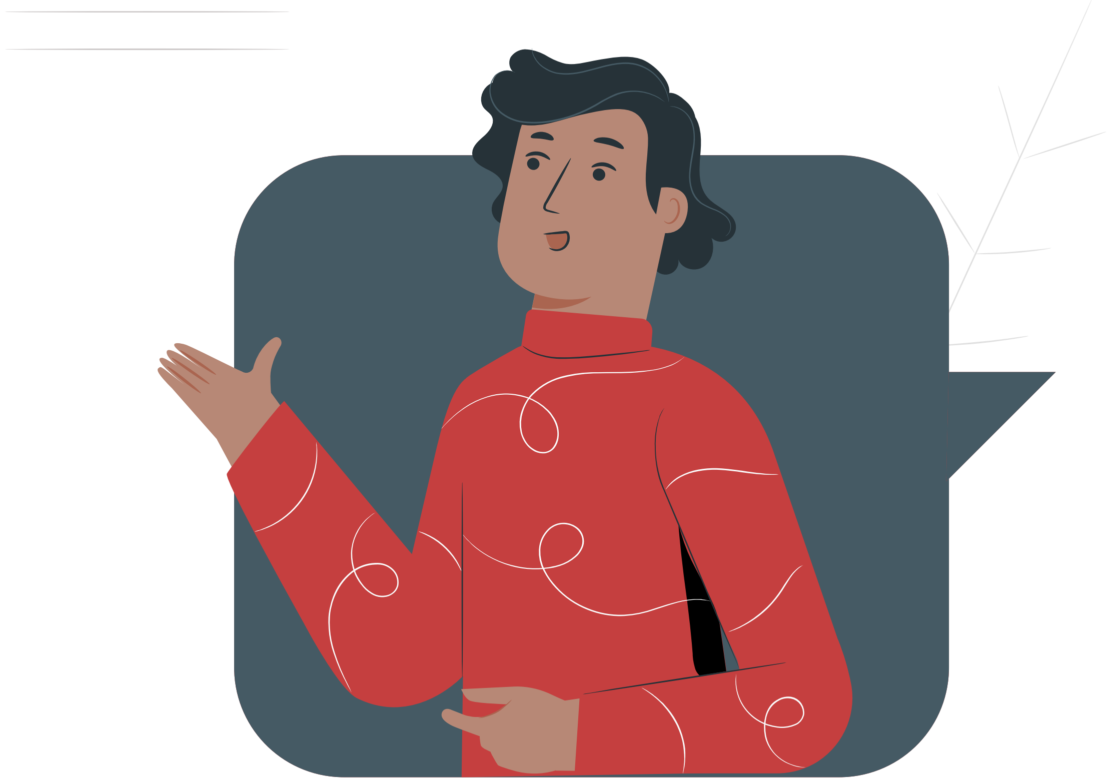
Lower RMSE reveals better real-world prediction accuracy.
higher RMSE shows greater discrepancies.

4

R²

Higher R² means the model captures more variance (good fit).
lower R² indicates poor fit.

Technical Hurdles and Solutions



Technical Hurdles and Solutions

Hurdle 1: Survey Design

- **Challenge**: Formulating precise and reliable questions to measure aspects of mental health such as Self-Esteem, Social Anxiety, Insomnia, FOMO, and Attention Span.

- **Solution:** Consulting mental health experts to ensure the validity and accuracy of the survey questions.

Hurdle 2: Arabic Survey and Translation

- **Challenge**: Translating responses from Arabic to English resulted in errors and loss of meaning.

- **Solution:** Applying data cleaning techniques using Pandas to handle unclear or missing responses and performing preprocessing steps to ensure consistency.

Technical Hurdles and Solutions

Hurdle 3: Data Collection

- **Challenge**: Reaching a diverse sample representing different age groups and regions.
- **Solution:** Distributing the survey widely across social media platforms to maximize participation.

