Project Part 3

a. **Include (the same) short description describing the project that you proposed in part 1.**

- The proposed project aims to conduct a correlational user study to investigate the relationship between independent variables (LinkedIn Premium Usage, Networking Activity) and a dependent variable (job search success rates through LinkedIn).
- The study focuses on two key research questions: the impact of LinkedIn Premium subscription on job search success and the influence of network activity on job search success.
- The study seeks to analyze the correlation between these variables and uncover potential patterns that can guide individuals in optimizing their job search strategies on LinkedIn.

b. **State what kind of analysis you did, and what your result is for every "test statistic" that you conducted (ie a 2 way ANOVA has a test statistic for each of the two main effects, and one for the interaction, so what was the result of each). This needs to be written like a results section: no commentary, just the results in APA format.**

Linkedin Premium User or Not vs Job Success Rate - Independent Samples t-test:

An independent samples t-test revealed no significant difference in job success rates between Linkedin Premium users and regular users, $t(70) = -0.69$, $p = 0.50$. The mean success rate for Premium users ($M = 25.67$, $SD = 3.21$) did not significantly differ from regular users ($M = 28.34$, $SD = 2.89$).The effect size, as measured by Cohen's d, was found to be $d = 0.43$, suggesting a moderate effect.

Network Activity Score and Job Success Rate – Correlation:

A significant positive correlation was observed between Network Activity Score and Job Success Rate, $r(N=72) = 0.27$, $p = 0.024$ ($p < 0.05$). This indicates that individuals with higher Network Activity Scores tended to have higher job success rates.

c. **Did the results support your hypothesis or not? That is, could you reject the null hypothesis, and thereby show support for your hypothesis? What type of error might you be making (type I or II), and what is the chance of that error? How do you know (ie where/what did you do to find the chance of that error you might be making)?**

1. **For LinkedIn Premium User or Not vs Job Success Rate:**

Test Used: Independent Samples t-test

The results did not support this hypothesis. The p-value is insignificant, so we cannot reject the null hypothesis. The value of p=0.50(high) and from this, we can conclude that LinkedIn premium users do not significantly differ in their job success rate from regular users.

The potential type of error in this case would be TYPE **- II error** because we might have failed to reject a false null hypothesis. The chance of error is **84% (β)**. We obtained the chance of error in Type-II by finding the power using power analysis and **β = 1-power.**

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
As a LinkedIn user, how successful have you been in your job search? Please select the option that most appropriately describes your situation / case.	72	1	5	2.33	1.061
Valid N (listwise)	72				

2. For Network Activity Score and Job Success Rate

Test Used: Correlation

The results achieved from this test supported this hypothesis. The p-value is significant and hence we can reject the null hypothesis. The value of p=0.024(low) and from this, we can conclude that Active users are significantly more successful in getting a job than less active users.

The potential type of error in this case would be **TYPE - I error** because we might have concluded that there is a significance when actually there is none. The chance of error is **2.4%.** We obtained the chance of error in TYPE-I error which is equal to the **p-value**.
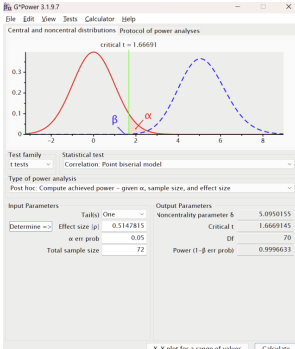
d. **If you did not need to do a power analysis for c above, conduct a power analysis now to see the likelihood you had of finding an effect as big as the one you found (also what was your effect size? – make sure you present Pearson's R or Cohen's D).**

We have performed power analysis for finding the first part of question c ie For LinkedIn Premium User or Not vs Job Success Rate.

