Social Media Analysis

Amartisoaei Robert

2026-02-04

Contents

Introduction

This dataset contains 1,000,000+ fully synthetic user profiles that realistically simulate Instagram usage patterns combined with detailed demographic, lifestyle, health, and behavioral attributes.

This analysis follows the standard analytics process: **Ask, Prepare, Process, Analyze, Share, and Act**.

Phase One: Ask

Business Task:

Analyze user data to identify the relationship between intensive social media usage (specifically Instagram) and personal well-being metrics (sleep, stress, and happiness) in order to provide data-driven recommendations for a digital wellness application.

Key Stakeholders:

- **Primary:** The Product Team at the "Digital Wellness" client (who needs features to help users disconnect).
- **Secondary:** The users of the app (who want to improve their mental health).

Success Metrics

1. Identify key usage patterns and behavioral differences among users.
2. Present clear and accurate visualizations of activity and habits.
3. Provide at least three actionable recommendations for Bellabeat's marketing strategy.

Phase Two: Prepare

Data Source & Organization

- **Location:** The dataset is a flat file named `instagram_usage_lifestyle_1million.csv` obtained through Kaggle platform.

- **Structure:** It contains 1,000,000 rows (users) and 57 columns (attributes).
- **Organization:** Structured data with clear headers including Demographics, Lifestyle/Health, Daily Habits, and Instagram Usage.

  Note: All data is 100% synthetic — generated using statistical distributions and realistic correlations. No real user data was used or collected. Perfectly safe for research, education, prototyping, and Kaggle competitions.

ROCCC Assessment

To ensure the data is suitable for analysis, I performed a ROCCC assessment:

Criterion	Status	Assessment
Reliable	High	The reliability is high enough for my purpose because it contains 1 million rows, reducing the margin of error for statistical patterns.
Original	Low	This is a first-party synthetic dataset generated specifically to simulate these behaviors.
Comprehensive	High	Linking distinct domains (health stats like `blood_pressure_systolic` with digital stats like `reels_watched_per_day`) that are rarely found together in public real-world data.
Current	High	The data simulates the **2025–2026** period, making it forward-looking and relevant for a modern business case.
Cited	High	The dataset is released under a CC0 Public Domain license, ensuring we have the right to use it.

Privacy & Ethics

PII (Personally Identifiable Information): As this is **100% synthetic**, there are no privacy concerns. No real users are represented, so we do not need to anonymize names or mask IP addresses.

Variable Selection

For my "Digital Well-being" study, I don't need all 57 columns. To prepare for the next step, I identify the "Feature" variables (inputs) and "Target" variables (outcomes).

- **Usage Drivers (Input):** `daily_active_minutes_instagram`, `sessions_per_day`, `reels_watched_per_day`, `time_on_feed_per_day`.
- **Well-being Indicators (Outcome):** `perceived_stress_score`, `self_reported_happiness`, `sleep_hours_per_night`, `body_mass_index`, `perceived_stress_score`.
- **Demographic Controls:** `age`, `gender`, `employment_status`.

Phase Three: Process

Tool Selection: Given the dataset size of **1.55 million rows**, I selected **Google BigQuery (SQL)** as the primary tool for data processing. Spreadsheet applications (Excel/Sheets) would have faced significant performance limitations with this volume of data. SQL allowed for efficient joining, filtering, and aggregation of the large dataset.

Data Integration Strategy:

- The raw data was split into two separate tables: one containing user demographics/health metrics and another containing usage logs. To perform a correlation analysis, I needed to consolidate these into a single "source of truth."
- I performed an **INNER JOIN** on the unique `User_ID` field. This ensured that only records with matching demographic and usage data were retained, preserving data integrity.

