I am one of the only people I have ever met who doesn't use social media. Instagram, tik tok, facebook, etc. In today's world the majority use it for entertainment as it allows us to ease boredom and it's something we don't have to think about. But I've never used it at all and many people ask me how I'm able to survive without it. Growing up my parents never allowed me to have it but it never made me mad or annoyed. I never cared for it, it was never something I had an urge to use and I'm quite happy for it. Furthermore, whenever an adult did talk about social media it was always about how bad it was, like smoking or drinking which I don't do either. Oftentimes we see documentaries, stories and statistics about how social media ruins lives and makes people feel jealous and inferior and is a waste of time, preventing us from actually interacting with the world. Although we can all make predictions about whether this is true or not I want to see through my own data. By tracking the usage of social media I want to see if people who use it more have higher rates of unhappiness, depression, anxiety, etc. **Does Social Media affect our wellbeing negatively?**

To answer this question I used the *Social Media User Behavior & Lifestyle – 1 Million Synthetic Users (Instagram 2025–2026 based).* Before all else it is important to note that this is not real data from real people. Instead the Dataset is described on kaggle as "**1,000,000+ fully synthetic user profiles** that realistically simulate Instagram usage patterns combined with detailed demographic, lifestyle, health, and behavioral attributes." It is safe for use and research and is incredibly large, providing a huge pool of data. The data was taken from statistical distributions and realistic correlations. It's grounded and math and will be the data set I use to see how social media (specifically instagram) stacks up against things like physical fitness, relationships, behaviour, and happiness.

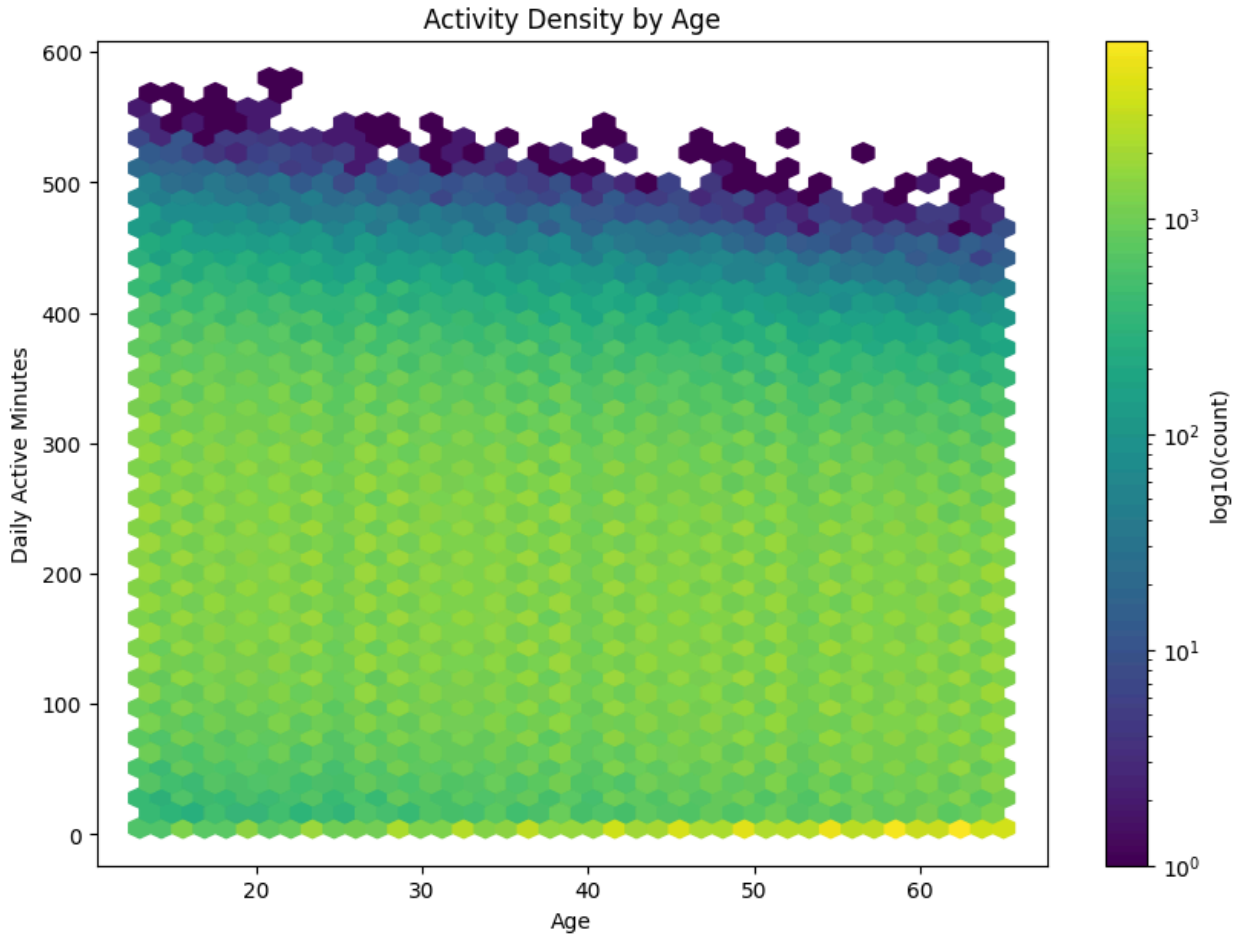
The first thing to think about was how many columns out of the 58 that I should keep. For the purposes of the question I want to stick to columns that accurately portray most importantly the length of time on social media  and level of activities outside of it. For this reason certain columns like content type, which specific social media app, and preferred content were removed. I wanted columns with integer/float values over booleans so we could better track change over time but kept all columns that increase a person's happiness and overall wellbeing like relationship status and income level. After sorting columns the ones I decided to keep were

