**Project Part 2**

**Samarth Saxena - saxenasa@usc.edu**

Q1. **Short description**: The proposed project aims to investigate the factors influencing participants' likelihood to suggest their favorite restaurant to others. The study involves collecting data on various variables to control for potential influences, such as cuisine preferences, social networks, and financial aspects. The research question focuses on whether offering a discount or opening a new location closer to the participant would increase the frequency of restaurant suggestions. The experimental design employs both between-subject and within-subject testing, randomly assigning participants to scenarios and measuring behavior before and after the scenarios. For Scenario 1 (discount offered), the average discount on the bill is tracked, while for Scenario 2 (new location opened), the average distance traveled is recorded. The dependent variable is the number of times the favorite restaurant is suggested. Operational definitions are provided for each variable to ensure consistency in data collection. The target population encompasses individuals of all age groups and genders, recruited from friends, family, coworkers, and extended networks. The analysis plan includes a 2x2 Mixed Analysis of Variance (ANOVA) to examine the effects of scenario type and time on the frequency of restaurant suggestions. Hypotheses are formulated for the main effects of scenario type and time, and a significance level of 0.05 is set for hypothesis testing. The study aims to shed light on whether situational changes influence participants' inclination to share their favorite dining spots, contributing valuable insights to understanding social dynamics and restaurant recommendations.

Q2. **Dataset Description**:

* **Independent Variables** (IVs):

1. **Scenario Type (Categorical)**: Discount Offered (Scenario 1), New Location Opened (Scenario 2)
2. **Manipulated Variables (Continuous)**: Average Discount on Bill (Scenario 1) - Percentage value, Average Distance (Scenario 2) - Distance in miles
3. **Control Variables (Continuous/Categorical)**: Cuisine (Categorical) - Type of food served by the favorite restaurant, Number of acquaintances the participant regularly interacts with (Continuous), Number of acquaintances that share the same interest (Continuous), Number of times the participant discusses dining options with acquaintances (Continuous), Average party size (Continuous), Monthly Income (Continuous), Monthly Expenses (Continuous)

* **Dependent Variable (DV)**: Number of times Favorite Restaurant is suggested (Continuous) - Count of how many times participants suggest their favorite restaurant as a dining destination.
* Categorical Variables: Scenario Type (Discount Offered, New Location Opened), Cuisine (Type of food served by the favorite restaurant)
* Continuous Variables: Average Discount on Bill (Percentage value), Average Distance (Distance in miles), Number of acquaintances the participant regularly interacts with, Number of acquaintances that share the same interest, Number of times the participant discusses dining options with acquaintances, Average party size, Monthly Income, Monthly Expenses, Number of times Favorite Restaurant is suggested

Q3. **Data Pre-processing**: In preparing our dataset for analysis, several pre-processing steps will be undertaken to ensure data integrity and reliability. For the independent variables (IVs), specific measures will be implemented for each scenario type. In data cleaning, entries will be checked to see if they fit with the defined categories, and unexpected values will be flagged. Imputation strategies will be considered for missing data, with categorical variables requiring no imputation. Further pre-processing steps, such as scaling for varying units, will be applied where necessary. Manipulated variables, including Average Discount on Bill and Average Distance, will undergo thorough data cleaning to confirm the validity of entries and potential imputation with mean or median for missing values. Control variables, such as Cuisine, will be checked for consistency in labeling. Continuous variables, like Number of Acquaintances, Monthly Income, Monthly Expenses, Number of Times the Participant Discusses Dining Options, and Average Party Size, will undergo data cleaning, imputation, and outlier checks. The dependent variable, Number of Times Favorite Restaurant is Suggested, will be subject to data cleaning, outlier checks, and no imputation as a continuous variable. General considerations, including missing data handling, outlier detection and treatment, consistency checks, coding and labeling uniformity, and data transformation and scaling, will be addressed for different variables as and when required. Throughout this process, detailed documentation will be maintained for transparency and reproducibility in subsequent analyses. Through these comprehensive pre-processing steps, the dataset's quality and reliability will be enhanced, laying a solid foundation for meaningful analysis and interpretation.

Q4. **Analysis**: We intend to use 2x2 Mixed ANOVA to analyze our data. We anticipate to have 3 test statistics:

* Scenario Interaction - Determine if there is a significant difference in restaurant mentions between Distance and Discount, regardless of time.
  + H0: There is no significant difference in restaurant suggestion count between Discount and New Location Scenarios.
  + H1: There is a significant difference in restaurant suggestion count between Discount and New Location Scenarios.
* Time Interaction - Determine if there is a significant difference in restaurant mentions before assigning a scenario, and after, regardless of scenario.
  + H0: There is no significant difference in restaurant suggestion count going from the Before timeframe to After in either scenario.
  + H1: There is a significant difference in restaurant suggestion count going from the Before timeframe to After in either scenario.
* Interaction Effect (Scenario x Time) - Determine whether there is a significant difference in restaurant mentions over time is different between the Distance scenario, and Discount scenario.
  + H0: There is no significant difference in restaurant suggestions going from the Before timeframe to After in neither the Discount nor New Location scenarios.
  + H1: There is a significant difference in restaurant suggestions going from the Before timeframe to After in either the Discount or New Location scenarios.

