TECHNISCHE UNIVERSITÄT BERLIN

Data Science with Python and R

Project Report

Danylo Ulianov, Batyr Ataev, Erdem Balli, Ömer Baycelebi
Matriculation number: 394635, 393289, 402431, 401549
Email: ulianov@campus.tu-berlin.de, ataev@campus.tu-berlin.de,
balli@campus.tu-berlin.de, o.baycelebi@campus.tu-berlin.de
Instructor: Guillermo Aguilar

Berlin, 04.03.2021
Winter Semester 2020/2021

Table of Contents

Executive Summary

The main question for the following research was the connection between social media usage and mental diseases all around the world. The idea was to connect the data from different sources and to find the correlation during the research.
The hypothesis of this project was based on the assumption that the increase of social media usage could increase mental disorders. Since there are a variety of mental disorders, it was decided to work mainly with three of those: ADHD, depression and anxiety.
During the project we made the decision to split up the task into two parts so that the results could either be proved or denied in various ways based on different data sets.
The results of this project are based on two different approaches. The first approach is based on the key variable “age” with which we compare the data sets. The second approach is based on the comparison of the World Happiness Index. This index plays a vital role since it is being used to create different country-pairs so that we were able to pair the countries in a rational way.

Chapter 1

Introduction

1.1 Motivation

Social media is a form of mass media, which allows users to globally interact and share information with each other. Thanks to it, we are more connected than ever before and are able to consume a huge amount of information very rapidly. However, just like everything social media has its disadvantages as well.
People are social creatures who in order to make progress in life require the companionship of the society. However, the desire for social connection to others can lead to a state where an individual feels depressed.[Kar+20]
By now there are over 4.2 billion social media users worldwide. It plays a huge role in our daily lives since the average person spends over two hours every day on social media platforms like Facebook or Instagram.
These platforms and their advertisers try to take advantage of this phenomenon by running ads to make a profit. Sean Parker, the first president of Facebook, disclosed that their platform was formed not to connect the users, but rather to distract them. Their goal was to consume the user's time as much as possible to get the maximum level of attention.
Since social media became a vital part of our lives, it is able to not only influence our mood, physical and mental health, but also the food we consume or the job we are working at.

1.2 Research question

In this paper, we are going to focus on one of the disadvantages by asking the following question:

Does the usage of social media increase mental disorders?

Different data sets which show the prevalence of mental disorders and social media usage in a year per country will be used in order to answer this question.

Chapter 2

Methodology

The structured work used during this project was organised in an agile way. Trying to analyse the main question, the goal was split in two approaches.

2.1 1. Approach: Using "age" as key variable

Our hypotheses is "Does the usage of social media lead to more mental disorders?" To test this hypotheses, we began by looking for data sets on the internet we could use to determine if there was some kind of a correlation between social media usage and mental disorder data. For mental disorder, we found a few data sets on an online publication called Our World in Data. The publication is edited by researchers from the University of Oxford and a non-profit organization named "Global Change Data Lab". From this publication, we chose the data set titled Share of the population with depression, 2017 which shows the prevalence of depressive disorders for each selected country.
After finding the data for mental disorder, we searched for data regarding social media. What came to our attention is that there was practically no data on social media as a whole, but most of the data sets focused only on individual social media platforms like Facebook or Instagram.
Looking at the data sets we came to the conclusion that in order to be able to compare the data sets we needed a key variable which was included in both data sets so that we could join the data sets. We agreed that the key variable should be "age" since our expectation was that the prevalence of mental disorder and the rate of social media users would be different for every age group.
This made the task to find fitting data sets even more challenging since we needed to not only find data sets for the year 2017 for both topics, but they also had to have "age" as a key variable in the data sets.

