King Saud University
College of Computer and Information Sciences
Information Technology Department

جامعة الملك سعود
King Saud University
كلية علوم الحاسب والمعلومات
قسم تقنية المعلومات

Term-1 1446 H

IT 362 – phase 2
Logbook for secondary dataset

Supervised by
Dr. Khulood Alyahya

First Term 1446 H

Exploratory Data Analysis (EDA): Secondary dataset

1. **Metadata Review:**

Source:

- Owner: The dataset is provided on Kaggle by Souvik Ahmed (owner) and edited by Muhesena Nasiha Syeda.
- Context: It was gathered as part of a university Statistics course project with a primary focus on investigating the correlation between social media usage and mental well-being.

Date Collected:

- The dataset was last updated a year ago, although no precise date of collection (start or end) is provided.
- As it is part of a university project, it is likely that the data was collected within the past 1-2 years, making it relatively recent.

Collection Method:

- The data was gathered through **surveys**, where participants self-reported their social media habits and mental health status.

Reliability of the Method:

- **Response Bias**: Since the data is self-reported, it may suffer from inaccuracies, such as participants underreporting or overreporting behaviors and feelings.
- **Selection Bias**: Given that the participants were likely **university students**, the findings might not be representative of the general population. Thus, caution is needed when generalizing the results beyond this group.

Despite potential biases, **survey-based data** is a widely accepted method in mental health research and can still provide valuable insights if the survey design is robust and the analysis is thoughtful.

2. **Bias Awareness in Secondary Data:**

In analyzing the secondary dataset from Kaggle, which investigates the relationship between social media usage and mental health, it's important to recognize several potential biases that could affect the accuracy of conclusions drawn. These biases can influence the validity of findings, especially when comparing this dataset to your primary data.

1. Sampling Bias:

- **Source**: The dataset includes 481 responses, potentially from a narrow population.
- **Concern**: If the survey was distributed via specific channels, it may over-represent **active social media users** and under-represent non-users or individuals with limited access to social media. The data may also over-represent **younger, tech-savvy users**, leading to **demographic bias**.

2. Response Bias:

- **Source**: Survey respondents may provide answers that align with socially accepted norms.
- **Concern**: Respondents may **underreport** mental health concerns, especially related to **depression or anxiety**, due to fear of stigma, leading to biased or incomplete insights into the negative impacts of social media.

3. Question Design Bias:

- **Source**: The design of the survey questions could skew responses by emphasizing certain experiences.
- **Concern**: If the questions focus heavily on **negative aspects** of social media, such as distraction or anxiety, respondents may overlook **positive or neutral experiences**.

4. Demographic Imbalance:

- **Source**: The dataset records demographic variables like age, gender, and occupation, but these groups may be imbalanced.
- **Concern**: Over-representation of specific groups (e.g., younger individuals or certain occupations) could skew results. A **gender imbalance**, with fewer responses from non-binary individuals, could also limit insights into diverse gender perspectives.

5. Cognitive Bias:

- **Source**: Respondents may interpret survey questions differently based on personal experiences.
- **Concern**: Social media users with different motives (e.g., professional vs. personal use) may perceive the impact of social media differently, leading to **inconsistent interpretations** across responses**.**

6. Omission Bias:

- **Source**: The dataset may omit relevant factors influencing the relationship between social media and mental health.

- **Concern**: Missing details on **specific social media platforms** used, **time spent** on them, or **motives for use** (e.g., socializing, work) could hide important nuances in how social media affects mental health.

7. Historical Bias:

- **Source**: The dataset captures data at a specific time without accounting for changing social media trends.
- **Concern**: Evolving trends in both social media and mental health awareness, such as the rise of new platforms, may not be reflected, leading to outdated conclusions.

8. Temporal Bias:

- **Source**: The survey captures responses from a single point in time.
- **Concern**: Without longitudinal data, it is difficult to account for **fluctuating social media habits** or **mental health trends** over time. This limits the ability to assess the **long-term effects** of social media usage on mental health.

Data Collection Methods

- **Dataset**: The dataset titled **"Social Media and Mental Health"** was collected from **Kaggle** on **April 18, 2022**.
- **Source**: Kaggle - Social Media and Mental Health
- **Objective**: The dataset was used as a secondary dataset to explore mental health patterns related to social media usage. This analysis aimed to compare findings with a primary dataset to identify trends and differences.
- **Sample Size**: Initially, the dataset contained **481 records** and **21 features**.

Processing and Cleaning Tasks

Initial Dataset Review

- **Data Types**:
  - **Numerical Features**: Age, frequency of social media usage, etc.
  - **Categorical Features**: Gender, occupation, social media activity status.
  - **Ordinal Features**: Likert scale responses (e.g., 1 to 5 scale for distraction, comparison to others).
- **Numerical Features of Interest**: Age, distraction frequency, social media usage without purpose.