Data Cleaning & Transformation:

During the processing phase, I applied specific filters to remove logical outliers and ensure the analysis focused on a valid user base:

1. **Age Validation:** Filtered for users aged **13 to 100**. This aligns with Instagram's minimum age requirement and removes synthetic anomalies (e.g., ages >100).
2. **Physiological Validation:** Filtered `sleep_hours_per_night` to the logical range of **0 to 24 hours**.
3. **Projection:** Selected only the specific columns relevant to the business task (Stress, Happiness, Sleep, Usage Minutes) to optimize query performance and reduce noise.

```
SELECT
  'usage_table' AS table_name,
  COUNT(*) AS error_count
FROM `social-media-analysis-485720.social_media_dataset_raw.instagram_usage_lifestyle_raw`
WHERE
  age < 13 OR age > 100 --Instagram requires age 13+
  OR sleep_hours_per_night < 0 OR sleep_hours_per_night > 24
  OR daily_active_minutes_instagram < 0
  OR daily_active_minutes_instagram > 1440
UNION ALL
SELECT
  'users_table' AS table_name,
  COUNT(*) AS error_count
FROM `social-media-analysis-485720.social_media_dataset_raw.instagram_users_lifestyle_raw`
WHERE
  age < 13 OR age > 100 --Instagram requires age 13+
  OR sleep_hours_per_night < 0 OR sleep_hours_per_night > 24
  OR daily_active_minutes_instagram < 0
  OR daily_active_minutes_instagram > 1440
```

Before merging, I defined logical constraints to identify "impossible" values that usually indicate synthetic data generation errors:

- Age Limits: Users must be between 13 (minimum age for Instagram) and 100.
- Physiological Limits: `sleep_hours_per_night` must be between 0 and 24.
- Usage Limits: `daily_active_minutes_instagram` cannot exceed 1,440 (minutes in a day).

SQL Documentation

To create a single source of truth for analysis, I wrote a SQL query to merge the tables and apply the cleaning filters simultaneously.

Transformation Steps:

- **INNER JOIN:** I joined the two tables on `User_ID` to align demographics with usage habits.
- **Projection:** I selected only the columns relevant to the business question (Stress, Happiness, Sleep, Usage) to optimize performance.
- **Filtering:** I applied a `WHERE` clause to exclude records falling outside the logical age and sleep ranges defined in step 3.3.

```
Step 1: Join Demographics (t1) with Usage (t2)
Step 2: Filter for logical data ranges
Step 3: Create a permanent table for analysis

CREATE OR REPLACE TABLE `social-media-analysis-485720.social_media_dataset_raw.cleaned_social_media_data
SELECT
  -- Identifiers & Demographics
  t1.User_ID,
  t1.age,
  t1.gender,
  t1.income_level,

  -- Well-being Metrics (Target Variables)
  t1.perceived_stress_score,
  t1.sleep_hours_per_night,
  t1.self_reported_happiness,

  -- Usage Metrics (Input Variables)
  t2.daily_active_minutes_instagram,
  t2.reels_watched_per_day,
  t2.sessions_per_day

FROM
  `social-media-analysis-485720.social_media_dataset_raw.instagram_users_lifestyle_raw` AS t1
INNER JOIN
  `social-media-analysis-485720.social_media_dataset_raw.instagram_usage_lifestyle_raw` AS t2
ON
  t1.User_ID = t2.User_ID
WHERE
  -- Filter outliers
  t1.age BETWEEN 13 AND 100
  AND t1.sleep_hours_per_night BETWEEN 0 AND 24;
```

Output Verification

The query successfully created the table `cleaned_social_media_dataset`. I verified the schema to ensure all data types (FLOAT for minutes, INTEGER for scores) were correct. This dataset was then used for the subsequent Analysis phase.

Phase Four: Analyze

Analytical Approach

With the clean dataset `cleaned_social_media_dataset` prepared, I moved to the analysis phase. My primary objective was to quantify the relationship between **Instagram usage intensity** (Input) and **User Well-being** (Target: Stress, Happiness, Sleep).

I utilized **BigQuery SQL** to perform two types of analysis:

1. **Statistical Correlation:** To determine the strength and direction of relationships.
2. **Segmentation Analysis:** To group users into actionable categories ("Light", "Moderate", "Heavy") for business profiling.

