

# Project Part 1

## 1. A short summary of

### a. Your interests

I am passionate about leveraging data to drive innovation and create efficiencies, particularly at the intersection of engineering and business. My core interests include natural language processing, computer vision, and geospatial analysis, focusing on impactful applications in healthcare, logistics, energy, and infrastructure. I am committed to exploring uncharted territories within these fields to enact meaningful change.

### b. The reasons why you choose your current degree and major

I am pursuing a Master's in Applied Data Science at the University of Southern California, a choice motivated by my belief in the transformative potential of AI. This degree is an ideal foundation for developing expertise in Data Engineering, Machine Learning, and MLOps, areas critical to translating data insights into strategic decision-making and deploying scalable solutions. My hands-on experience with startups in building robust data and ML pipelines further supports my educational journey, as the program equips me with advanced tools to create reliable, efficient, and impactful data systems.

### c. The reasons why you decided to take this class,

I chose DSCI 517 to better understand how statistics can reveal patterns in human behavior and people's choices, offering insights essential for user-focused data science. Human behavior is complex and evolves over time, making understanding it challenging and rewarding. This course aligns with my goal of integrating psychological insights into data analysis by exploring how statistical methods apply to human-centered studies. Through these skills, I aim to approach data as more than numbers, using it to reflect the nuances of human behavior and decision-making that drive meaningful, impactful research.

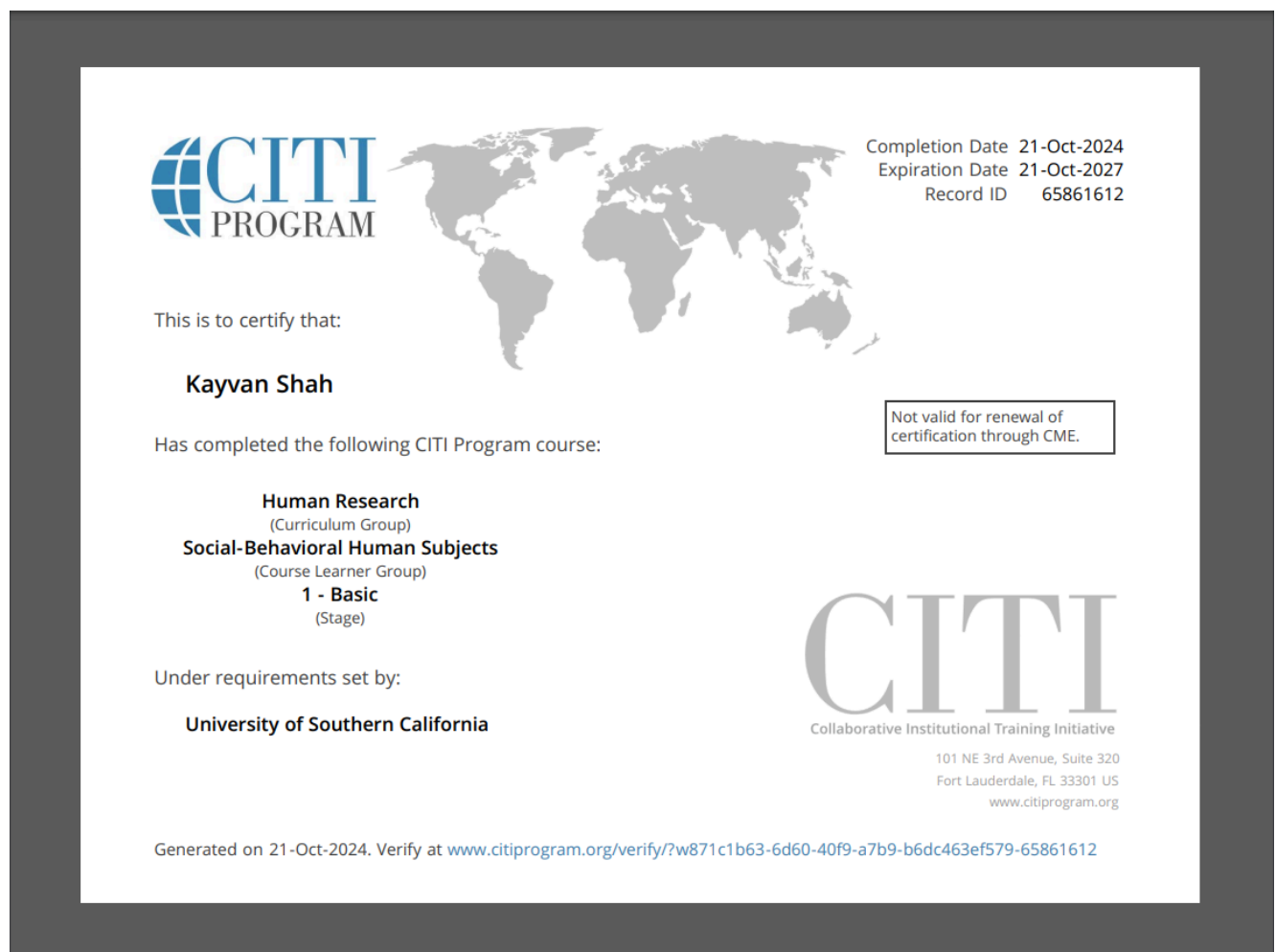
### d. Your personal ambitions to change the world, and

