How might we encourage teens/pre-teens to be mentally and physically healthy?

Kaushal Gianchandani, Abbie Tolon

First off, kudos on the choice of problem! This is indeed an important issue.

While using data science to tackle a social or healthcare related problem, one has to strike a balance between the size of the dataset and the level of details. Building datasets with several details is both expensive and time consuming. Moreover, it often relies on picking a small sample from a very large population. These small samples may not be representative of the actual population and an analysis based on such small samples can be skewed. Alternatively, large datasets that span a big chunk of the population can be used to draw more generalized insights, however, these datasets may not be detailed enough and can often lead to spurious correlations/conclusions.

Keeping this in mind, we propose to work with data provided by Google Trends, which is a website that provides the popularity of different search terms in different geographical domains. Given Google's reach, it is fairly certain that the information conveyed by this data is representative of a large chunk of the population. Furthermore, the Interest in each search term over time is quantified on a scale of 0 to 100, which makes it convenient to draw a comparison between two or more search terms. However, Google's privacy policy does not let us access details such as age, gender, ethnicity, education background and other personal details of the users.

We will use Google Trends data to try and identify the internet search patterns of people struggling with mental/physical health issues. Based on that information, we will compile a list of suggestions that will help teens/pre-teens focus on their mental/physical health.

In the subsequent part of this handout, we discuss a piecewise outline for the project.

Task 1: Data identification

To begin this analysis, let us compile a list of search terms that you think are relevant to the problem. You can think of it in terms of two questions:

1. What "search term(s)" can best quantify the mental health of a person? For instance, a person who is in a foul mood for multiple days can google "I am depressed. What can I do to be happier?", the key word in this sentence is "depressed" because that is what we are trying to predict. There can be multiple such search terms and we need to make a list of these search terms.
2. What search term(s) can affect the mental health of a person? For instance, a person can be in a foul mood because they are stressed at work, so the relevant search term in this case is "work stress"

Suggestions: Physical fitness, Anxiety, Work stress, Instagram

Task 2: Data acquisition

Once you have identified a bunch of relevant search terms, the next step is to make a google sheet (let's call it *trends_data*) in which you document the popularity of each search term relevant to the problem. Here is an example of how to do it:

1. Go to Google Trends
2. Type in the search term, e.g. 'Physical fitness'
3. Change the parameters of the search:
    a. country: United States,
    b. time: 2004 - present,
    c. categories: All categories,
    d. Search type: Web Search.
4. Click the download button to download the data as a .csv[1] file (that can be opened on google sheets/excel) and copy the data in that file to *trends_data*, where the first column is Year-Month and the next column contains the number representing the interest from 0 to 100
5. Repeat this for multiple search terms. Note that after you have stored the first term, you only need to copy the second column from each individually downloaded csv file. Once you are finished, your excel file should look like this (the values quoted below are merely placeholders):

Time stamp	Physical fitness	Mental health	Depressed	Work stress	<other relevant terms>
2004-01	89	54	26	58	…
2004-02	80	57	31	100	…
2004-03	…	…	…	…	…

Task 3: Visualize the data and draw insights

Based on the data you have collected, we will do the following:

1. Calculate correlations between the possible cause and probable outcomes. For instance, to check if a fall in search of *Physical health* led to a rise in the search of *Anxiety* we will calculate the correlation between the interest in these two terms.
    a. If the number is close to 1, we can say that increased interest in *Physical health* is associated with increased interest in *Anxiety*
    b. If the number is close to 0, we can say that increased interest in *Physical health* search terms have no association with interest in *Anxiety* search terms
    c. If the number is close to -1, we can say that increased interest in *Physical health* search terms is associated with a decrease in interest in *Anxiety* search terms

2. After identifying the relevant correlations, we will use this data to make a chart to visualize the contribution of each of the relevant terms that can indicate that a person is not doing too well physically/mentally.

3. ***[If time permits]*** After identifying the terms with high correlations, we will build a linear regression model to identify the relevant terms which can indicate that a person is not doing

[1] csv: Comma-separated values

too well physically/mentally. The linear regression model will hopefully help us predict these problems.

This objective analysis will serve as a basis upon which we will then make a set of guidelines to help teens/pre-teens to stay healthy.

Task 4: Identify the limitations of our approach

To conclude our analysis we will try to identify the limitations of our approach. For examples, we encourage you to think about the validity of this statement cited in the second paragraph of this handout: "*information conveyed by this data is representative of a large chunk of the population.*" According to you, is this statement true or false? Give your reasons to support your answer.

Another limitation to note is that we cannot definitively conclude why a person uses a search term. For example, a person might search the term *Anxiety* because they are writing a research paper, or they simply want to learn more about the topic, rather than because they themselves have anxiety. Similarly, searching for *Physical fitness* terms does not necessarily mean a person is exercising. Moreover, when comparing the trends between *Anxiety* and *Physical fitness*, we cannot assume the two are actually related without additional analysis and without controlling for other variables. Instead, this can serve as a good primary analysis for generating additional questions for future research. We can also generate initial hypotheses.

Key takeaways from the discussion:

Moving averages
[KG: Add notes and examples]