Coco Cheng

Mingyang Xue

Short Sleep Duration (Insomnia) and its Impact

Summary of research questions and Results

1. As time goes, do the number of people with the problem of short sleep duration or insomnia in the U.S. increase or decrease?

   Over time, the number of people who have short sleep duration in the U.S. decreases. In the year 2015, the number reached the lowest, and it rebounded back in 2016.

2. What kind of characteristics of people like their ages and sexes attribute to sleep problems?

   Sex has little contribution on sleep problems. Old people have better sleep quality or fewer sleep problems than young people. People with higher education levels generally have worse sleep quality or more sleep problems than people with lower education levels. Students in university have the worst sleep quality or most sleep problems in this dataset. Fat people have better sleep quality or fewer sleep problems than thin people.

3. What is the impact of insomnia have on people like damaging people's health?

   People with insomnia are more likely to experience some mental health problems like anxiety and depression. As insomnia severity index increases, anxiety index and depression index also increase. There is a negative relationship between insomnia severity index and health problem index. There is not an obvious linear relationship between bmi and insomnia severity index.

Motivation and background

We want to do research on short sleep durations because we are college students and most of the college students either do not have enough sleeping time due to coursework or having psychological related sleeping troubles, leading to short sleep durations.

The first research question is essentially helping people to understand the importance of the problems. By knowing how many people or what percentage of the population are struggling to sleep, we will understand it is actually a serious issue that people need to pay attention to. Knowing the answer for the second research problem will help us figure out what kind of people are more likely to have sleep problems. For instance, we know that older adults also need 9 hours per day of sleeping time but age-this characteristic may suggest that older people are more likely to have troubles falling asleep. This is worth computing after knowing the contributing factors to sleep problems, people who are with those characteristics will be aware of the possibility and it may also be helpful for figuring solutions for the issue. The third research question. It has similar use as the first research question to remind people to be aware of the short sleep duration problem. Understanding the potential impact of the problem like a higher possibility of having high blood pressure, diabetes and depression suggests why the problem of short sleep duration needs some immediate action. Getting the answer for this question also helps to conclude that people with depression or diabetes may be healthier if they can improve their sleeping quality.

Dataset

1. https://data.world/makeovermonday/2019w23
2. https://www.kaggle.com/feraco/sleep-deprivation/data?select=data_sleepybrain.csv

For the one on data world website, the data file is called "Time Americans Spend Sleeping." For the data in Kaggle, there are two datasets - "data_sleepybrain.csv", "demdata_160225_pseudonymized.csv," and the explanatory document "demdata_codebook_160226.txt."

Methodology (algorithm or analysis)

To analyze the first research question, we are using the dataset called “Time Americans Spend Sleeping.” We filtered the data by taking out the parts not during the time from 2004 to 2016 and the data with missing values like NaN. After filtering, We sorted the data by the order of the year. We counted people who are under 25 years old and sleep less than 8.6 hours per day and people who are above 25 years old and sleep less than 9.3 hours per day as people with short sleep duration or so-called insomnia. With the counting number, we plotted out a line chart to indicate as time goes, if the people with the problems of short sleep duration increase or decrease.

The datasets that we use for the second research question are the “demdata_160225_pseudonymized.csv” and “data_sleepybrain.csv” from Kaggle. In these two datasets, we have about 100 patients and lots of information about them is collected and listed in different columns in these datasets. First, we use some python codes to clean these two datasets in order to merge them to one single sheet of data, which is easier for us to handle. We edit one column named “id” in “data_sleepybrain.csv”, so we can merge these two datasets to one by connecting the two columns named “id” in these two sheets. Then, we clean the merged data by making some columns readable and concise. More specifically, we firstly select 11 columns that we will use in this sheet and drop all other columns. Secondly, since values stored in the “BMI1” column are strings and contain “,”, we replace the “,” with “.” and therefore we are able to convert those values to floats. Then, because there are three columns, named “KSQ_HealthProblems”, “KSQ_OverallSleepQuality” and “EducationLevel” seperately, contain Swedish, we make them into columns that only contain numbers as representation. Numerical values are much easier and, in all sense, enough for our analysis of this problem. After the cleaning of data, we export the new dataset which will be used later to a csv file named “data2.csv”.

