STA130 Individual Project Proposal

Groupmate request: Emma DeCouto

1.

Research Question	Do students studying their bachelor degree sleep more than 8 hours or more a day?
Null Hypothesis	Students studying their bachelor degree sleep less than 8 hours per day
Alternative Hypothesis	Students studying their bachelor degree sleep 8 hours or more per day

Variables:
2021_cross - DEMO_age
2021_cross - LIFESTYLE_time_use_balance_sleeping

Population:
I plan on using the specific age range given in my question along with the time used to sleep to make up the population. For simplicity's sake, I will say that the age range is from 17-23 years old, assuming that the person left for university right after their high school education.

Visualisations:
I believe the bar plot would be the best way to visualise this because the data is simple and this would be an effective way. The X axis would be the age and the y axis would be how many hours they sleep.

Methods
First, I would discard the data from the people whose ages are outside of the range. Then, using the correlation between the two variables, I would perform hypothesis testing. If students do sleep 8 or more hours per day, then the null hypothesis can be discarded.

2.

Research Question	Do older people take up longer to answer surveys than younger people?
Null Hypothesis	Older people take up a longer duration
Alternative Hypothesis	Older people take up shorter duration

Variables:
2021_cross - DEMO_age
2021_cross - SURVEY_duration_seconds
2021_cross - SURVEY_progress

Population:
I plan on taking into account the given ages, the completion percentage, and the time taken to make up the population.

Visualisations:

A series of box charts would best represent this correlation visually. The X axis would be the ages in intervals and the Y axis would be the time as the box chart would be vertical. Furthermore, having a variance of linear regression analysis through providing a scatterplot (also considering there's more than 2 variables being used) would give more intel on the sample's patterns. A bar plot could also be used, along with a line to show linear regression analysis by including a scatter plot (additional visualisations like this may be added depending on how the data changes). The Y axis being the ages and the X axis being the time.

Methods:

First, only particular data will be used, as those outside of the age range of the population will be irrelevant to the research question. If the completion progress is less than 100%, then the duration would be essential, as to give the most accurate data possible. Through hypothesis testing, the null hypothesis and alternative hypothesis can be concluded through the data.

3.

Research Question	Do older individuals who live with their parents have better mental health as opposed to those who don't live with their parents?
Null Hypothesis	Older individuals who live with their parents do have better mental health
Alternative Hypothesis	Older individuals who live with their parents have worse mental health or the same scale of mental health

Variables:

2021_cross - DEMO_age
2021_cross - GEO_housing_live_with_parent
2021_cross - WELLNESS_self_rated_mental_health

Population:

I plan on using 3 variables which is more than the standard 2 or less we've been using in this course. The 3 variables would be how old people are, if they live with their parents or not, and how their mental health is rated.

Visualisations:

Quite like the previous proposal, There are two possible methods, analysing them without intervals through a scatter plot or with intervals, through a box plot. For the box chart, we could make the x-axis have 2 variables, with one of them being whether the person lives with their parents or not and the other being their age. That means there would be 2 bars right next to each other for each age. The y-axis would be the rated mental health.

Methods

There are two possible methods, analysing them with or without intervals. You could also interpret the data as either numerical or ordinal depending on if it’s categorised or not. Also, there would be a specific range for the population. From here you would do hypothesis testing to accept or reject the null hypothesis.

Results for Every Proposal:

Potential results will either prove or disprove the null hypothesis

It is important to note that there would be data points that serve as outliers or inaccurate data from the data extracted from the dataset. For the first proposal, students with extremely high hours (health issues, irregular sleep patterns for example) and students with extremely low hours (academic life, sleeping disorders, for example) would distort the average of the data and a linear regression model if it were to be drawn. For the second proposal, if some individuals may skip some responses which would provide inaccurate data. For the third proposal in particular, there may be external forces that impact the mental health of students, whether it’s due to their school environment, work environment, or another factor.