Feedback: Great work! **No changes made**

**Project Part 1**

**Samarth Saxena - saxenasa@usc.edu**

**1.** **Introduction:**

1. **Interests:** Academically, my interests lie in Data Science, especially NLP, and the recent boom of LLMs. I have always wanted to become a successful data scientist working towards a more significant cause directly impacting and improving society. Apart from this, I am hugely interested in sports and gaming. I was a state-level recognized gamer (esports) in my high school and university, and I was involved in multiple sports on a state level as well. I like to work out and love anything involving physical activity. Material things that give me happiness are- cars, shoes, and collector’s knives (you’d be surprised to find out how many people are into that!).

1. **Why I chose this degree and major:** I decided to pursue a master’s degree in data science because I wanted to upskill in this field, learn more, and work on myself to get to a stage in my career where I can significantly contribute to a cause or an organization. I remember sitting in front of my jupyter notebooks, playing with hyperparameters, and enjoying myself while uncovering and discovering the reasons behind the changes in a neural network’s performance when I make a particular change, and that’s how I knew this was the career for me.

1. **Why I took this class:** I have a vested interest in research, to the point where I might consider doing a PhD at the end of my graduate program. I have done multiple research studies, and I chose this course to improve my understanding of research on human subjects and learn more about it. Moreover, these data science topics piqued my interest from an early stage, which is why I felt compelled to take this course.

1. **Personal ambitions:** I desperately want to be part of something that creates a difference in the world. Be it natural impact, societal impact, or anything else, all I want is to be involved and know that my efforts are going towards something bigger than myself. Even when I apply to internships or jobs, I look at their vision pages to see what they are working on and their future goals. I do not want to be a part of an organization whose only purpose is to generate revenue and grow as a business. In this day and age, we need to be more conscious of our community, and I take that very seriously.

1. **The reason why I’m interested** in my topic is because it is close to my heart. My best friend, during my undergraduate program, eventually quit his studies to open a nationwide restaurant chain, and it did not go as planned. He surmounted a tremendous amount of debt, his restaurants failed, and I saw him go from being a joyful presence to becoming clinically depressed. I know I cannot bring back his dreams, but doing this research is my way of paying respect to him and his courageous efforts.
2. 

**2.** **Study Structure**

1. The variables we must collect will allow us to standardize the number of times the participant suggests their favorite restaurant. Some variables we need to control include:

● Cuisine: We want to know what type of food the participant’s favorite restaurant serves. This will help us understand if they have highly unique tastes or a more common palette.

● Number of acquaintances the participant regularly interacts with: We need to know if participants interact with many people to know if they can share their restaurant with other people.

● Number of acquaintances that share the same interest: Of these acquaintances, we also want to know how many are at least open to consuming the same cuisine the restaurant serves to know if any biases prevent the participant from sharing their restaurant.

● Number of times the participant discusses dining options with acquaintances: We should also know how often the participant shares the decision-making, which provides an opportunity to present their favorite restaurant.

● Average party size: There is the possibility that the participants require their favorite restaurant to be able to serve a group of people, so we should track if the participant typically dines by themselves or with others.

● Monthly Income: We can evaluate if the participant can dine at a high-end location that may be out of economic reach for the general population or opt for more affordable options.

● Monthly Expenses: By comparing how much the participant spends at their favorite restaurant to their monthly expenses, we can distinguish if they treat the restaurant as an occasional treat or a regular stop.

● Number of times Favorite Restaurant is suggested: Measuring this before assigning either scenario will help provide a baseline so we know if our scenarios affect the number of times the restaurant gets suggested.

For scenario 1, where a discount is offered, we will collect:

● Average Discount on Bill: How much, if any, discount is applied to the final bill? This will allow us to determine if the participant regularly visits the restaurant during a promotion. This should be a percentage value, and estimates are allowed.

For scenario 2, where a new location is opened,

● Average Distance: How far does the participant travel to reach the restaurant? They are allowed to self-define where the starting point for their travel is, i.e., whether they start from home, work, or elsewhere. This will be in miles, and estimates are allowed.

For every scenario, we have the same dependent variable:

● Number of times Favorite Restaurant is suggested: We will track how many times the participants offer their Favorite Restaurant as a place to dine. No matter how often the participant mentions the restaurant in a conversation, we will only count it as one occurrence when the participant decides on a place to dine. Estimates are allowed.

1. We will run an experimental design where we compare participants using between-subject and within-subject testing. Participants will be randomly assigned to one of two groups corresponding to our scenarios. This provides the between-subject testing to compare whether one scenario has a more significant effect. The within-subject aspect also comes from measuring the before and after of these scenarios: how much do participants suggest their restaurant in the current situation, and how much will they suggest after one of the scenarios? Because of the multifaceted design of this experiment, our design is between-subject and within-subject testing.