2.2 2. Approach: Using the World Happiness Index (WHI) to eliminate the confounders

In this approach we tried to eliminate as many confounders as possible by taking the World Happiness Index into consideration during our research.
The index is contained in the World Happiness Report, a publication of the United Nations Sustainable Development Solutions Network (UN SDSN), which mobilizes scientific expertise in support of the Sustainable Development Goals, for example developing a more sustainable future. The data was compiled from the World Gallup Poll while the happiness index was calculated by taking an average rate of 1000 people per country who evaluated their lives as a whole, with the best possible life being a 10/10 and the worst possible being a 0/10.
Economists also analyzed six major factors (Figure 2.1) which, according to them explain a significant part of the WHI. These factors are just an attempt to provide answers why it could be the case that some countries are ranked higher than others however they do not influence the total score. For example the GDP per capita factor of Norway is higher than the life expectancy factor, which means the former could have played a bigger role in the decision of a surveyed Norwegian than life expectancy. The residuals are "unexplained components, differ for each country, reflecting the extent to which the six variables either over- or under-explain average [...] life evaluations".

	Country	Happiness.Score	GDP.per.Capita	Social.support	Life.expectancy	Freedom	Generosity	Corruption	Dystopia.Residual
1	Norway	7.537	1.61646318	1.5335236	0.796666503	0.63542259	0.36201224	0.315963835	2.2770267
2	Denmark	7.522	1.48238301	1.5511216	0.792565525	0.62600672	0.35528049	0.400770068	2.3137074
3	Iceland	7.504	1.48063302	1.6105740	0.833552122	0.62716264	0.47554022	0.153526559	2.3227153

Figure 2.1: World Happiness Score for the first 3 countries of 2017

But those factors did not play a role in our analysis. Since we could not include all of these given factors in our analysis (due to the required huge amount of data) we tried to cover them roughly by the World Happiness Score.
We decided to use the data from the WHI report of 2017 on Kaggle since we found the most amount of data regarding usage of social media in each country for that particular year.
In this approach we wanted to take a closer look at countries with a similar happiness index while focusing on four certain variables. Thus we collected data sets regarding the usage of social media in each country and the corresponding depression rate, anxiety prevalence's and ADHD prevalence's, all from the year 2017. However, since the ADHD and the anxiety incidents in each country were given as prevalence's we had to divide the prevalence's by the population to have rates so that the data could be in the same format we could compare them.
Within this approach we had two different methods to look closer at a correlation between the mental illnesses and the usage of social media.
In the first method we selected all the countries with a happiness score between 6 and 7. Firstly, we wanted to take a closer look at countries with similar happiness indexes without making the span too big. Otherwise, this attempt of eliminating confounders would not make any sense.

Secondly, most of the countries had an index between 6 and 7. This is why we chose the range 6-7 and not 5-6, for example. The last step for preparing the data for the visualisation was to merge the resulting 14 countries with the given mental disorder and social media usage rates into a single table (Table 2.1).
In this method we visualized the data in 3 different scatter plots where the y-axis shows the mental illness category, while the x-axis shows the social media usage rate.

Country	**Social_Media_Usage**	**DEPRESSION_Rate**	**ANXIETY_Rate**	**ADHD_Rate**
Argentina	70	3.665488	5.9945073	1.45568760
Austria	47	3.260970	0.5243249	0.07030494
Brazil	58	3.297368	6.3512641	1.25477264
France	56	4.253807	6.1875270	0.74380756
Germany	41	3.959866	6.3808133	0.26317547

Table 2.1: Variables used for the second approach

Within the second approach we conducted a second methodology. Instead of defining a scope to choose countries from for further analysis, we this time created country pairs with a similarly Happiness Score. We did this to put the countries on the same level and eliminate the confound to make them more comparable.

The country pairs are:

- Nigeria and Vietnam (Happiness Score: 5.9)
- Brazil and United Arab Emirates (Happiness Score: 6.6)
- Italy and Russia (Happiness Score: 5.0)
- Austria and USA (Happiness Score: 7.0)
- Mexico and Singapore (Happiness Score 6.6)

Chapter 3

Results

3.1 Results of the first approach: Depression rate and social media usage in the USA

There were only two data sets to analyze for the first approach, so the final results consist of two separate bar charts representing the social media usage and the depression rate in the USA for different age groups.
The figures 3.1 and 3.2 show the visualised results after cleaning and analysing the data sets.