- **Categorical Features**: Gender, occupation, affiliated organization type.

Handling Missing Data

- **Observation**: The column **"What type of organizations are you affiliated with?"** had **30 missing entries (6.23%)**.
- **Decision**: Dropped the rows containing missing data.
- **Rationale**: Given the small percentage (6.23%), removing rows was a better option than imputation to avoid introducing bias or altering the dataset.

Dealing with Unwanted Entries in Gender

- **Observation**: The **Gender** column had unwanted entries apart from "Male" and "Female."
- **Decision**: Removed rows with invalid or missing gender entries.
- **Rationale**: Ensuring a consistent focus on the two major gender categories simplified the analysis and improved the clarity of gender-related results**.**

Outlier Detection and Removal

- **Target Feature**: **Age**.
- **Issue**: There were unrealistic values in the **Age** column, especially values above **75 years**, which were unlikely given the study context (social media usage).
- **Method**: Applied **Interquartile Range (IQR) analysis** to identify and remove outliers.
- **Decision**: Removed all records with age outliers.
- **Rationale**: Extreme values in the age column could distort statistical results and affect model performance**.**

Normalization of Numerical Features

- **Target Features**: Likert scale features (e.g., distraction, validation seeking, comparison to others).
- **Method**: Applied **decimal scaling** to normalize all numerical values (on a 1-5 scale) to a **0-1 range**.
- **Rationale**: Normalization ensured comparability across features, especially when conducting analyses like regression or correlation, without one feature dominating others.

Data Integrity Check

- **Action**: Performed a complete data integrity check post-cleaning to ensure:
  - No missing values in any column.
  - No inconsistencies or invalid entries remained.
- **Rationale**: This ensured that the dataset was ready for analysis, with no gaps, duplicated, or corrupted entries.

Decisions Made and Rationale

Decision 1: Dropping Missing Data

- **Why**: The missing data accounted for a small portion of the dataset.
- **Rationale**: Removing rows ensured a clean dataset without introducing bias from imputation methods, maintaining overall data integrity.

Decision 2: Removing Gender Anomalies

- **Why**: The dataset included unwanted values in the Gender column.
- **Rationale**: By removing these anomalous records, the data became easier to manage and analyze, focusing on male and female respondents.

Decision 3: Outlier Removal in Age

- **Why**: Unrealistic age values, especially those exceeding 75 years, were likely errors.
- **Rationale**: Outlier removal ensures the dataset reflects realistic social media usage patterns, without extreme values distorting analysis.

Decision 4: Applying Normalization

- **Why**: Likert scale features (e.g., distraction, comparison) had varying ranges.
- **Rationale**: Normalizing these features ensured that all variables were comparable, eliminating the risk of certain features skewing the results.

Challenges Faced and How They Were Overcome

Challenge 1: Identifying Outliers in Age

- **Issue**: Extreme values in age raised concerns about their validity.
- **Solution**: Used **IQR analysis** to statistically detect and remove outliers without arbitrary assumptions.
- **Outcome**: The cleaned data now accurately reflected the expected age distribution for social media users.

Challenge 2: Handling Skewed Distributions

- **Issue**: Several features (e.g., Age, frequency of distraction by social media) exhibited right-skewed distributions.
- **Solution**: Considered using **log transformation** or **Box-Cox transformation** for skewed distributions to normalize the data and improve interpretability.
- **Outcome**: Skewness was reduced, improving the dataset's suitability for further analysis.

Challenge 3: Managing Discrete vs. Continuous Features

- **Issue**: The dataset contained both **ordinal** (Likert scale) and **continuous** (Age) features.
- **Solution**: Treated categorical and ordinal features appropriately without normalization, while applying scaling only to continuous features.
- **Outcome**: This approach preserved the integrity of categorical data, maintaining relationships within the dataset.

Challenge 4: Avoiding Bias Through Imputation

- **Issue**: Handling missing data in a way that didn't distort patterns or introduce bias.
- **Solution**: Chose to **drop missing values** rather than impute, avoiding potential bias.
- **Outcome**: Ensured dataset quality was preserved without altering inherent distributions or relationships.

Logbook Entry: EDA Process

1. Exploratory Data Analysis (EDA) Objectives

- Understand general data distribution and key trends.
- Identify relationships between mental health attributes and social media use.
- Investigate correlations between ordinal/discrete features using appropriate methods.