1. Age
2. Gender
3. Country
4. Income_level
5. Employment status
6. Relationship status
7. Has children
8. Exercise hours per week
9. Sleep hours per night
10. Diet
11. Perceived stress score
12. Self reported happiness

13. Hobbies count
14. Daily activity minutes

After deciding what to keep and what not to do I need to clean my data set. WIth over 1 million rows, each one an individual synthetic user I decided first to check for and erase any duplicates. And to my surprise, there weren't any! Maybe because it was a synthetic dataset but there was not a single instance of any duplicate data sets. The data is incredibly clean and doesn't contain any missing values either. The kaggle data set is already clean to begin with. With that out of the way it's time to sort.

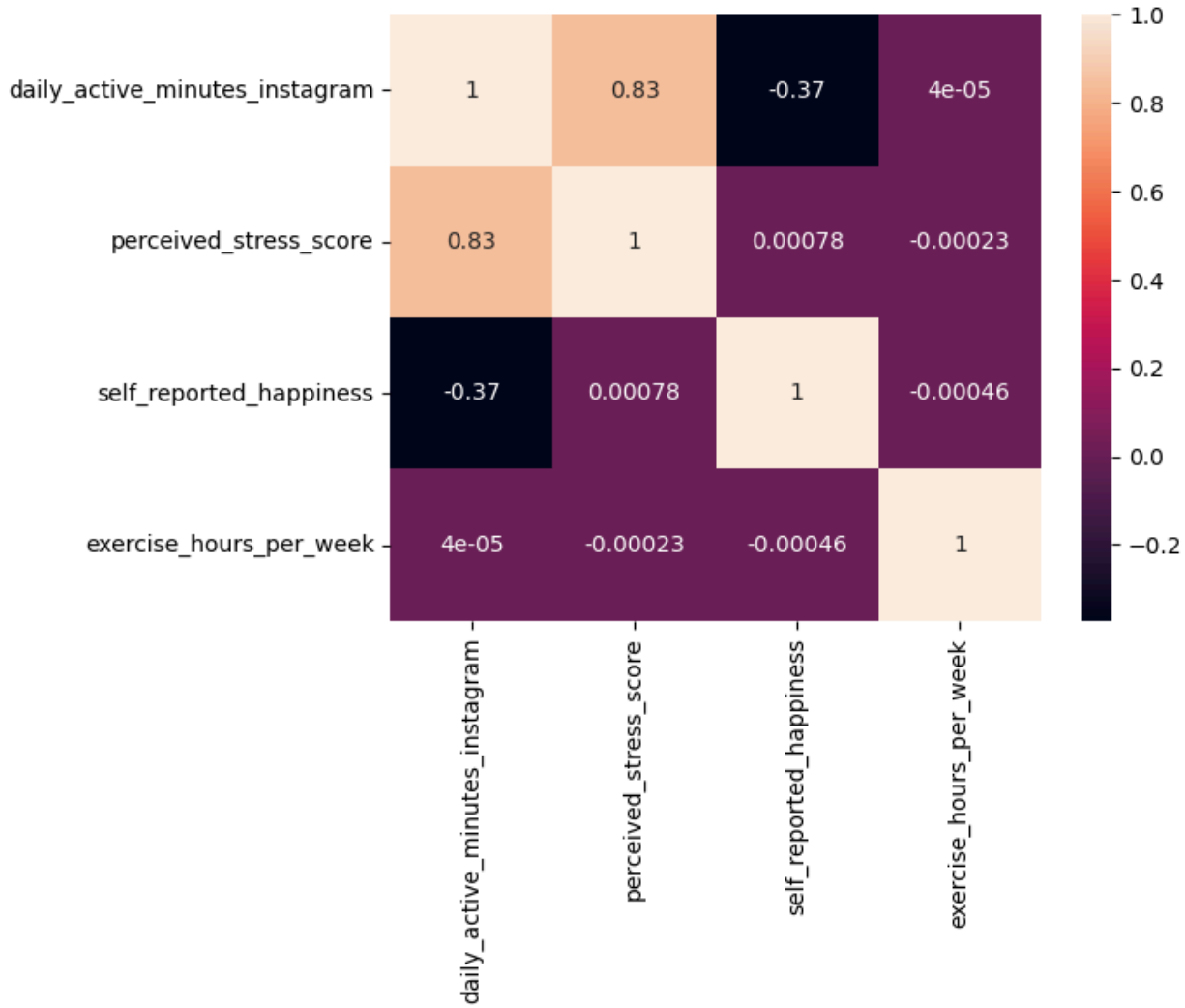
Since I'm basing my question off social media usage I sorted the data from least to greatest using the Daily activities minutes column. I decided to first try and graph age and social media usage as a test but ran into a sharp problem. One million is a huge amount of data and on a scatterplot it's just a complete mess. After doing some research I decided a hexbin would be the better graph and after two attempts came upon a decent looking graph.