Now , for the second part of question c ie For Network Activity Score and Job Success Rate:

We know that we have a TYPE-I Error and we do not need to find power analysis. However, if we had to conduct a power analysis to see the likelihood of finding an effect as big as the one we found before, then the power analysis is as mentioned below:

We calculate the effect size from Pearson's R which is reported as 0.265. The effect size comes to 0.514. Power calculated in G*Power shows that it comes 0.99966 (~1) and hence, beta is approximately zero. This is as expected since this is a type 1 error and the probability of making a type 2 error show be minimum.

e. **Write a short discussion section like what you might see in a paper. What are the *big* take away conclusions from your results? I.e., what do they mean? Interpret them for readers. This is the commentary that you left out of b above. <u>What are the limitations of your study?</u> *For example: Did you have any confounds? Was your sample representative? Other limitations? Also mention from d above if your power was too low (If your power was high but you still didn't find an effect, what does this mean?).* Think critically about your own design and study.**

The big takeaway conclusions from our results are as follows:

1. There is no significant association between having an active LinkedIn Premium subscription and a job search success rate through Linkedin. Based on the statistical analysis conducted, paying for a Premium subscription on LinkedIn does not appear to be a key factor influencing one's success in finding a job through the platform.Users with a LinkedIn Premium subscription and those without seem to have similar levels of success in their job searches and investing in LinkedIn premium might not provide them a significant advantage in their job success journey..

2. There is a significant correlation between Network Activity Score and Job Success rate through Linkedin. Active engagement on LinkedIn significantly influences job success rates. Individuals who demonstrate higher network activity levels, more connections, increased daily usage hours, a higher number of profile views, and frequent posting, tend to experience greater success in job searches. So, investing more productive time on LinkedIn boosts your chances of finding success in your job search.

The limitations and confounds of our study are as follows:

1. Our study sample consisted of University of Southern California (USC) student population (22-26 year olds). There is a lack of representation in our sample from experienced individuals who are seeking new employment opportunities hinders the generalizability of our findings.

2. Additionally, limiting the study majorly to USC means that the job-seeking landscape can significantly differ across regions and institutions, and our findings may not be universally applicable.

3. A confounding factor is the reliance on personal networks for job searches. Some people might have personal connections through working relatives etc. Personal connections in job opportunities can create bias because the results may vary from those achieved through formal applications.

4. Another confounding factor is that employers might hold preconceived notions about the capabilities or experience levels of younger individuals, affecting their chances in the job market. Since the majority of our sample are younger individuals with little to no experience, this age-related bias adds an additional layer of complexity to the interpretation of our findings

Our power that we found for the analysis between Linkedin Premium User or Not and Job Success rate is low (power 0.16 and β =0.84) . A possible reason could be the lack of a large enough sample size. Our sample size was small with N=72 participants only.

f. Include as an appendix that includes both previous assignments (part 1 and 2), which you can update based on comments you've received and/or new understanding making sure you highlight all changes in red font. If you don't have any changes you can make, then put a note in red at the top of the appendix indicating that. As a second appendix, also provide a screenshot of G*power showing one of your power analyses. These appendices dont count towards your page "limit"

Appendix -1:

G*Power analysis of the following hypothesis:

Null Hypothesis (H0): There is no significant association between having an active LinkedIn Premium subscription and job search success.

Alternative Hypothesis (H1): There is a significant association between having an active LinkedIn Premium subscription and job search success.
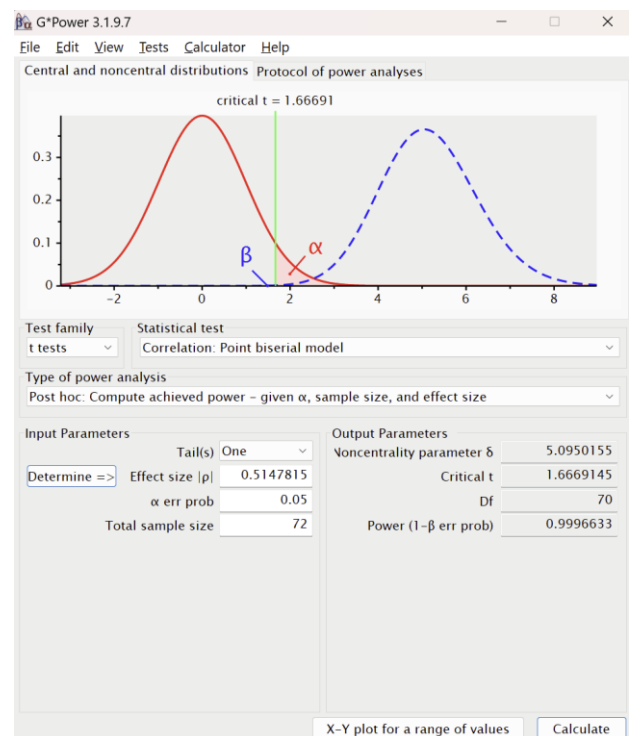
Appendix -2:

Project Part – 2:

g. **Include short description outlining the project that you proposed in part 1.**

- The proposed project aims to conduct a correlational user study to investigate the relationship between independent variables (LinkedIn Premium Usage, Networking Activity) and a dependent variable (job search success rates through LinkedIn).

- The study focuses on two key research questions: the impact of LinkedIn Premium subscription on job search success and the influence of network activity on job search success.
- The study seeks to analyze the correlation between these variables and uncover potential patterns that can guide individuals in optimizing their job search strategies on LinkedIn.

h. Describe the dataset you will have. What are the variables? Which are IVs and which are DVs? Within IVs and DVs, which are categorical and which are continuous?

The data would be extracted from the Google Form and organized into an Excel sheet as a table. Each row will serve as an index for individual participants, and the columns will contain their responses to the Google Form questions. The columns for the dataset would be LinkedIn premium user, Number of connections, Number of posts in the last 6 months, Number of profile views in the last 90 days, Daily Usage, and Job Search Success.

The variables for the study would be:
- LinkedIn premium user (Categorical): Either the subject is a LinkedIn Premium User or not.
- Network Activity (Continuous): Categorical variable that indicates if a user is an active user based on factors: approximate number of LinkedIn posts in the last 6 months, number of connections, daily usage in hours, number of profile views in the last 90 days.

    Formula:
    - **Number of Connections**: <100 = 1, 100-250 = 2 , 250-500 = 3, 500-1000 = 4 , >1000 = 5
    - **Approximate Number of LinkedIn Posts in the last 6 months**: 0-1 = 1, 2-3 = 2, 4-6 = 3, 7-10 = 4, >10 = 5
    - **Daily Usage in Hours**: <0.5 = 1, 0.5-1 = 2 , 1-2 = 3 , 2-4 = 4 , >5 = 5
    - **Number of Profile Views in the last 90 days**: 0-20 = 1, 20-60 = 2, 60-120 = 3, 120-200 = 4, >200 = 5
    - **Activity Score**: [0.15 * Approximate Number of LinkedIn Posts in the last 6 months] + [0.35 * Daily Usage in hours] + [0.3 * Number of Connections] + [0.2 * Number of Profile Views in the last 90 days]

- **Job Search Success** (Continuous): Likert scale -> No response, Connection, Referral, Interview, Landed job which corresponds to numbers from 1-5.

- IV: LinkedIn Premium User, Network Activity

- DV: Job Search Success
    ### i. Describe the data pre-processing steps that you believe you will need to do. What data cleaning? Will you impute values from missing data? If so, how? What other pre-processing might you need to do, if any?

Data Preprocessing Steps:

- Missing Values: If the column having the missing value is numerical (e.g. Number of connections), impute the missing value by the mean value of the column. Else if it is categorical, impute the missing value by mode of the column.

- Data Cleaning: Removing ',' or any other symbols from numerical factors like number of connections. For example, clean '1,000' to '1000'.