2. Tools and Libraries Used

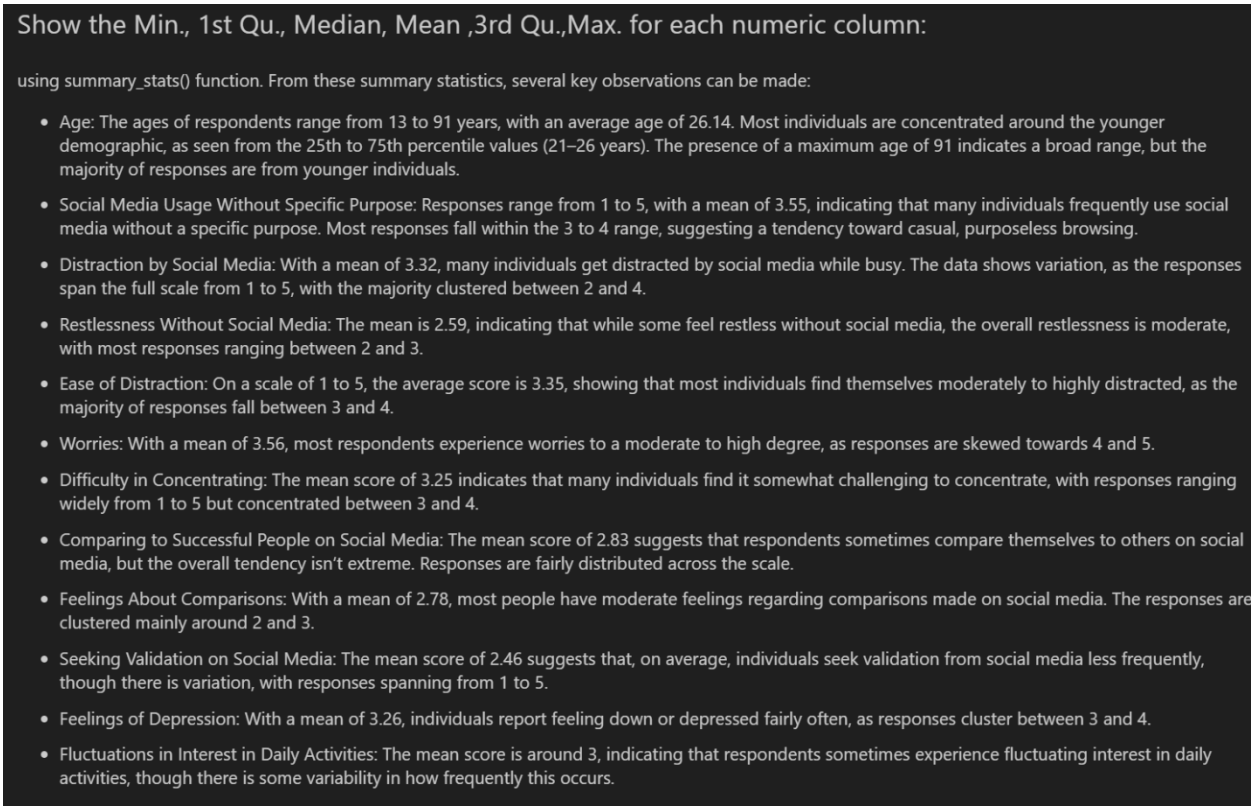
- **Pandas**: For data manipulation and analysis.
- **NumPy**: For numerical operations.
- **Matplotlib** and **Seaborn**: For visualizations, including distribution plots and heatmaps.
- **SciPy**: For statistical tests (e.g., Spearman correlation).

3. Types of Analysis Performed

Summary Statistics

- **Objective**: Calculate basic statistical measures (e.g., Min., Median, Mean, Max.) to understand key trends in numerical columns.

- **Outcome**:



Variance Analysis

- **Objective**: Measure the variability within each feature to understand the spread of responses.
- **Outcome**: Variance was moderate to high across most features, with **Age** having a particularly high variance (98.31), indicating a wide age distribution.

Distribution Visualization

- **Objective**: Identify skewness in the data, particularly for features like age, distraction frequency, and usage without purpose.
- **Outcome**: Age and distraction frequency were **right-skewed**. Log or Box-Cox transformations may be applied in further modeling phases.

Correlation Matrix (Spearman Method)

- **Objective**: Compute the correlation between ordinal features to assess relationships.

- **Outcome**: The **Spearman correlation method** was used to handle ordinal features like distraction, comparison to others, and depression.

Heatmap Visualization

- **Objective**: Visualize the correlation matrix to easily identify patterns and relationships between features.
- **Outcome**: The heatmap helped highlight significant correlations, such as between **distraction** and **difficulty concentrating**.

4. Interesting Findings and Anomalies

Skewed Distributions

- **Age**: Concentrated around younger age groups, with a few older outliers.
- **Distraction by Social Media**: Most respondents report high levels of distraction, clustering toward the higher end of the scale.

Significant Correlations

- **Distraction and Difficulty Concentrating**: A positive correlation was identified between how distracted individuals were by social media and their difficulty in concentrating.
- **Social Media Comparison and Feelings of Depression**: Moderate positive correlation suggests that those who compare themselves to others on social media more often tend to report higher feelings of depression.

Variance in Sleep Issues

- **Sleep Issues**: High variance suggests significant variation in how sleep problems are distributed across the population, possibly related to other mental health factors.