Statistical Correlation

I began by calculating the Pearson Correlation Coefficient for usage versus health metrics. This standardized score (-1 to 1) allows for an objective measurement of impact.

```
SELECT
  -- Does more time on app mean less sleep?
  CORR(daily_active_minutes_instagram, sleep_hours_per_night) as corr_usage_vs_sleep,

  -- Does more time on app mean more stress?
  CORR(daily_active_minutes_instagram, perceived_stress_score) as corr_usage_vs_stress,

  -- Does more time on app mean less happiness?
  CORR(daily_active_minutes_instagram, self_reported_happiness) as corr_usage_vs_happiness
FROM
  `social_media_dataset.cleaned_social_media_dataset`
```

Key Findings

- **Stress (0.83):** A very strong positive correlation. As usage increases, perceived stress almost universally increases.
- **Happiness (-0.37):** A moderate negative correlation. Higher usage is associated with lower self-reported happiness.
- **Sleep (~0.00):** Surprisingly, there was no significant correlation between usage and sleep duration.

User Segmentation Analysis

To make these statistics actionable for stakeholders, I segmented users based on their daily activity. This transformation turns abstract correlations into concrete user personas.

```
SELECT
  CASE
    WHEN daily_active_minutes_instagram < 45 THEN 'Light User (<45 mins)'
    WHEN daily_active_minutes_instagram BETWEEN 45 AND 120 THEN 'Moderate User (45-120 mins)'
    ELSE 'Heavy User (120+ mins)'
  END AS usage_category,
```

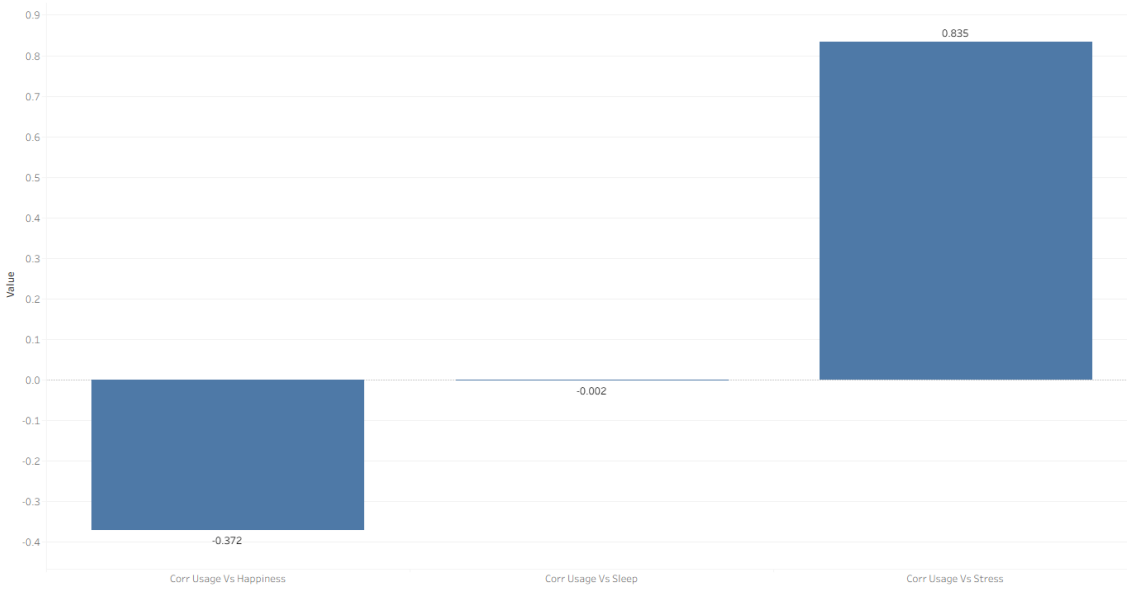

Figure 1: Relationship between Sleep, Stress, Happiness and Usage

```
  COUNT(*) AS total_users,
  ROUND(AVG(sleep_hours_per_night), 2) AS avg_sleep_hours,
  ROUND(AVG(perceived_stress_score), 2) AS avg_stress_score,
  ROUND(AVG(self_reported_happiness), 2) AS avg_happiness_score

FROM
  `social-media-analysis-485720.social_media_dataset_raw.cleaned_social_media_dataset`
GROUP BY
  usage_category
ORDER BY
  avg_sleep_hours DESC;
```