1. For our experiment, we will track different variables as independent and dependent variables depending on the scenario we present. For our first scenario, our independent variable is the average discount on their bills. For our second scenario, the independent variable is the average distance the participant travels to the restaurant, letting the participant decide where their trip originates. Our dependent variable stays the same in both scenarios: How often will the participant suggest their favorite restaurant? Will the number of suggestions increase if the restaurant offers a discount? What about if the restaurant opens a new location closer to the participant? We hope to answer these questions with our scenarios.

1. For our independent and dependent variables, we want to present operational definitions that allow us to capture the data we need in a way we can analyze it. For scenario 1, where a discount is offered, our operational variable is the Average Discount on the Bill, where the participants report how large of a discount they usually receive when dining at their favorite restaurant, expressed as a percentage. We will manipulate this variable by offering a theoretical percentage discount to lower the participant’s bill and increase the Discount on the Bill. For scenario 2, where a new location is opened, Average Distance is defined as the distance the participant travels to the restaurant, expressed in miles. Participants won’t have to specify the starting point, which allows for favorite restaurants near work, home, or any other location. We will manipulate this variable by presenting a theoretical new location for the participant that shortens the distance the participant has to travel. For both scenarios, we have the same dependent variable: the number of times Favorite Restaurant is suggested. We will ask participants how often they offered their favorite restaurant as a dining destination. This value is limited to once per dining consideration, which means multiple mentions in the same conversation or deliberation only count as one time. In contrast, multiple dining events will accumulate multiple occurrences. To give a concrete example, mentioning the participant’s favorite restaurant multiple times for a breakfast meal on a particular day would be counted once, but presenting the choice for different meals such as lunch, afternoon tea, dinner, and midnight snack would be counted four times.

1. The population we aim to collect from will be of all age groups, genders, and so on. We believe that regardless of socioeconomic status, everyone will have a favorite restaurant that they want to share with others. With this in mind, the data we collect will help us determine which method will result in the participants sharing their favorite restaurant more often. We will be recruiting from our friends, family, and coworkers. They are invited to share the study with their acquaintances as well. We aim to collect 30 responses from our target population.

**3.** **Analysis Structure**

1. Research Question: With respect to the participant’s favorite restaurant, would opening a closer location or offering a discount cause them to suggest their favorite restaurant more often?  
   We will answer our question by measuring the number of times the participant suggests their favorite restaurant to their acquaintances. We will assign participants to one of two scenarios and compare the number of suggestions to determine which scenario has the greater effect. The two scenarios we will assign participants will either be (Scenario 1), the restaurant providing them a discount, or (Scenario 2), the restaurant opens a location closer to the participant. By comparing the number of suggestions before (Before Time Point) and after (After Time Point) assigning these two scenarios, we should be able to determine which, if any, scenario increases the likelihood of the participant suggesting their favorite restaurant as a dining option the most.

1. By assigning participants to two different situations, Discounts vs New Locations, we can determine which method of these two would inspire more participants to mention their favorite restaurant. We would measure the independent variable before and after assigning a random scenario. We will employ a Mixed Analysis of Variance (ANOVA) design to address the research question, specifically a 2x2 Mixed ANOVA. This approach allows us to examine the effects of two independent variables, Scenario Type (Discount vs. New Location) and Time (Before vs. After), as well as their interaction on the dependent variable, the number of times the favorite restaurant is suggested.

## Hypotheses:

### Main Effect of Scenario Type:

Purpose: This will allow us to determine which scenario has a larger effect on the number of times the participant suggests their favorite restaurant.

Null Hypothesis (H0): There is no significant difference in the number of restaurant suggestions between the Discount and New Location scenarios.

Alternative Hypothesis (H1): There is a significant difference in the number of restaurant suggestions between the Discount and New Location scenarios.

### Main Effect of Time:

Purpose: This will allow us to determine if either scenario affects the number of times the participant suggests their favorite restaurant.

Null Hypothesis (H0): There is no significant difference in the number of restaurant suggestions between the Before and After time points.

Alternative Hypothesis (H1): There is a significant difference in the number of restaurant suggestions between the Before and After time points.

Significance Level:

We will set the significance level (alpha) at 0.05, indicating a 5% chance of Type I error. This threshold is commonly used in hypothesis testing.

Scenario Type (Discount vs. New Location):

This factor represents the experimental conditions for randomly assigning participants. It enables us to evaluate whether introducing a discount or opening a new location significantly impacts the frequency of restaurant suggestions.

Time (Before vs. After):

This within-subject factor accounts for the temporal aspect of our study. By measuring participant behavior before and after the implementation of the assigned scenario, we can assess any changes that occur over time.

Our analysis aims to provide insights into whether opening a closer location or offering a discount would lead participants to suggest their favorite restaurant more frequently. By examining the main effects and interaction, we can discern which scenario has a more pronounced impact on participant behavior.