I aspire to bridge the gap between machine learning and real-world applications, focusing on data governance, quality, and community-driven insights. By gathering and understanding data, doing groundwork with local communities, and engaging in conversations to address real issues, I aim to deliver personalized and scalable solutions—from local to national impact. Strategic data insights, combined with solid engineering practices, can drive decisions that profoundly benefit communities through healthcare optimization, environmental sustainability, or infrastructure development. I aim to harness data and AI to solve meaningful challenges, enhancing access to information and boosting resilience and efficiency in essential systems.

- e. The reasons why you are interested in the topic you have chosen for your project.

This project explores how everyday habits—like caffeine, exercise, and sleep—affect productivity, particularly for students balancing heavy academic demands. By examining variables such as caffeine intake, sleep quality, and exercise frequency, I aim to understand whether these routines align with improved focus and productivity. Since I'm observing natural behavior rather than intervening, this approach allows for a straightforward look at how lifestyle choices link to productivity. Analyses will reveal any predictive relationships between these habits and productivity, shedding light on which routines may be most impactful. I hope these findings highlight simple, effective changes for students seeking to enhance their focus and well-being.

- f. Show me a screenshot of your CITI certification for human subjects research.



## 2. Sketch out the plan for the user study that you will conduct this term, including details such as:

### a. What variables are you going to collect?

#### i. **Caffeine Intake:**

1. Frequency of Consumption(Never, Rarely, Occasionally, Frequently, daily)
2. Timing (Morning, Afternoon, Evening, Late-Night, Never)

#### ii. **Physical Activity:**

1. Frequency(Scale 0-7)
2. Timing(Morning, Afternoon, Evening, Late-Night)
3. Variations in Energy levels(Yes, No, Somewhat, Unsure)

#### iii. **Sleep Quality:**

1. Hours of sleep per night
2. Consistency of sleep schedule (Consistent, Somewhat Consistent, Somewhat in-consistent, Very in-consistent)

#### iv. **Daily Productivity:**

1. Self-reported productivity rating (scale from 1 to 5)
2. Productivity Consistency (Consistent, Somewhat Consistent, Neutral, Somewhat in-consistent, Very in-consistent)
3. Variation in Productivity (Morning, Afternoon, Evening, Late-Night, Constant)

### b. What design is your study (experimental vs. correlational; if experimental, what factors are between subjects vs. within subjects)?

**Study Design:** Correlational Study

**Justification:**

- i. It's correlational as there is no active manipulation of variables, and it does not involve random assignment. Also, the participants' daily routine is observed as it occurs naturally.
- ii. Our study aims to examine the relationship between daily habits (caffeine intake, physical activity, sleep quality) and productivity.
- iii. We aim to identify trends and associations between these variables without manipulating them. Participants will self-report their habits and productivity over time, and we will analyze naturally occurring variations to see if they correlate.

### c. Given those answers, out of those variables, which are your IV(s) and DV(s)?

#### i. **Independent Variables:**

##### 1. **Caffeine Intake:**

- a. Frequency of Consumption(Never, Rarely, Occasionally, Frequently, daily)

- b. Timing (Morning, Afternoon, Evening, Late-Night, Never)

**2. Physical Activity:**

- a. Frequency(Scale 0-7)
- b. Timing(Morning, Afternoon, Evening, Late-Night)
- c. Variations in Energy levels(Yes, No, Somewhat, Unsure)

**3. Sleep Quality:**

- a. Hours of sleep per night
- b. Consistency of sleep schedule (Consistent, Somewhat Consistent, Somewhat in-consistent, Very in-consistent)

**ii. Dependent Variables:**

**1. Daily Productivity:**

- a. Self-reported productivity rating (scale from 1 to 5)
- b. Productivity Consistency (Consistent, Somewhat Consistent, Neutral, Somewhat in-consistent, Very in-consistent)
- c. Variation in Productivity (Morning, Afternoon, Evening, Late-Night, Constant)

- d. What are the operational definitions going to be for your IV(s) and DV(s)? (ie how are you going to measure or manipulate the variables)?