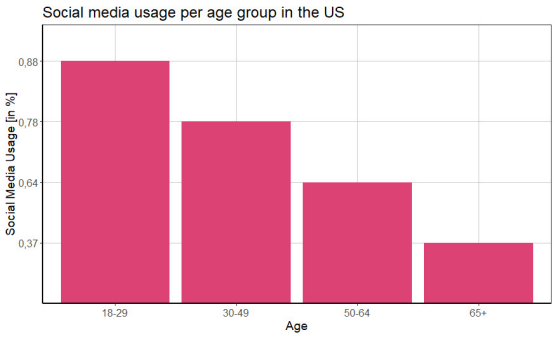

Figure 3.1: Social media usage in the USA

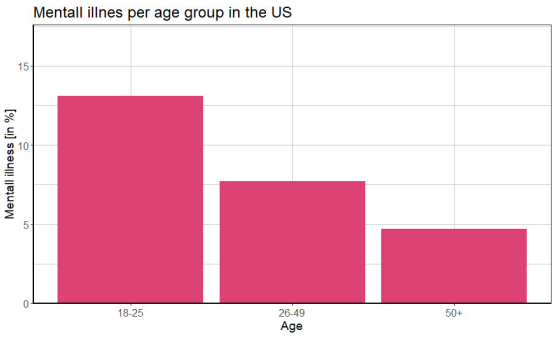

Figure 3.2: Depression in the USA

Since the age groups are different in the data sets, we came to the conclusion that the comparison between the data sets was not possible. However, it was still possible to see the pattern, that young people use social media and suffer at the same time from depression more than other respondents. This conclusion is not significant by itself, but it could be an additional argument for the case the second approach would show the same development of the results.

3.2 Results of the second approach: Mental disorders in view of social media usage in different countries

1. Method: Mental illness rate of the country span

In figure 3.3 is the expected curve course of the linear regression line regarding all mental illnesses (in this example the ADHD rate). We expected a pretty high positive correlation between the mental disorder rate and the usage of social media in a country, which means a higher social media usage in a country should lead to a higher mental illness rate. But our expectations did not match the actual results.

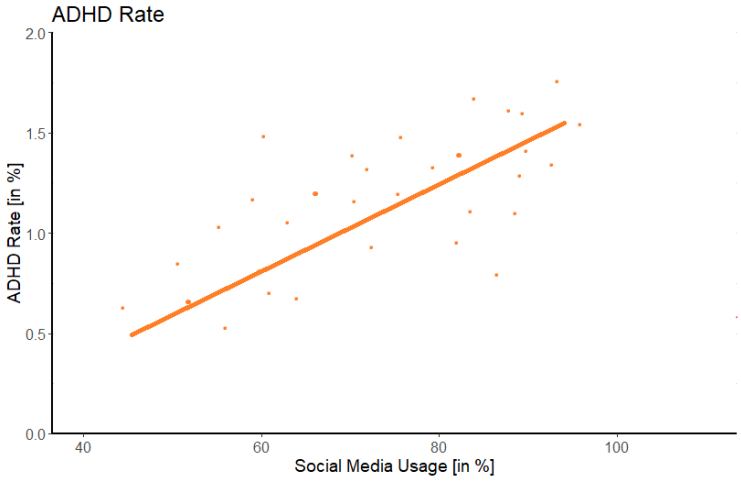

Figure 3.3: Expected ADHD curve course

The actual curve course shown in figure 3.4 was different from the expected curve course. There are statistical outliers such as the United Arab Emirates (with a social media usage rate of 99% and an ADHD rate of nearly 0.5%), which lead to a result we cannot be very certain about. This is the case due to the confidence interval represented by the grey boundaries. If we were to repeat the sampling over and over, in this case, 95% of the regression lines would be in that grey zone. But in this case the regression lines could also be very different with a different slope, due to the space and large span of the confidence interval. However, looking at the regression line there seems to be almost no relationship between the ADHD rate and the usage of social media.