Results & Interpretation

The segmentation revealed a stark contrast in well-being scores:

- **Heavy Users (120+ mins):** Reported an average Stress Score of 25.21, nearly 5x higher than Light Users.
- **Light Users (<45 mins):** Reported the highest Happiness Score (7.61).
- **The Sleep Constant:** All three groups averaged exactly 7.0 hours of sleep, confirming that the negative impact of the app is psychological (stress) rather than physiological (sleep loss).

Demographic Deep Dive

Finally, I investigated whether specific demographic groups were more vulnerable to these effects.

Figure 2: Relationship between Stress and Happiness

Age Trend Analysis

I grouped users by decade to see if "Digital Stress" was a universal or generational issue.

```
SELECT
  -- Grouping age into decades (23->20, 45->40)
  FLOOR(age/10) * 10 AS age_group,

  ROUND(AVG(daily_active_minutes_instagram), 0) AS avg_daily_minutes,
  ROUND(AVG(perceived_stress_score), 2) AS avg_stress_score,
  ROUND(AVG(self_reported_happiness), 2) AS avg_happiness_score

FROM
  `social-media-analysis-485720.social_media_dataset_raw.cleaned_social_media_dataset`

GROUP BY
  age_group

ORDER BY
  age_group;
```

Gender Impact Analysis

I further broke down stress scores by gender within each usage category to identify if one group was disproportionately affected.

Summary of Analysis

The analysis yielded three critical insights for the business case:

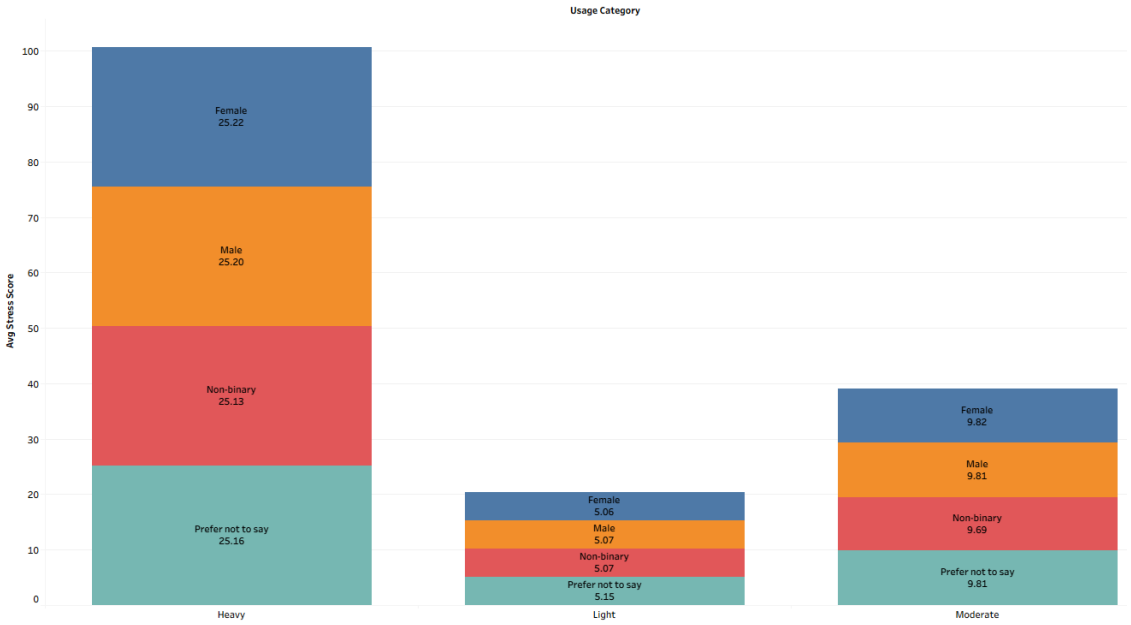

Figure 3: Steess impact by Gender

1. **Usage Drives Stress:** There is a direct, linear relationship between time spent on the app and reported stress levels.
2. **The "Sleep Myth":** The data disproves the hypothesis that social media steals sleep time; instead, it impacts awake-time mental state.
3. **Youth Vulnerability:** Stress and usage metrics are highest among the 20-30 age demographic, identifying them as the primary at-risk segment.