Now, we need to know whether their characteristics contribute to sleep problems, so we choose the column named “KSQ_OverallSleepQuality” as the measurement for sleep problems of patients. Here, after reading the instruction of the original dataset, we assume

that an individual with higher KSQ_OverallSleepQuality value sleeps well and has little sleep problems, and vice versa. Then, we put all the patients into different categories according to their sexes, ages, BMIs and education levels using the “groupby” method in python and calculate the mean of KSQ_OverallSleepQuality within each group. Then, with the sample statistics we receive from the data, we observe there are only two categories of sex in this dataset, which are female and male, and they have slightly different but very similar value of mean of KSQ_OverallSleepQuality according to the data we have, so we want to set up a two sided hypothesis test to show if the female population has the same level of overall sleep quality as male population. Let **μ_f** be the mean of KSQ_OverallSleepQuality for females and **μ_m** be the mean of KSQ_OverallSleepQuality for males. In this case, the null hypothesis is **μ_f = μ_m**, and we make the size of test, which is also known as alpha, be 0.05, which means if we get a p value from this data greater than 0.05, we fail to reject the null hypothesis, implying sexuality does not affect overall sleep quality; if we get a p value from this data less than 0.05, we reject the null hypothesis, implying sexuality has some effects on overall sleep quality. We use the function ttest_rel() in scipy.stats to help us calculate p value from the data. As long as we plug in the data of KSQ_OverallSleepQuality for females and males separately as two parameters, the function returns the p value to us automatically. Then, we choose to plot the best fit lines, to show the correlation between KSQ_OverallSleepQuality and ages, education levels, BMIs of patients. After finishing these steps on the above, we are able to come up with some ideas about which characteristics of people contribute to sleep problems and short sleep duration.

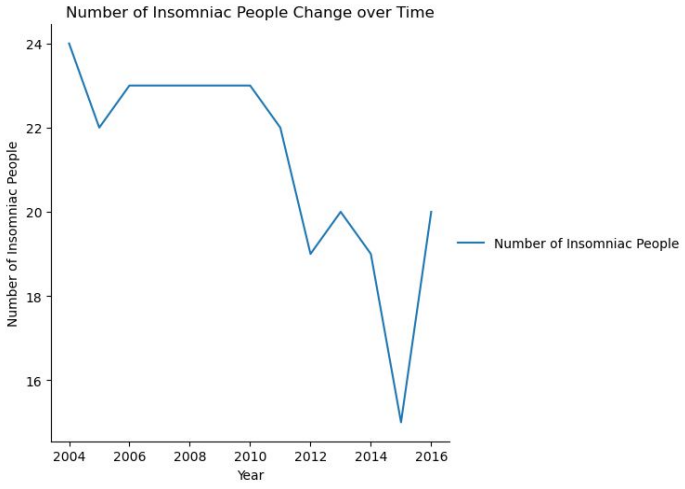
The data we will use for the third research question is “data2.csv” that we already cleaned up from “demdata_160225_pseudonymized.csv” and “data_sleepybrain.csv”. First, we do some research on how sleeping may affect people’s mental health. After understanding the domain knowledge, we essentially set ISI as the independent value which is typically x value and have anxiety index, depression index, bmi, and health problem index as the dependent value which is y. We plot them in the same figure to see if either depression index and ISI or anxiety index and ISI have positive linear or negative relationships to analyze if short sleep duration have influences on people’s mental health like depression and anxiety.

Using scatter plot may not necessarily show the relationship since it may overfit so that we also plot out the best fit line.

