Research Question 1:
Does increased time spent on social media predict higher levels of anxiety among Canadians?

Q55, Q66_1 - From var_names.csv

Null Hypothesis:

- There is no linear association between time spent on social media and feelings of anxiety in Canadians.
    - If the null hypothesis is supported, then the answer to the research question is no, increased time spent on social media predicts no difference in levels of anxiety among Canadians

Alternative Hypothesis:

- There is a positive linear association between time spent on social media and feelings of anxiety in Canadians.
    - If the null hypothesis is rejected and the alternative hypothesis is supported, then the answer to the research question is yes, increased time spent on social media predicts higher levels of anxiety among Canadians

Data Source:
All the data will be from the CSCS data, found in the two .csv files var_names and CSCS_data_anon, for the data collected from surveyed Canadians

Outcome Variable (Y): Level of Anxiety - Q66_1 from 2021_cross dataset

- As my analysis aims to predict the levels of anxiety, this is my outcome variable - continuous data
- Measured in the CSCS (Canadian Social Connection Survey) in question 66_1, which asked "Over the PAST TWO WEEKS, how often - have you felt nervous, anxious or on edge?"
- The responses returned a number of days - all ordinal data will be converted into numerical, in order for the data to be able to be visually represented in a graph
- Visualisation:
    - A histogram could be used to see the distribution of anxiety levels across the data in my sample, which is useful as it allows analysis to be done on the concentration of scores and check for the presence of outliers, and additionally allows for the visual comparison of the distribution of anxiety levels between different subgroups of data - in this case, low vs. high social media users.
        - Could be created using plotly, as I prefer its interactive features

Predictor Variable (X): Time spent on social media - Q55 from 2021_cross dataset

- My null hypothesis suggests time spent on social media has no influence on anxiety levels.
- Measured in the CSCS in question 55, which asked "In the past week, on average, approximately how much time PER DAY have you spent actively using social networking websites?"
- The responses returned were in minutes/hours per day - all ordinal data will be converted into numerical, in order for the data to be able to be visually represented in a graph
- Visualisation:

- A scatter plot could be utilised with the the predictor (X) and outcome (Y) variable on their respective axes to visualise there is a trend between the two variables, which could be useful as it allows one to see if there are any patterns within the data, and also allows for the checking of linearity of the trend of the data
    - Created using plotly, for reasons stated prior.

Analysis Plan:
I intend on using various Python libraries for various tasks, such as pandas for data manipulation, plotly or seaborn for data visualisation, and statsmodels for statistical modelling. Using these libraries, I will be able to clean up my data - such as by removing missing data (using the df.dropna() function), and get descriptive statistics using the .describe() method. Following such analysis, I will be able to visualise my data using graphs and plots and visually assess the relationship between my two variables, and try to fit a simple linear regression line to my data, using .fit() and .summary(). I am choosing to try and fit simple linear regression as I am doing an analysis on just two numeric variables, which is a suitable usage for the regression.

Assumptions of simple linear regression:
- Linearity: The relationship between my predictor and outcome variables should be linear in nature
- Homoscedasticity: there is a constant variance throughout all my data points and the regression line
- Independence of Errors: The errors should be independent of each other.

My null hypothesis assumes that there is no influence of the time spent on social media on the levels of anxiety felt, and thus the slope of the regression line would be 0 in the case that my null hypothesis is supported.

The findings from this analysis may be used to inform the general public about the possible detrimental nature of our usage of social networking websites and media on our mental health due to its influence on anxiety levels, and may be used to find a better balance for individuals that will promote healthier lifestyles as a treatment for individuals suffering from generalised anxiety disorder.

Research Question 2:
Does a higher level of educational attainment (Bachelor's, Master's, or Doctoral degree) predict higher household income in the Canadian population, based on data from the Canadian Social Connection Survey?

Q53, Q96 - From var_names.csv

Null Hypothesis:

- There is no association between the level of educational attainment and the household income in the Canadian population.
    - If the null hypothesis is supported, then the answer to the research question is no, attaining a higher level of education predicts no difference in household income in the Canadian population.

Alternative Hypothesis:

- A higher level of educational attainment predicts a higher household income in the Canadian population.
    - If the null hypothesis is rejected and the alternative hypothesis is supported, then the answer to the research question is yes, attaining a higher level of education predicts higher household incomes in the Canadian population.

Data Source:
All the data will be from the CSCS data, found in the two .csv files var_names and CSCS_data_anon, for the data collected from surveyed Canadians

Outcome Variable (Y): Household Income - Q96 from 2022_cross dataset

- As my analysis aims to predict the household income, this is my outcome variable - continuous data
- Measured in the CSCS (Canadian Social Connection Survey) in 2022 in question 96, which asked "What is your best estimate of your total household income received by all household members, from all sources, before taxes and deductions, during the year ending December 31, 2021? Note: Income can come from various sources such as from work, investments, pensions or government. Examples include Employment Insurance, Social Assistance, Child Tax Benefit and other income such as child support, spousal support (alimony) and rental income."
- The responses returned numerical responses, with some being specific numbers, and others giving a range. For the responses which returned a range, the median value of the range was taken.
- Visualisation:
    - As income data often tends to be right-skewed, as the majority of values are clustered at the lower end, with a long tail extending towards higher values as a result of the influence of high earning outliers, I could apply a log transformation in efforts to reduce the possible right-skewness of the data and also improve the interpretability of the data. After the log transformation, the data could be represented in a simple histogram displaying the distribution of the data.