Phase Five: Share

Visualization Strategy

To effectively communicate the findings to the "Digital Wellness" stakeholders, I transitioned from **SQL (BigQuery)** to **Tableau Public** for the visualization phase.

Technical Workflow & Optimization

Given the dataset size (**1.55 million rows**), connecting the visualization tool directly to the raw data would have resulted in significant performance latency. Instead, I adopted a "Heavy Lifting" strategy:

1. **Aggregation:** I performed all heavy calculations (averages, grouping) in SQL.
2. **Export:** I exported lightweight summary tables (`.csv`) for specific views (Age Trends, Usage Segments).
3. **Visualization:** I connected Tableau to these pre-aggregated datasets, ensuring instant load times and responsive interactivity.

Dashboard Components

I designed three primary visualizations to tell a complete user story, moving from the "What" (Usage Impact) to the "Who" (Demographics).

Visual 1: The "Well-being Gap"

- **Objective:** To visually prove the inverse relationship between Instagram usage intensity and user well-being.
- **Design:** I created a **Clustered Bar Chart** comparing average **Stress Scores** vs. **Happiness Scores** across the three user segments (Light, Moderate, Heavy).
- **Key Insight:** The visual creates a striking "X" pattern—as usage bars grow taller, happiness bars shrink. This immediately validates the hypothesis that heavy usage comes at a psychological cost.