Results

1. **As time goes, do the number of people with the problem of short sleep duration or insomnia in the U.S. increase or decrease?**

   The line plot below shows that as time goes, the number of people with the problem of short sleep duration decreases in the U.S. It reaches the minimum at the year of 2015 and rebounds after that. Every year there are 63 participants participating in this experiment which suggests that even though there's a downtrend of the insomniac population, there are still roughly 30 percent of the population struggling to sleep. The reason that the insomniac population significantly dropped at the year of 2015 can be that there were two new drugs that had been approved by the FDA which were tasimelteon in January 2014 and suvorexant in August 2014.

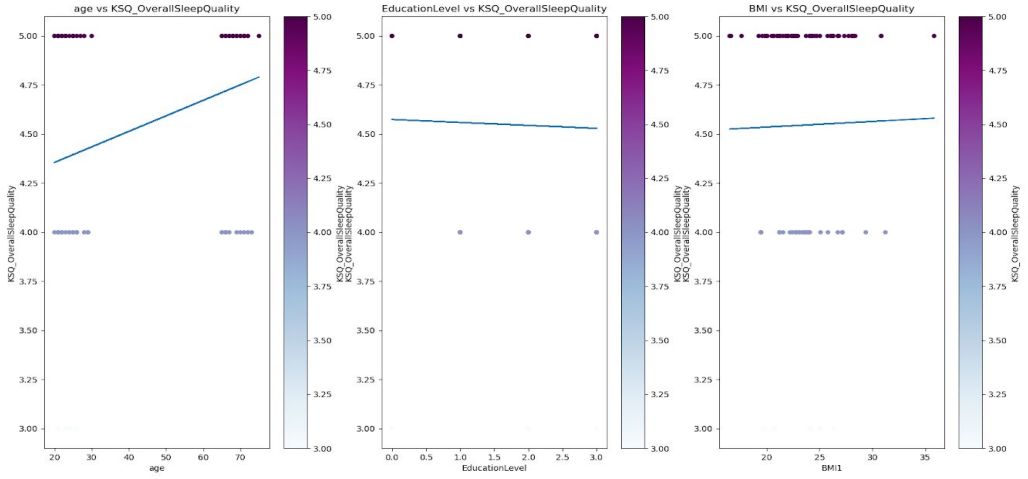


2. **What kind of characteristics of people like their ages and sexes attribute to sleep problems?**

Average sleep quality measurement for different groups of individuals are listed below:

- average sleep quality for female: 4.614
- average sleep quality for male: 4.476
- average sleep quality for young people: 4.383
- average sleep quality for old people: 4.744
- average sleep quality for individuals who finished primary school: 4.5
- average sleep quality for individuals who finished secondary school: 4.667
- average sleep quality for individuals who are students in university: 4.387
- average sleep quality for individuals who finished university: 4.625

For the hypothesis test, we set null hypothesis H_0 to be **$\mu_f = \mu_m$,** where **μ_f** is the mean of KSQ_OverallSleepQuality for females and **μ_m** is the mean of KSQ_OverallSleepQuality for males. And the alternative hypothesis H_1 is **$\mu_f \neq \mu_m$.** The p value we get is 0.418. Since 0.418 > 0.05, we fail to reject the null hypothesis and conclude that there is no statistically significant difference between overall sleep quality of females and males. The three plots of three different characteristics are displayed below.