        - This visualisation would be useful as it can be used to break the incomes down into brackets with the use of bins, which would allow the understanding of the sample size in each bin (income bracket).
        - Would be created using plotly

Predictor Variable (X): Level of Higher Education - Q53 from 2022_cross dataset

- My null hypothesis suggests time spent on social media has no influence on anxiety levels.
- Measured in the CSCS (Canadian Social Connection Survey) in 2022 in question 53, which asked "Do you have a Bachelors, Masters, or Doctoral degree from a 4 year university?"
- The responses returned categorical responses of No, Bachelors, Masters, or Doctorate - different variations (such as PhD for Doctorate) of each of these responses would be organised together
- Visualisation:
    - Multiple violin plots could be used to combine the elements of a box plot with kernel density estimations, in which the violin plots would plot the household incomes for the surveyed individuals with the categorical education levels as the grouping variable.
        - This would be useful as it allows one to to visualise the median, IQR, density distribution, and also potential multi-modality of household incomes within each education level
        - Would be created using the Python library plotly due to its interactive features

Analysis Plan:

I intend on using various Python libraries for various tasks, such as pandas for data manipulation, plotly or seaborn for data visualisation, and statsmodels for statistical modelling. Using these libraries, I will be able to clean up my data - such as by removing missing data (using the df.dropna() function), and possibly apply log transformation if necessary. Following this, I will convert my categorical IV "Level of Higher Education" into numerical indicator variables, with the use of the pandas method pandas.get_dummies():, creating new columns in my df - such as Education_Bachelors, Education_Masters, and Education_Doctorate - with the column having values 0 or 1 - indication of the presence of absence of the specific education level for each data observation. Then, I will utilise a simple linear regression model and fit it with the .fit() method, and .summary() to examine regression results, with a focus on the p-values to identify statistically significant differences between the observed statistics and the regression model.

Upon conducting my analysis, I will summarise my findings, and answer my research question, indicating which of the two hypotheses were supported.

The findings from this analysis may be used to demonstrate to the general public the beneficial (for at least the economic aspect) nature of higher education - and its different levels, to encourage the increase of individuals undertaking further studies.

Research Question 3:
Is there a difference in life satisfaction between individuals who are single and those who are in relationships among Canadians?

Q8_1, Q63 from 2023_cross in var_names.csv

Null Hypothesis:

- There is no difference in average life satisfaction scores between individuals who are single and those who are in relationships among Canadians.
    - If the null hypothesis is supported, then the answer to the research question is no, there is no difference in life satisfaction between individuals who are single and those who are in relationships among Canadians

Alternative Hypothesis:

- There is a difference in average life satisfaction scores between individuals who are single and those who are in relationships among Canadians.
    - If the null hypothesis is rejected and the alternative hypothesis is supported, then the answer to the research question is yes, there is a difference in life satisfaction between individuals who are single and those who are in relationships among Canadians

Data Source:
All the data will be from the CSCS data, found in the two .csv files var_names and CSCS_data_anon, for the data collected from surveyed Canadians

Outcome Variable (Y): Life Satisfaction - Q63 from 2023_cross dataset

- As my analysis aims to predict the levels of life satisfaction on a scale, this is my outcome variable - discrete ordinal data
- Measured in the CSCS (Canadian Social Connection Survey) in 2023 in question 63, which asked "On a scale of 1 to 10, How do you feel about your life as a whole right now?"
- The responses returned a number on a scale from 1 to 10 - 10 being maximum satisfaction and 1 being the least satisfaction possible.
- Visualisation:
    - This variable can be visualised using a bar plot, as it allows the visualisation of the counts for how many times a unique value appears in a column - in this case the scores (1 through 10) would be individual columns, and the frequency of each score would be represented by the height of the bar of the column.
        - This is useful as it emphasises the discrete nature of the data, while presenting a straightforward and easily interpretable visual representation of the frequency of each score for the sample.
        - Could be created using plotly, as I prefer its interactive features

Predictor Variable (X): Relationship Status: Single - Q8_1 from 2023_cross dataset

- My null hypothesis suggests relationship status has no influence on life satisfaction levels.

- Measured in the CSCS (Canadian Social Connection Survey) in 2023 in question 8_1, which asked “What is your current relationship status? (Check all that apply) - Single”
- The responses were collected by individuals either checking or not checking the box for this question - translated into yes or no
- Visualisation:
    - Box Plots for comparison with Single and Taken as the two boxes and the Outcome Variable (Life Satisfaction) to be visualised within the two box plots
        - This would be useful as it allows one to to visualise the median, quartiles, potential outliers for each group - in this case Single, Taken
        - Would be created using the Python library plotly due to its interactive features

Analysis Plan:
I intend on using various Python libraries for various tasks, such as pandas for data manipulation, plotly or seaborn for data visualisation, and statsmodels for statistical modelling. Using these libraries, I will be able to clean up my data - such as by removing missing data (using the df.dropna() function). Following this, I will utilise a simple linear regression model with life satisfaction as the outcome variable and the “Single” indicator variable as the predictor, as specified above - and the p-value associated with the coefficient for the indicator variable would allow me to examine the regression results, testing the null hypothesis to identify statistically significant differences between the observed statistics and the regression model, to see if there is a difference in the average life satisfaction between single and non-single individuals.

Upon conducting my analysis, I will summarise my findings, and answer my research question, indicating which of the two hypotheses were supported.

The findings from this analysis may be used to demonstrate to the general public, especially those seeking happiness in relationships to possibly reconsider - and maybe find that true satisfaction is found in oneself, or perhaps encourage people to find happiness in relationships - which could be used in ads for dating app campaigns.