A decently cool looking graph which makes no sense upon first examination. This Hexbin helps us visualize the correlation between age and daily activity measurements. Because the data set is so large the hexbin graph color codes entire chunks of people and places them where they fit in the data. The Yellow to light green hexagons are the hotspots where the biggest chunks of

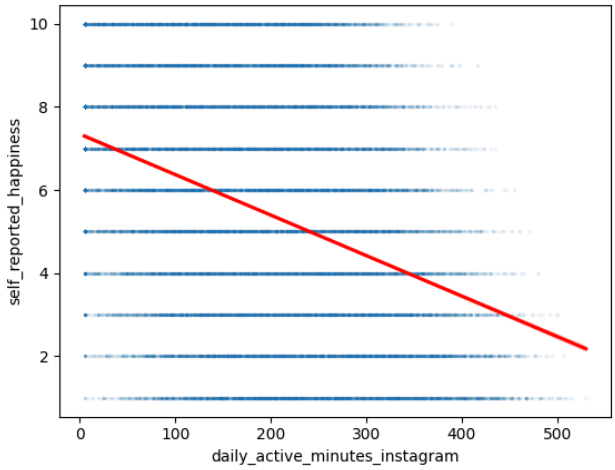
people lie (10^3). They all contain between 1000 to 3000 people depending on the amount of data points. The darker the hexagons the less amount of data is contained in each one. At the very top the black hexagons only contain about 1 to 10 people per hexagon. What this graph is able to tell us is that a small few people in their 20s have almost 600 minutes (10 hours) on instagram daily. This tends to die down as age increases with even the most seasoned social media enthusiasts only reaching just over 500 minutes (8.3 hours). As expected there are more yellow dots at near 0 as age increases meaning that the older you are the less likely you are to use social media. As age decreases the dots get darker meaning that less and less people are remaining at near 0 minutes of usage. What surprised me is that even the darkest hexagons in the 0 to 100 range are green meaning each hexagon is around 10^2 data points or above. That's an impressive number of people who don't use instagram at all. This graph tells us that the younger you are, the more likely you are to use Instagram. Something to remember