Hurdle 4: Open-Ended Response Analysis

- **Challenge**: Analyzing open-text responses with varying lengths and meanings, which made classification and interpretation complex.
- **Solution:**
  - Using VADER Sentiment Analysis from the NLTK library to classify responses into emotional tones.
  - Applying thematic analysis to categorize responses into key themes.
  - Utilizing Regex (re library) to clean and preprocess text.

Technical Hurdles and Solutions

Hurdle 5: Categorical Variable Encoding

- **Challenge**: Converting categorical data (e.g., age, regions, favorite apps) into numerical formats for model compatibility.
- **Solution:** Using One-Hot Encoding from Pandas and removing one category per group to avoid multicollinearity issues.

Hurdle 6: Calculating the Mental Health Index (HMI)

- **Challenge**: Combining multiple mental health dimensions into a single index.
- **Solution:** Assigning appropriate weights to each dimension based on expert recommendations and normalizing values using z-scores with Scikit-learn’s StandardScaler.

Technical Hurdles and Solutions

Hurdle 7: Poor Model Performance

- **Challenge**: Initial models showed poor performance due to noisy data or weak relationships between independent variables and the dependent variable.

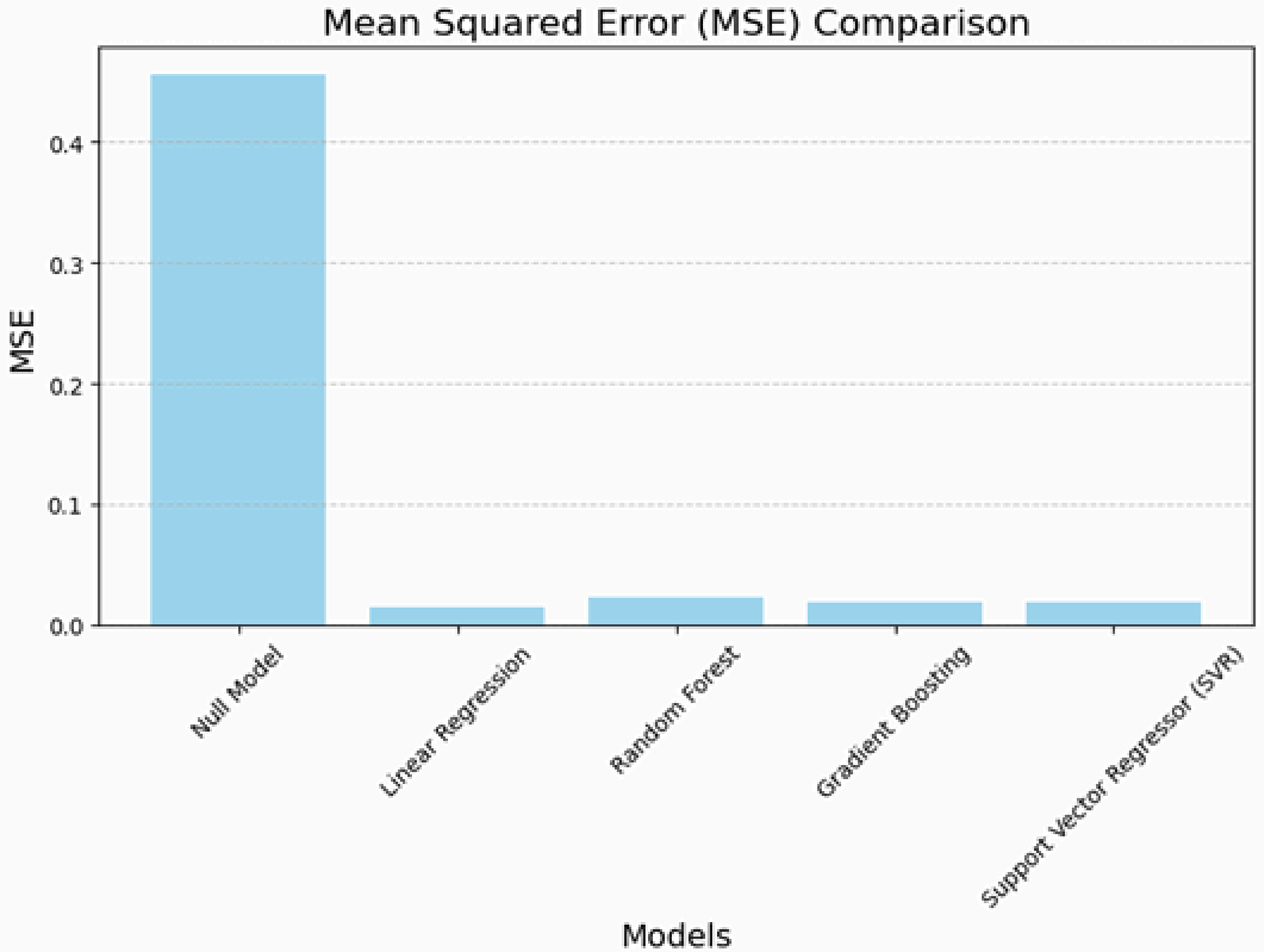
- **Solution:** Conducting correlation analysis using Pandas to identify features most relevant to the Mental Health Index (HMI), removing weakly correlated features, and improving data quality for better model accuracy.