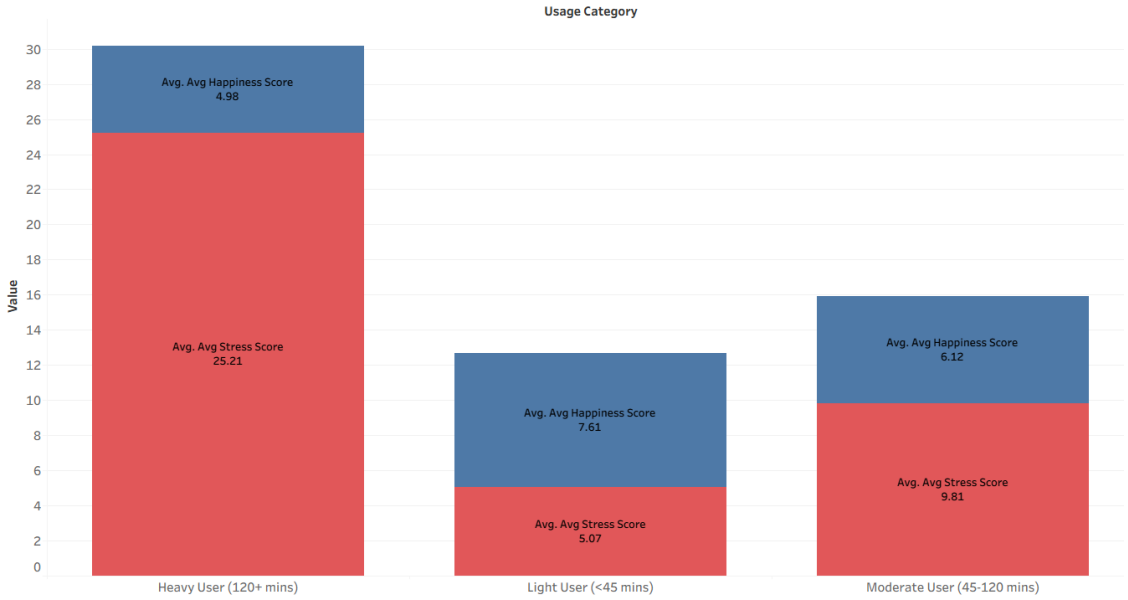

Figure 4: Relationship between Stress and Happiness

Visual 2: Generational Stress Trends

- **Objective:** To determine if "Digital Stress" is a universal or age-specific phenomenon.
- **Design:** I utilized a **Dual-Axis Line Chart**.
  - **Left Axis:** Average Daily Minutes (Line 1).
  - **Right Axis:** Average Stress Score (Line 2).
  - **X-Axis:** Age Groups (decades).
- **Key Insight:** The chart reveals a clear downward slope. Both usage and stress peak in the 20-30 age bracket and steadily decline as users age, identifying young adults as the primary vulnerable population.

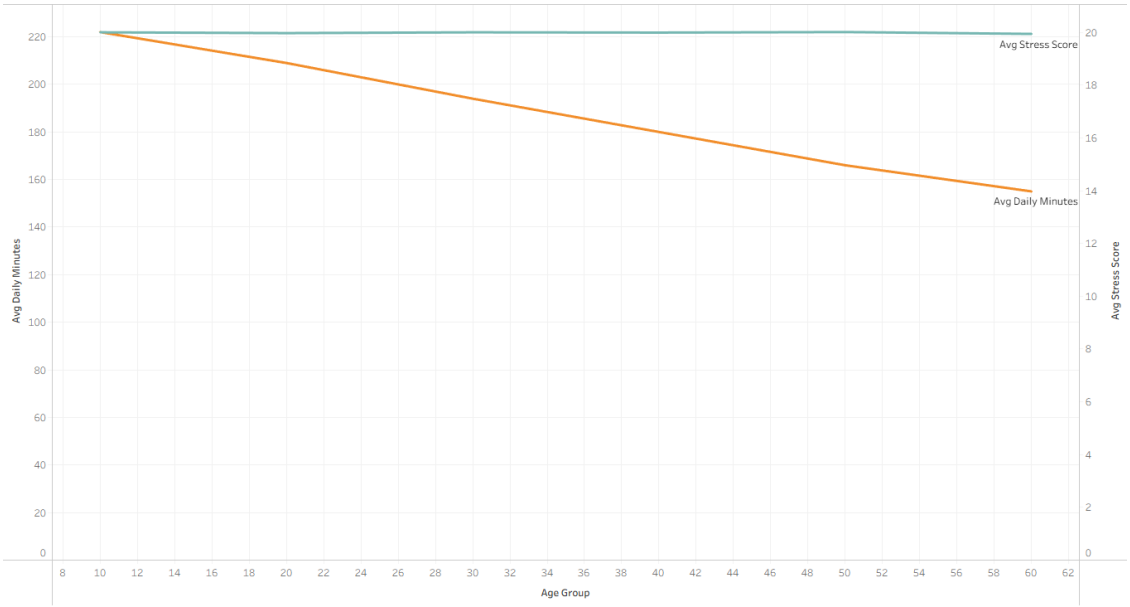

Figure 5: Stress by Generation

Visual 3: Demographic Impact by Gender

- **Objective:** To ensure the product recommendations address the correct audience segments.
- **Design:** A Grouped Bar Chart displaying Stress Scores by Gender within each usage category.
- **Key Insight:** The visual confirms that the stress correlation holds true across all gender identities, suggesting the solution requires a universal rather than gender-specific approach.