- Other pre-processing:

- o LinkedIn Premium User: Convert Yes to 1, No to 0
- o Network Activity: Convert values falling in the specified range as follows.
  - Number of Connections: <100 = 1, 100-250 = 2, 250-500 = 3, 500-1000 = 4, >1000 = 5
  - Approximate Number of LinkedIn Posts in the last 6 months: 0-1 = 1, 2-3 = 2, 4-6 = 3, 7-10 = 4, >10 = 5
  - Daily Usage in hours: <0.5 = 1, 0.5-1 = 2, 1-2 = 3, 2-4 = 4 , >5 = 5
  - Number of Profile Views in the last 90 days: 0-20 = 1, 20-60 = 2, 60-120 = 3, 120-200 = 4, >200 = 5
  - Activity Score: [0.15 * Approximate Number of LinkedIn Posts in the last 6 months] + [0.35 * Daily Usage in hours] + [0.3 * Number of Connections] + [0.2 * Number of Profile Views in the last 90 days]

  Job search activity: There are 5 categories for job search activity namely No response, connections, referrals, interviews, and job offer. These 5 categories will be encoded as numbers from 1 to 5 which would in turn be used for the independent samples t-test and correlation test.

j. State what kind of analysis you intend to do, and what your hypothesis is for every "test statistic" that you will generate (i.e a 2 way ANOVA has a test statistic for each of the two main effects, and one for the interaction, so generate a hypothesis for each).

We will perform a correlational user study, and the research questions we define for the study are:

1. Is there a significant association between having an active LinkedIn Premium subscription and a job search success rate through LinkedIn?

For this question, we can used the independent samples t-test, as the IV (LinkedIn Premium User) is categorical and DV is continuous.

Null Hypothesis (H0): There is no significant association between having an active LinkedIn Premium subscription and job search success.

Alternative Hypothesis (H1): There is a significant association between having an active LinkedIn Premium subscription and job search success.

This test will help us to determine if there's a statistically significant association between LinkedIn Premium usage and job opportunities. The LinkedIn Premium Usage Independent Variable and Job Search Success dependent variable can answer this research question.

2. Is there an association between LinkedIn premium usage and users' networking activity?

Similarly, for this question, we can use correlation test, as both the IV (Networking Activity) and the DV (job search success rate) are continuous.

Null Hypothesis (H0): There is no significant association between job success rate and Networking Activity.

Alternative Hypothesis (H1): There is a significant association between job success rate and Networking Activity.

This test will help us to determine if there is an association between a user's networking activity and job search success rate. The Network Activity Independent Variable and Job Search Success dependent variable can answer this research question.

Project Part – 1 Appendix:

Overall Feedback

How exactly are you going to use that information to categorize people as active or inactive? We need to know here the exact cutoffs/formula to determine this. google forms is not recruiting. You collect data that way, but how will you get this google form link out to the USC student body for people to complete it?

<u>"Are LinkedIn Premium Users More Likely to Find Job Opportunities?"</u>

1. A short summary (1/2 page to 1 page) of:

    a. **Your interests:** I am passionate about data, and I have always enjoyed working with math, I found the field of Data Science creative and fascinating. Apart from academics, I enjoy cooking and exploring various cultures and cuisines.

    b. **The reasons why you choose your current degree and major:** My grandmother lost her eyesight due to glaucoma. While trying to help her, I came across OrCam MyEye, a product that assists the visually challenged using technologies like deep learning and computer vision. This encounter significantly kindled my interest in data science. The realization that data can be used to address critical issues with substantial social impact motivated me to pursue a master's program in Data Science.

    c. **The reasons why you decided to take this class:** I came across the concept of AB testing user design in multiple applications during my internship search and was curious about how statistical tests are used. I was also interested in understanding how a research study is designed, especially when the user data is so sensitive and confidential, and the ethical ways to conduct a study.

    d. **Your personal ambitions to change the world, and:** I want to work on services or products that directly impact society. Tackling problems that could either make life more convenient, solve critical issues in healthcare through research, or improve the environment through the use of data is something I am ambitious about.

    e. **The reasons why you are interested in the topic you have chosen for your project.** I have been on my job search journey for the last year in the USA and have utilized LinkedIn for it both premium and basic. Especially given the current job market, I feel that such a study could be helpful to analyze if there is something like purchasing a LinkedIn premium or being actively involved in networking on LinkedIn that could impact Job search quality.

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

1. **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?**

Ans:

- LinkedIn Premium Usage: This variable indicates whether a user is a LinkedIn Premium user or not.
- Networking Activity: Categorical variable which indicates if a user is an active user based on factors: approximate number of LinkedIn posts in the last 6 months, number of connections, daily usage in hours, and number of profile views in the last 90 days.