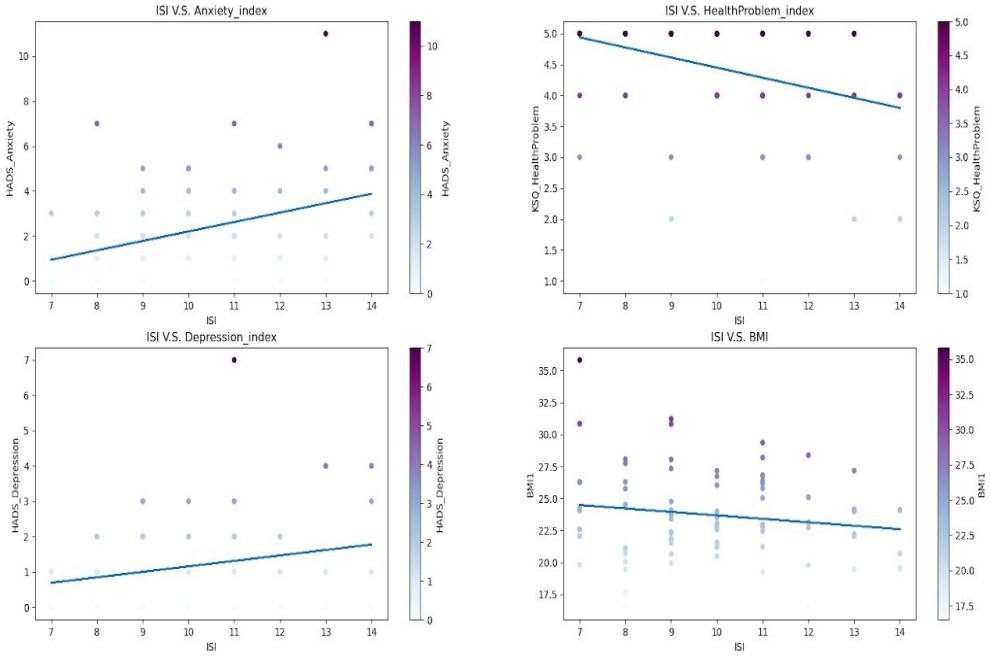
From the computed means and plots we get, we derive several conclusions.

1. Sex has little contribution on sleep quality
2. Old people have better sleep quality than young people on an obvious scale.
3. People with higher education levels generally have worse sleep quality than people with lower education levels. Students in university have the worst sleep quality in this dataset.
4. People with higher BMI have better sleep quality than people with lower BMI. In other words, fat people have better sleep quality than thin people.

3. What is the impact of insomnia have on people like damaging people's health?

People with insomnia are more likely to experience some mental health problems like anxiety and depression. Based on the graph shown below, as insomnia severity index increases, anxiety index and depression index also increase. In other words, there is a positive relationship between the variables which implies that the impact of having severe insomnia is harming people's mental health. There is a negative relationship between insomnia severity index and health problem index which is not matching what psychologists and other scientists have suggested. The reason for it may be that the KSQ_HealthProblem indexes are collected from the Karolinska Sleep Questionnaire. There may be a report bias since most of the people

tend to not admit that they are not healthy so that they may not tell the truth when answering the questionnaire. There is not an obvious linear relationship between bmi and insomnia severity index as the fourth plot shown.



Challenge Goals.

1. **Multiple Datasets:** We find two datasets which can be used as the sources for our project. These two datasets are both relevant to short sleep duration as we described above. We wish to put them together in analysis and come up with a richer conclusion.
2. **Result Validity:** We want to use a statistical hypothesis test to show that the conclusion we get from analyzing those two datasets is valid at size equals 0.05(alpha = 0.05). In this project, we use the T test to do the hypothesis test and set the null hypothesis to be **μ_f = μ_m,** where **μ_f** is the mean of KSQ_OverallSleepQuality for females and **μ_m** is the mean of KSQ_OverallSleepQuality for males. Then we use the function ttest_rel() in scipy.stats to help us calculate p value from the data and conclude that there is

no statistically significant difference between overall sleep quality of females and males in a very rigid way.

Work Plan Evaluation

Plan

We will use **workplace on Ed** to collaborate finishing the research and making sure both of us have access to source code. We are dividing the tasks into three main parts based on the research questions.