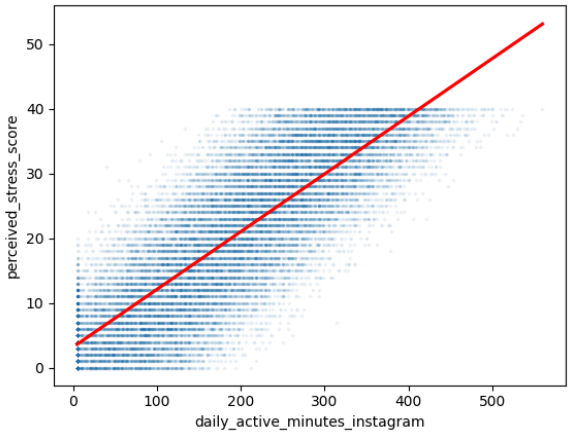
But then I took a step back and realized I had to find correlations that had to do with wellbeing. I needed to know what to look for. I changed my strategy and instead decided to first track key relationships using a heat map.




This heat map shows relationships between the columns and variables. The higher the number the higher the correlation up to 1 and the lower the number the lower the correlation. What jumps out immediately is the strong positive correlation between daily active minutes and perceived stress core. 0.83 is pretty high and means that as instagram time goes up stress score almost always goes up too. At the same time there is a moderately negative correlation between daily active minutes and self reported happiness, at -.37. That means to a certain but very real degree more instagram time is associated with lower happiness. All the other values are extremely close to 0 meaning there is no correlation. In this particular data set exercise hours have little correlation with any of the corresponding columns.

I want to continue to check the correlation between active minutes and perceived stress score and self reported happiness. The best way to do that is through a scatter plot but with a million data points we'll have to take a sample instead. Let's look at a sample of 30000:




Both graphs track their respective y axis over time, the amount of minutes spent on instagram daily. This is between 0 (none at all) and 500 which is over 8 hours a day. The y axis on the first graph to the left tracks happiness on a simple scale of 1 to 10. But the key finding is the red line, a regression line confirming what the heatmap showed us. There is a negative correlation. Those who spend less time on social media tend to report higher happiness (around 7-8) while increased daily exposure to instagram leads to decreased self reported happiness. The graph on the right supports the same narrative as there is a very clear positive correlation between instagram time and perceived stress. The narrative points to social media (instagram specifically) not being good for our happiness or stress levels.

To clearly define my question from earlier I'm asking: Does increased social media time affect our wellbeing (happiness, stress, etc)? After looking at the graphs I would have to say that it does. Each graph took a sample of 30000 and the correlation was still prevalent. According to the data, social media use does affect our happiness and stress. Also according to the data it has a negative influence.

But this is not the be all and end all. It's important to note that there are hundreds of other variables that can affect happiness. Correlation does not equal causation. To get an even more accurate interpretation we would need to factor in as many variables as we could. My data only factors in so many variables and even this dataset is synthetic, based on statistical analysis. Even so, what I found lines up with the common interpretations that people more skilled than I have found. If I had more time I would seek to explore every single variable through heatmaps and scatterplots and multiple regression analysis' but those are for another day.