.

Modeling Efforts and Results

Modeling Efforts and Results

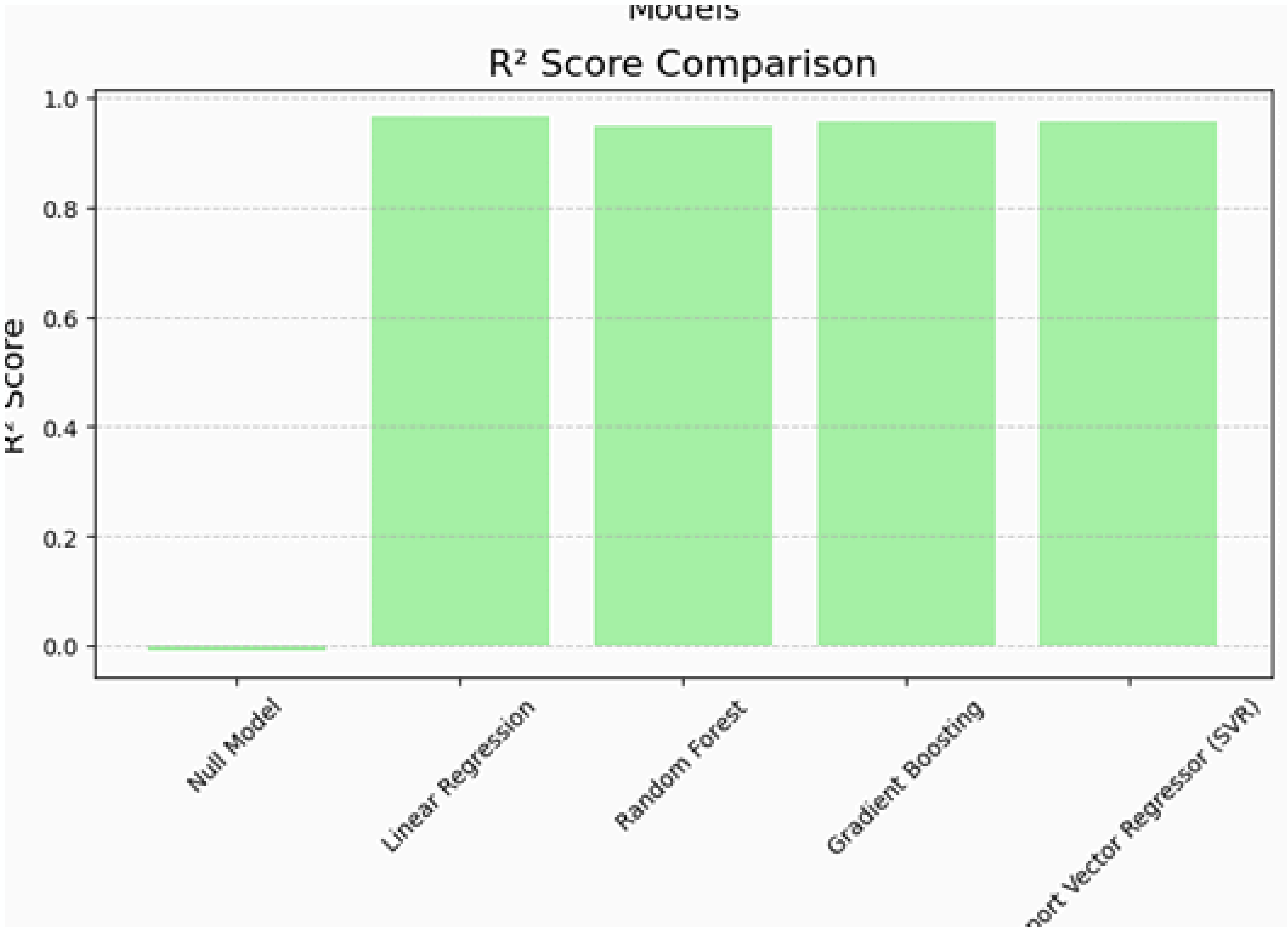
Performance Visualization for **Primary Data**