Our study is correlational. So, operational definitions for each independent variable (IV) and dependent variable (DV) will be measured. Our data will be collected through self-reported surveys.

**i. Independent Variables (IVs):**

These variables relate to everyday habits, and we will ask participants to report on them using specific questions designed to capture the required data.

**1. Caffeine Intake:**

**a. Frequency of Consumption:**

- i. Operational Definition: Self-reported frequency of caffeine consumption.
- ii. Measurement: Participants will choose one of the following options (Never, Rarely, Occasionally, Frequently, daily)

**b. Timing of Consumption:**

- i. Operational Definition: Self-reported time of day caffeine is consumed.
- ii. Measurement: Participants will indicate the time(s) they typically consume caffeine (Morning, Afternoon, Evening, Late Night, Never)

**2. Physical Activity:**

**a. Frequency of Activity:**

- i. Operational Definition: Number of days per week participants engage in physical activity.
- ii. Measurement: Participants will report the number of days they are physically active in a typical week, rated on a scale from 0 (no activity) to 7 (daily activity).

**b. Timing of Activity:**

- i. Operational Definition: Self-reported time of day participants engage in physical activity.

- ii. Measurement: Participants will indicate when they typically exercise (Morning, Afternoon, Evening, Late Night)
  - c. **Variations in Energy Levels:**
    - i. Operational Definition: Self-reported variation in energy levels during the day.
    - ii. Measurement: Participants will choose from the following options(Yes, No, Somewhat, Unsure)
- 3. **Sleep Quality:**
  - a. **Hours of Sleep per Night:**
    - i. Operational Definition: Average hours of sleep participants get per night.
    - ii. Measurement: Participants will report the typical number of hours they sleep each night.
  - b. **Consistency of Sleep Schedule:**
    - i. Operational Definition: Self-reported consistency of participants' sleep schedules.
    - ii. Measurement: Participants will rate their sleep schedule consistency(Consistent, Somewhat Consistent, Somewhat Inconsistent, Very Inconsistent)
- ii. **Dependent Variables (DVs):**
  - 1. **Daily Productivity:**
    - a. **Self-Reported Productivity Rating:**
      - i. Operational Definition: Participants' subjective rating of their daily productivity.
      - ii. Measurement: Rated on a scale from 1 (very unproductive) to 5 (very productive).
    - b. **Productivity Consistency:**
      - i. Operational Definition: How consistent participants feel their productivity is over time.
      - ii. Measurement: Rated with the following options Consistent, Somewhat Consistent, Neutral, Somewhat Inconsistent, Very Inconsistent)
    - c. **Variation in Productivity:**
      - i. Operational Definition: Self-reported variations in productivity throughout the day.
      - ii. Measurement: Participants will indicate when they feel most productive(Morning, Afternoon, Evening, Late-Night, Constant)
- e. What is your population? How are you going to get participants from that population? How many are you planning to recruit for the study?
  - i. The population for this study consists of university students who are likely to consume caffeine daily, engage in physical activity, and experience varying sleep habits. This demographic is ideal because they often manage their productivity in academic or work environments.
  - ii. Recruitment will target students from the University of Southern California (USC), where the study is being conducted. We will share the survey through Email, Student organization platforms, and word of mouth.
  - iii. A sample of 30-50 participants is a reasonable size for detecting patterns in correlational research while maintaining manageable data collection and analysis efforts for our study.

### 3. Sketch out your plan for analysis:

a. State your research question(s), and discuss how it could be answered by analyzing the data that you listed in the previous question. That is, affirm for me that your research question is answerable using the data you will collect.

i. **Research question:** *How do caffeine intake, physical activity, and sleep quality influence self-reported productivity levels among university students, both on a daily and weekly basis?*

ii. We will answer this question by analyzing the correlations between our independent variables and dependent variables (productivity). Our collected survey data will measure specific aspects of each habit and productivity metric, allowing us to examine relationships between these variables. We can determine which factors are strongly associated with productivity outcomes by comparing the relationships between different habits and productivity measures.

iii. **Data Collection Mapping:**

**Primary DVs from Survey:**

- Daily productivity rating (1-5 scale)
- Productivity consistency rating
- Time-of-day productivity variations

**Primary IVs from Survey:**

- Caffeine consumption frequency and timing
- Physical activity frequency (0-7 scale) and timing
- Sleep hours and consistency rating

b. Describe in your own words what kinds of analysis could be done with the data to answer each question. Be specific about what analysis -within null hypothesis significance testing- you would use and why.