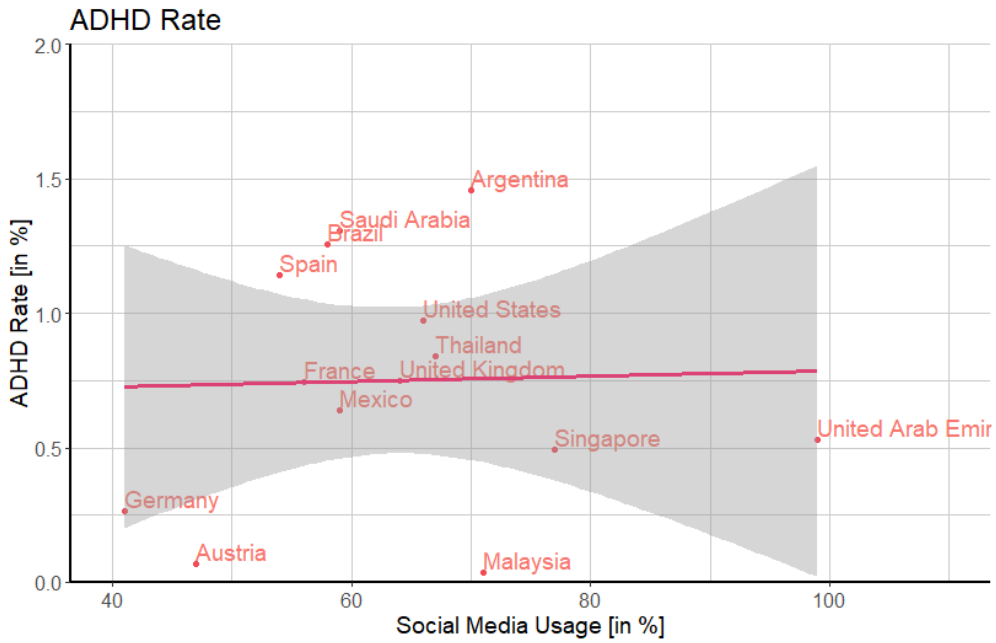

Figure 3.4: ADHD curve course in view of social media usage

The same issue applies in the second graph in figure 3.5. There is a little negative relationship between the anxiety rate and the social media usage, but again with a large confidence interval scope.