Preprocessing:

- Number of Connections: <100 = 1 , 100-250 = 2 , 250-500 = 3, 500-1000 = 4 , >1000 = 5
- Approximate Number of LinkedIn Posts in the last 6 months: 0-1 = 1 , 2-3 = 2 , 4-6 = 3 , 7-10 = 4 , >10 = 5
- Daily Usage in hours: <0.5 = 1 , 0.5-1 = 2 , 1-2 = 3 , 2-4 = 4 , >5 = 5
- Number of Profile Views in the last 90 days : 0-20 = 1 , 20-60 = 2 , 60-120 = 3 , 120-200 = 4 , >200 = 5
- Activity Score: [0.15 * Approximate Number of LinkedIn Posts in the last 6 months] + [0.35 * Daily Usage in hours] + [0.3 * Number of Connections] + [0.2 * Number of Profile Views in the last 90 days]
  - Active User if Activity Score >= median(Activity Score)
  - Inactive Users if Activity Score < median(Activity Score)

- Job Search Success: This variable indicates the job search success on a scale of 5.

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

Design of the Study: **Correlational**

In this case, we want to understand the relationships between LinkedIn Premium usage, networking activity, and job search success. There are no experimental manipulations, and participants' characteristics are observed as they naturally exist.

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

Ans:

Independent Variable (IV): LinkedIn Premium Usage, Networking Activity

Dependent Variable (DV): Job Search Success

   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)?**

Ans:

1. LinkedIn Premium Usage(Categorical): Either the subject is a LinkedIn Premium User or not.

2. Networking Activity (Categorical): Users can be categorized as either "active" or "inactive" based on various factors, including the number of connections, the extent of following, the frequency of posts, the post impressions, daily app usage, the quantity of InMails sent, and more.

2. Job Search Success (Categorical): Likert scale -> No response, Connection, Referral, Interview, Landed job

No response	connections	referrals	interviews	Job offers

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?

Ans:

Population: All LinkedIn users will form the population for this study.

Participants: We are collecting data from USC students who are LinkedIn users and individuals in our LinkedIn network using Google Forms.

We are planning to recruit approximately a minimum of 30 participants for our research.

We are collecting data from USC students who are LinkedIn users and individuals in our LinkedIn network using Google Forms. We will email the link to USC students in our classes as well as other usc students that we know to make sure that the link reaches to USC students. We are also going to post the link on our respective Linkedin page to reach out to people in our Linkedin network.

2. 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.

Ans. We will perform a correlational user study, and the research questions we define for the study are:

- Is there a significant correlation between having an active LinkedIn Premium subscription and a job search success rate through LinkedIn?

This question assesses whether there's a positive relationship between LinkedIn Premium usage and job opportunities. The LinkedIn Premium Usage Independent Variable and Job Search Success dependent variable can answer this research question.

- Is there a significant correlation between Network Activity and job search success rate through LinkedIn?

This question assesses whether there's a positive relationship between Network Activity and job opportunities. The Network Activity Independent Variable and Job Search Success dependent variable can answer this research question.

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.

Ans.

1. Is there a significant association between having an active LinkedIn Premium subscription and a job search success rate through LinkedIn?

For this question, we can use the Chi-Squared Test of Independence, as both the IV and DV are categorical.

Null Hypothesis (H0): There is no significant association between having an active LinkedIn Premium subscription and job search success.

Alternative Hypothesis (H1): There is a significant association between having an active LinkedIn Premium subscription and job search success.

This test will help us determine if there's a statistically significant association between these two categorical variables.

2. Is there a significant correlation between Network Activity and job search success rate through LinkedIn?

Similarly, for this question, we can use the Chi-Squared Test of Independence, as both the IV (Networking Activity) and DV (job search success rate) are categorical.

Null Hypothesis (H0): There is no significant association between job success rate and Networking Activity.

Alternative Hypothesis (H1): There is a significant association between job success rate and Networking Activity.

This test will help us to determine if there is an association between a user's networking activity and job search success rate.