Linear Regression =0.014992 had a slightly lower MSE

Modeling Efforts and Results

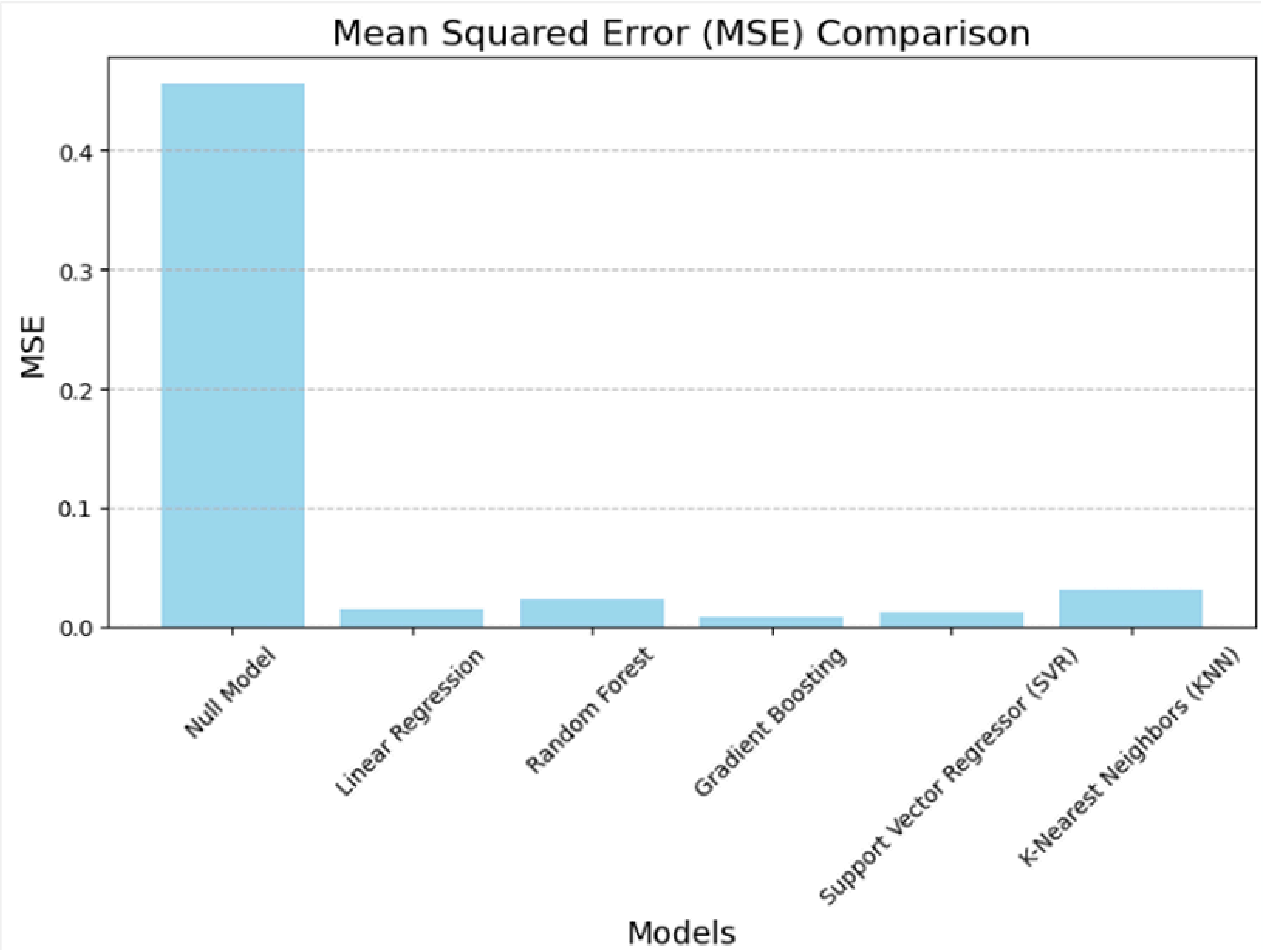
Performance Visualization for **Primary Data**