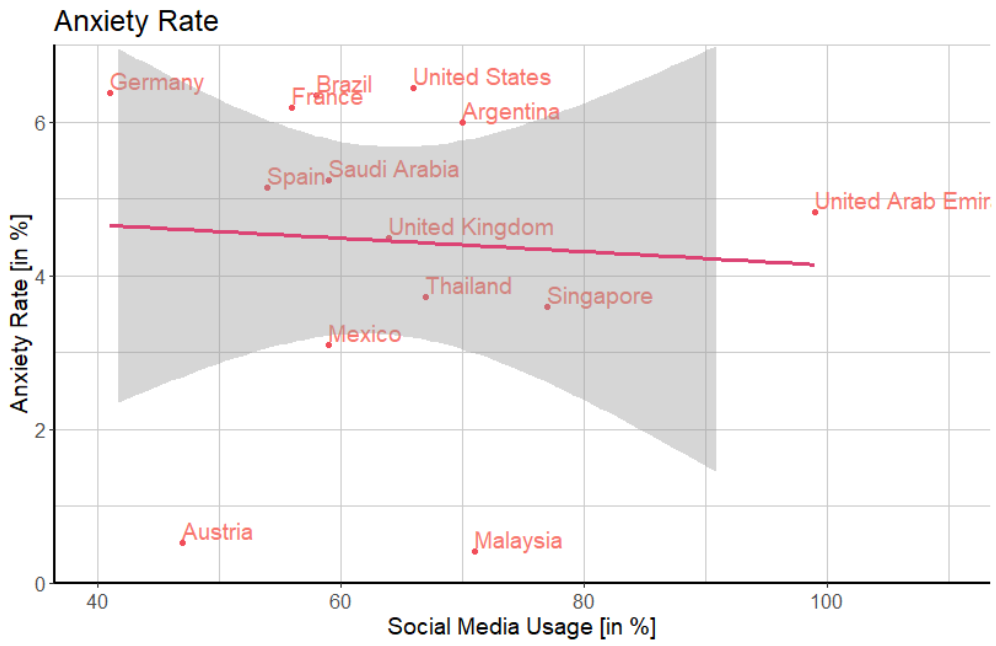

Figure 3.5: Anxiety curve course in view of social media usage

Nevertheless, we got a preciser regression line regarding the depression rate (figure 3.6). Almost all countries have a depression rate between 3 - 4.5%. This time there were not too many outliers which led to a confidence interval with a small span, so we can be more certain with the correlation coefficient.

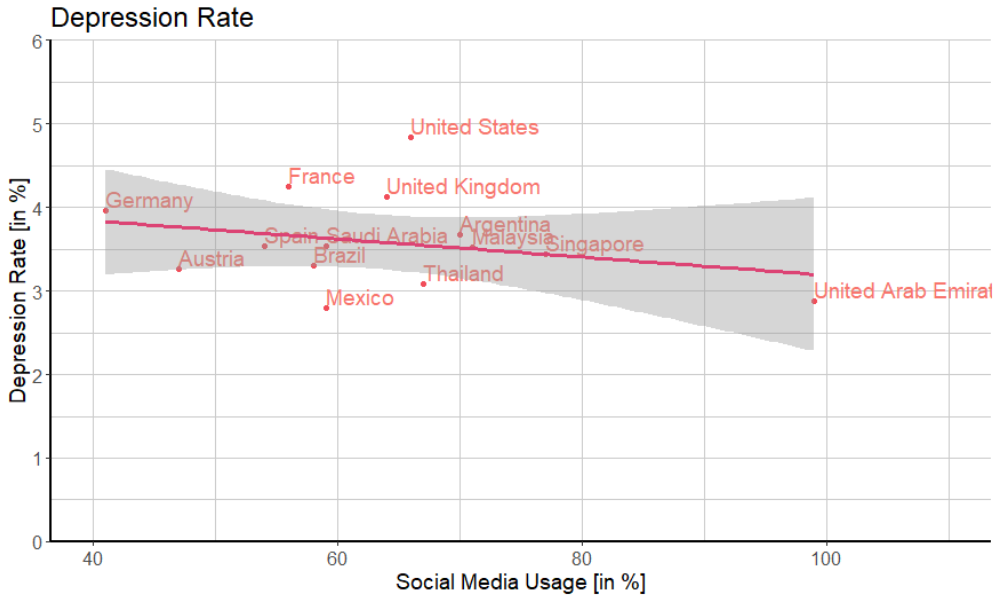

Figure 3.6: Depression curve course in view of social media usage

The computed correlation coefficients for all mental illnesses are:

- R(Social media usage; ADHD rate) = 0.0311
- R(Social media usage; Anxiety rate) = -0.061
- R(Social media usage; Depression rate) = -0.2723

Even if the correlation coefficient of social media usage and depression rate is negative, which means that higher social media usage leads to a lower depression rate, we cannot say radically that -0.27 indicates a high negative correlation. Furthermore, the coefficients of the other 2 mental illnesses and social media usage are almost 0.
All in all there is approximately no correlation between the usage of social media and the 3 mental disorders, with regards to our data sets and the country span method.

2. Method: Mental illness rate of the country pairs

Like in the first method we have expected a high positive correlation between the mental disorder rates and the usage of social media in a country. In figure 3.5, 3.6, 3.6 it is not feasible to see a pattern. The country pairs are plotted completely with no structure and give no insight in any correlation.