Since this is a correlational study with multiple variables, we will employ several statistical analyses within null hypothesis significance testing:

i. **Multiple Regression Analysis - Combined IV Test**

1. **Purpose:** To examine the combined influence of caffeine intake, physical activity, and sleep quality on productivity ratings.

2. **Hypotheses:**

- a. Null Hypothesis (H0): The combination of caffeine intake, physical activity, and sleep quality does not significantly predict productivity ratings ( $R^2 = 0$ )
- b. Alternative Hypothesis (H1): At least one daily habit significantly predicts productivity ratings ( $R^2 > 0$ )

3. **Significance Level:**  $\alpha = 0.05$  (as per general trend)

4. **Independent Variables (IVs):**

- a. Caffeine Intake: Frequency of Caffeinated Beverage Consumption

- b. Physical Activity: Frequency of Physical Exercise per Week
- c. Sleep Quality: Average Sleep per Night and Consistency of Sleep Schedule
- 5. **Dependent Variable (DV):**
  - a. Overall Productivity Rating
- ii. **Pearson Correlation Analysis - Detect Continuous DV Effects over Productivity (positive, negative, or neutral)**
  - 1. **Purpose:** To determine the strength and direction of relationships between continuous variables such as xyz
  - 2. **Main Effect of Caffeine Intake:**
    - a. **H0:** There is no correlation between caffeine intake frequency and productivity ratings ( $r = 0$ )
    - b. **H1:** There is a significant correlation between caffeine intake frequency and productivity ratings ( $r \neq 0$ )
    - c. **IV:** Frequency of Caffeinated Beverage Consumption
    - d. **DV:** Overall Productivity Rating
  - 3. **Main Effect of Physical Activity:**
    - a. **H0:** There is no correlation between physical activity frequency and productivity ratings ( $r = 0$ )
    - b. **H1:** There is a significant correlation between physical activity frequency and productivity ratings ( $r \neq 0$ )
    - c. **IV:** Frequency of Physical Exercise per Week
    - d. **DV:** Overall Productivity Rating
  - 4. **Main Effect of Sleep Quality:**
    - a. **H0:** There is no correlation between hours of sleep and productivity ratings ( $r = 0$ )
    - b. **H1:** There is a significant correlation between hours of sleep and productivity ratings ( $r \neq 0$ )
    - c. **IV:** Average Sleep per Night
    - d. **DV:** Overall Productivity Rating
- iii. **Chi-Square Tests of Independence:**
  - 1. **Purpose:** To analyze relationships between categorical variables
  - 2. **Variables to be analyzed:**
    - a. Timing of habits (Morning/Afternoon/Evening/Late-Night) vs. Productivity variations
      - i. **IVs:** Timing of Caffeinated Beverage Consumption, Time of Day for Physical Activity
      - ii. **DV:** Productivity Consistency (e.g., Consistency in Productivity Throughout the Week)
    - b. **Consistency measures vs. Productivity consistency**
      - i. **IVs:** Consistency of Sleep Schedule, Energy Level Patterns
      - ii. **DV:** Consistency in Productivity Throughout the Week
  - 3. **Hypotheses:**
    - a. **H0:** No association exists between timing/consistency variables and productivity patterns

- b. H1: Significant associations exist between timing/consistency variables and productivity patterns

**iv. Weekly Analysis:**

**1. Purpose:** To examine weekly patterns and aggregate effects or to analyze changes across weeks

**2. Repeated Measures ANOVA:**

**a. Variables:**

- i. Weekly average productivity ratings  
DV: Overall Productivity Rating (assessed weekly)
- ii. Weekly patterns of caffeine consumption  
IV: Frequency of Caffeinated Beverage Consumption (assessed weekly)
- iii. Weekly physical activity frequency  
IV: Frequency of Physical Exercise per Week (assessed weekly)
- iv. Weekly sleep consistency  
IVs: Average Sleep per Night, Consistency of Sleep Schedule (assessed weekly)

**b. Hypotheses:**

- i. H0: No significant differences in productivity patterns across weeks
- ii. H1: Significant differences exist in productivity patterns across weeks

The analysis focuses on three key components:

- 1. Overall predictive relationships through multiple regression
- 2. Individual correlations between continuous variables
- 3. Associations between categorical timing and consistency measures

This comprehensive analysis approach will allow us to:

- Identify which daily habits have the strongest relationships with productivity
- Understand how the timing of different activities relates to productivity patterns
- Determine the role of consistency in habits and its relationship with consistent productivity
- Account for both continuous and categorical variables in our dataset

By examining these relationships through multiple statistical approaches, we can provide robust insights into how caffeine intake, physical activity, and sleep quality correlate with productivity among university students.