Linear Regression = 0.966787 achieved the highest R^2 score

Modeling Efforts and Results

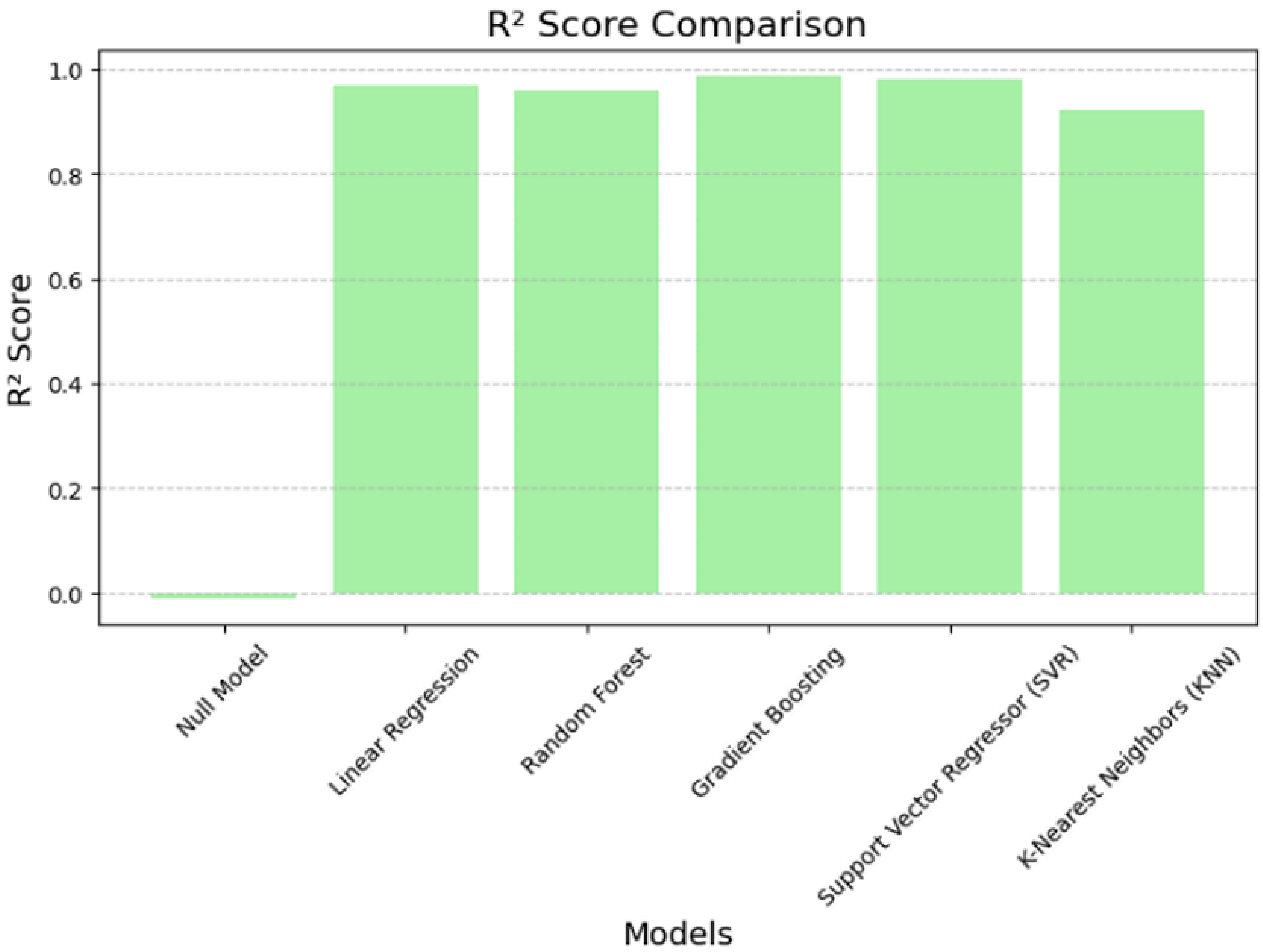
Performance Visualization for **Secondary Data**