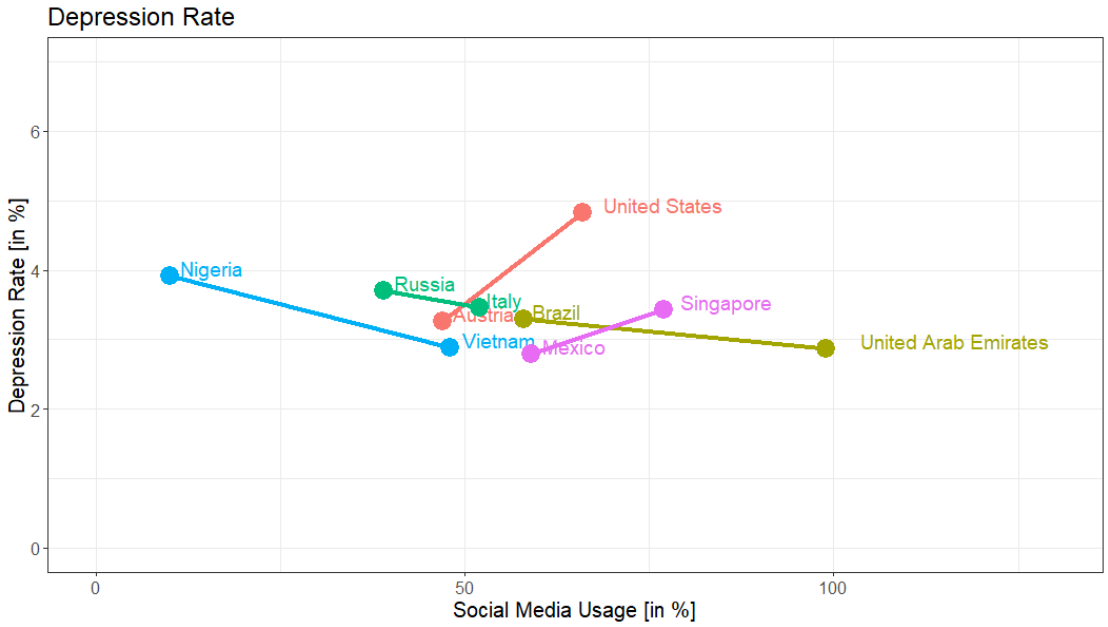

Figure 3.7: Grouped Plot for Depression Rate

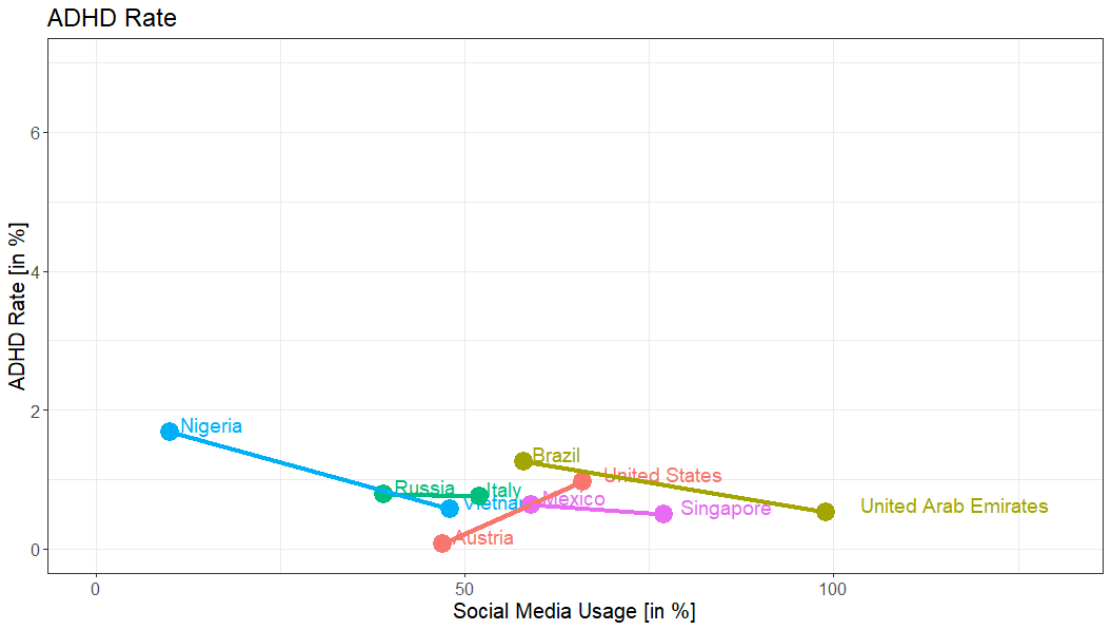

Figure 3.8: Grouped Plot for ADHD Rate

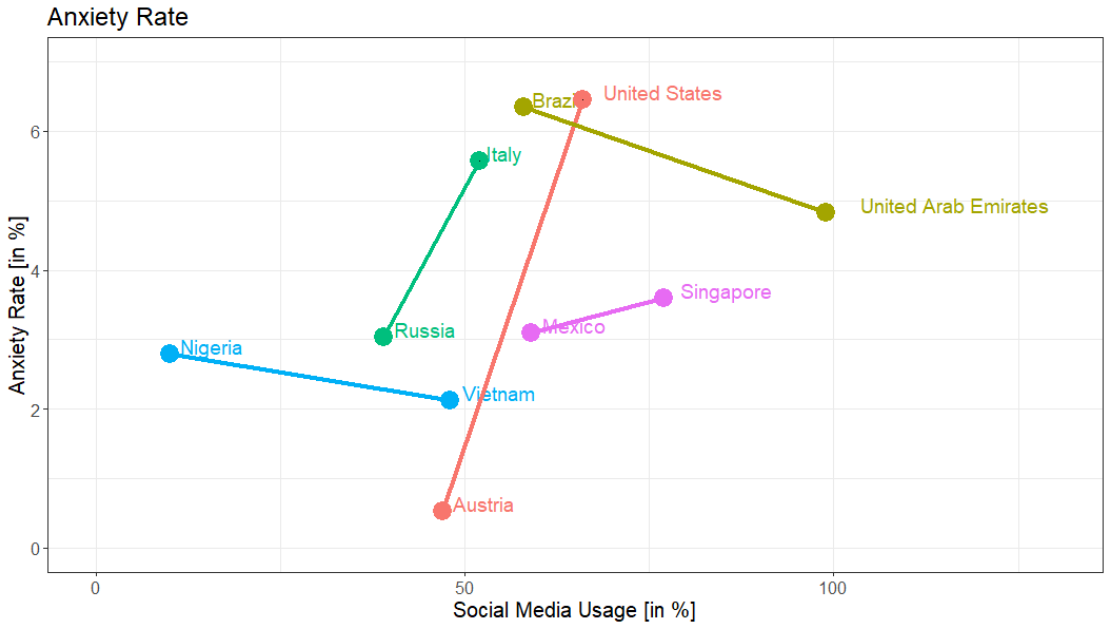

Figure 3.9: Grouped Plot for Anxiety Rate

Interesting to see is that some pairs kept its structure, i.e. Vietnam was always lower than Nigeria, United Arab Emirates was always lower than its counterpart Brazil and Austria was always lower than the United States. So there have to be factors we did not consider that would have an impact on the rates.

Chapter 4

Discussion

4.1 Evaluation

Before we started with the project we made the assumption that in the last few years the cases of social media and mental disorders rose. This eventually led to our hypothesis that usage of social media increases mental disorders. To test this hypothesis, we had two approaches.
In the first approach in which we used "age" as the key variable it was challenging to find data for both of our topics, social media and mental disorders. The data had to fulfill the following three criteria.

1. The data had to include "age" as a column or a row.
2. The data had to be relatively current.
3. The data had to be from the same year.

After searching the Internet for data that fulfills our criteria we only found suitable data sets for the United States of America. One country is still just not enough to make a general conclusion. Furthermore the age groups are not comparable. This is seen when comparing the age groups of the depression prevalence data 4.1 and usage of social media data 4.2 of the US.