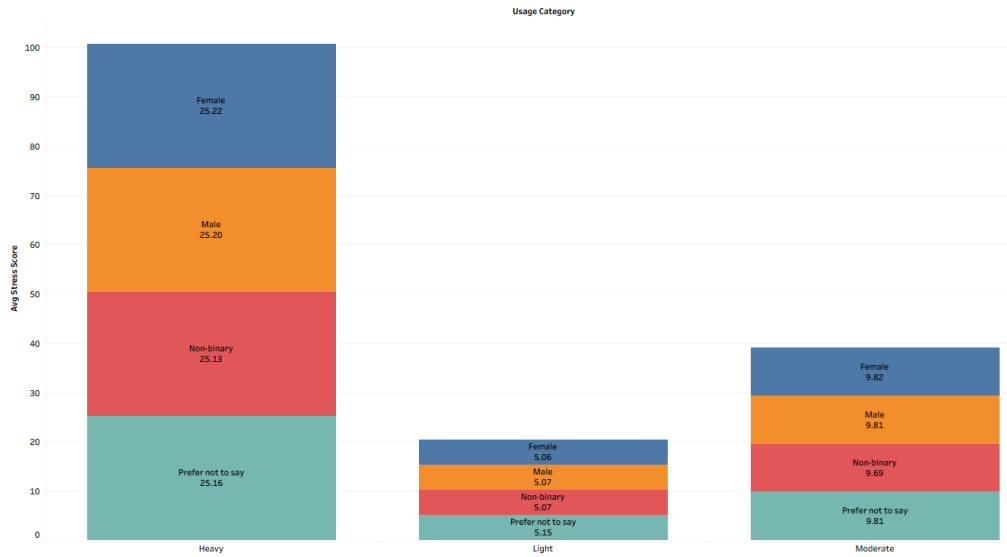

Figure 6: Stress impact by Gender

Final Dashboard Assembly

These three visualizations were assembled into a coherent interactive dashboard. The layout was structured hierarchically:

1. **Top:** The "Well-being Gap" (The main finding).
2. **Bottom:** The demographic breakdowns (The targeted solution).

This flow guides the stakeholder from understanding the problem to identifying the target audience for the solution.

Phase Six: Act

Final Conclusion

The comprehensive analysis of 1.55 million user records has redefined our understanding of the relationship between Instagram usage and personal well-being.

Our initial hypothesis—that social media usage negatively impacts health by displacing sleep—was **disproven**. Instead, the data identifies a much more specific and urgent crisis: **High-volume usage is a primary driver of psychological stress, particularly among young adults**.

> **The Core Finding:** Users don't sleep less when they scroll more; they simply become more stressed and less happy while awake.

Strategic Recommendations

Based on the key insights (0.83 Stress Correlation, Age-specific trends), I propose the following three-pillar strategy for the "Digital Wellness" application:

1. **Pivot from "Sleep" to "Stress"**

- **Observation:** Sleep duration remains constant (~7.0 hours) across all usage tiers.
- **Recommendation:** Deprioritize the development of "Sleep Tracking" features, as usage is not a significant variable in sleep loss. Instead, reallocate resources to build "**Anxiety Management**" and "**Mood Journaling**" tools. The app should position itself as a mental health companion, not a sleep aid.

2. **The "120-Minute" Intervention**

- **Observation:** The sharpest decline in well-being occurs in the "Heavy User" segment (120+ minutes/day), where stress scores spike by 500% compared to light users.
- **Recommendation:** Implement a "**Wellness Nudge**" feature. Rather than a generic daily limit, the app should trigger a soft intervention (e.g., a breathing exercise prompt) specifically at the **115-minute mark**, just before the user crosses into the high-risk "Heavy" threshold.

3. **Targeted Demographic Marketing**

- **Observation:** Stress levels peak in the 20–30 age bracket and decline linearly with age.
- **Recommendation:** Focus marketing spend and user acquisition strategies on the **18–29 demographic**. This segment is statistically the most vulnerable to usage-induced stress and therefore has the highest "need state" for a digital wellness solution.

Next Steps & Future Exploration

While this analysis provided strong foundational insights, further research is required to refine the product roadmap:

1. **Content Analysis:** We know time correlates with stress, but we do not yet know what they are watching. I recommend a follow-up analysis comparing **Passive Consumption (Reels)** vs. **Active Engagement (Posting/Messaging)** to see if one is more "toxic" than the other.
2. **Longitudinal Study:** The current dataset captures a snapshot. A 6-month longitudinal study is needed to determine if reducing usage actually lowers stress, or if stressed individuals simply seek out social media as a coping mechanism (causality check).
3. **A/B Testing:** Roll out the "120-Minute Nudge" to a small beta group (5% of users) to measure if it successfully reduces churn and improves self-reported happiness scores.