Gradient Boosting = 0.007362 the lowest MSE, indicating the best accuracy.

Modeling Efforts and Results

Performance Visualization for **Secondary Data**



Gradient Boosting = 0.986705 achieves the highest R^2.

Modeling Efforts and Results

Best performer with the highest accuracy for MHI prediction.

Primary	Secondary
Linear Regression	Gradient Boosting

Findings on Large Language Models (LLMs)

Findings on Large Language Models (LLMs)

Large Language Models (LLMs) are AI systems based on deep learning that understand, analyze, and generate human language.

Role in Analysis:

- Supported efficient processing of textual responses and thematic analysis.

Insights:

Strengths:

1- Automated sentiment scoring

2- hypothesis generation.

Challenges:.

1- Computational cost

2- potential biases in pre-trained data

Recommendations and Future Directions

Recommendations and Future Directions

Key Recommendations:

- Develop personalized interventions for time management and digital well-being.

- Promote awareness campaigns on healthy social media use.

Recommendations and Future Directions

Model Improvements:

- Explore advanced feature engineering for richer insights

.

- Combine predictions using ensemble methods for robustness.
- Investigate deep learning models (e.g., RNNs, Transformers) for behavioral and textual analysis.

Recommendations and Future Directions

Expansion Opportunities:

1. Broaden participant demographics for diverse perspectives.
2. Conduct longitudinal studies to track long-term effects.
3. Incorporate multimodal data sources (e.g., engagement metrics) for comprehensive analysis.

Conclusion

Conclusion

This study highlights the complex relationship between social media use and mental health. By integrating primary and secondary datasets, we validated key findings and identified key behavioral impacts. Despite challenges, feature refinement and modeling yielded valuable insights.

Conclusion

Correlation analysis was crucial in improving the Mental Health Index (MHI) model's performance. Initially, irrelevant features caused overfitting and poor accuracy. By selecting only relevant features, the model's performance significantly improved. Linear Regression achieved the highest R^2 (96.68%) and lowest error metrics, while Gradient Boosting and SVR were strong alternatives. Random Forest performed the weakest due to its sensitivity to noise. Overall, correlation analysis helped enhance model accuracy and identify key mental health drivers.

THANKS