Firstly, Coco will work on research question 1 to show the reason why we want to do this research since it is actually a more serious problem than most people will assume. Coco will need about **half an hour** to do some searching and reading articles about the importance of the sleeping problems as the background information to solve the research question. Then Coco will need **an hour** to clean and filter the data to the number of people who have bad sleeping quality or short sleep duration out from the sample data using python. Lastly Coco will need **two hours** to use data visualization like bar charts or pie charts with Python to show the percentage of the people who have sleeping troubles to show the importance of understanding there are actually how many people are struggling with short sleep duration.

Secondly, Mingyang will work on the research question 2. This question is very important for our research, because only if we figure out how people with different characteristics behave differently, further analysis and speculation can be done with the data we have. Mingyang will spend about **one hour** filtering patients in the dataset into several categories according to their physical characteristics. Then, Mingyang will spend about **half an hour** designing hypothesis tests about the physical characteristics. That is to say, we want to know if female and male patients have identical sleeping hours and sleeping cycles. After making the hypothesis tests, Mingyang will spend about **two hours** trying to apply the data he receives for different groups of people into hypothesis tests and draw conclusions whether we should consider people with different characteristics behave differently regarding their sleeping information.

Lastly, Coco and Mingyang will work together on the research question 3. Firstly, Coco will do some background research on what are some potential impacts of short sleep duration which will take about **an hour**. After having some domain knowledge, Coco will clean up and filter the data that she thinks will contribute to getting the answer for this question and share her thoughts in google documents with Mingyang which will take about **two hours**. Mingyang will show the data about the impact of short sleep duration on different groups of people using various plots in Python. After doing the data visualization, Mingyang will set up the hypothesis tests and try to draw conclusions about how short sleep duration impacts health of people, which will take about **two hours**. Mingyang will search and read research articles on this topic to come up with explanations and supporting evidence for her result analyzing the given data which will take about **an hour**.

Evaluation

The predictions for how many hours we would spend on each question are not so precise. An important part that we did not take into consideration is the time for writing the report and other supporting documents. We ended up spending a much longer time for the project than we had planned. Firstly, cleaning the dataset from Kaggle took a longer time than an hour. Secondly, we spent longer time on coding since there are many parts that we are learning from online resources like how to calculate P-value and find the best fit line in Python. Also we did not think about the testing part in advance which is time-consuming. Lastly, even though our result like the data visualization shows the answers for all of our research questions, it is still hard for us to present the results in words which takes at least two more hours than planned.

Testing

For Q1, we are testing to see if our standard for short sleep duration is fair or not and accuracy of the line plot. We calculate the average sleeping hours for the whole dataset including the years that we did not include for the graph. The average sleeping hours for the entire dataset is 8.8 hours which shows our standard that both 8.6 hours and 9.3 hours are

closed to 8.8. In terms of accuracy, I test if the year with the minimum number of insomniac people is 2015 and it is 2015 which shows that the graph is accurate.

For Q2, the most important part that we need to test is the computation of average sleep quality of different groups of patients. As long as we get the correct mean values, we can make the plots and calculate the p value using some handy and reliable functions in python, which make the process and result unambiguous and inerrable. Thus, we use another method to compute average sleep quality of different groups of patients and print them out. We also print out the results we get in Q2, so we can easily compare these two groups of values in the output and see whether we do something wrong in Q2.

For Q3, I use the idea of testing on a similar data set. Both insomnia severity index and sleeping quality index are a way to measure the severity of the sleeping problem. In the testing, I change the independent value from ISI to sleeping quality index and get similar scatter plots shown to show that my result for Q3 should be credible. For instance, my testing plots of bmil and sleeping quality index also do not have an obvious linear relationship.

Collaboration

We do not have anyone who is not in the class CSE 163 help us. We did use online resources like Stackflow to solve some difficulties with python coding. We used the websites listed below as background knowledge to help us analyze the results we got for the research questions.

https://onlinelibrary.wiley.com/doi/10.1111/crj.12095

https://www.acc.org/latest-in-cardiology/articles/2015/06/16/08/40/new-insomnia-drugs-in-the-context-of-cardiovascular-disease