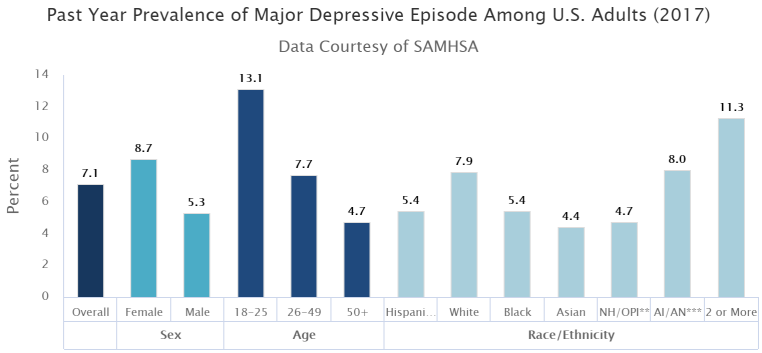

Figure 4.1: Data set regarding mental depression rates in the US

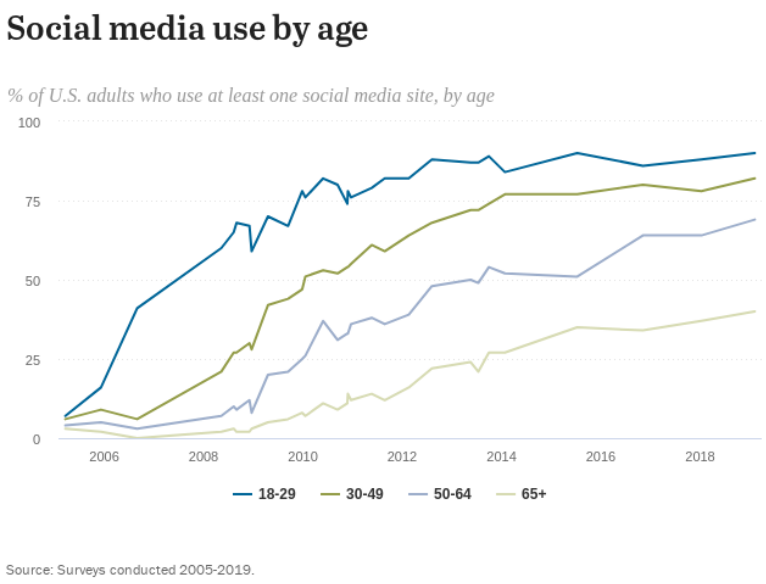

Figure 4.2: Data set regarding social media usage in the US

In the second approach in which we used the World Happiness Index as the key variable we came to the conclusion that the correlation coefficients for ADHD, anxiety and depression in relation to social media usage are approximately zero. This means that there is almost no correlation between the two variables in each of the three cases. However, there could be multiple factors that prevent us from seeing the correlation. Firstly, the World Happiness Index is subjective since differences between cultures exist. For example, some nationalities have higher expectations which need to be fulfilled so that they are happy.
Another reason could be that the pressure of fellow citizens influences the opinion given to pollsters. For instance, in Scandinavia people are socially pressured to be happy.
In addition, the surveyed group consisted only of 1000 people. This is just simply not representative enough to make a general conclusion.
Finally, another reason may be that the World Happiness Index is simply not the right index to calculate the confounders since it does not take factors like the political or weather conditions into account. You can also be a very happy person having anxiety issues. To get the best possible data set to be able to answer the research question it would require a dedicated survey, which may only be possible in about 10 years when there is enough data available and we are able to distinguish between the age groups that either use or do not use social media.

4.2 Conclusions

Since we had difficulties finding data sets for mental disorders and social media usage during our first approach we suggest for further notice to apply field instead of desk research. In the latter the relevant respondents could be chosen directly by the data researchers and problems regarding different age groups from the first approach could be avoided. This would also make the data more comparable, more accurate and help

against misinformation or manipulation.
It could be helpful to connect just one disease with the social media usage. Making the research questions preciser makes also the results a bit more meaningful for all the questions.
It is also recommendable to split the project into different parts so the hypothesis could be proved in different ways. Concentrating on just one approach could lead to the situation in which the results could be falsely interpreted.

List of Contents

Table Directory

Literature

[Kar+20] Fazida Karim et al. “Social Media Use and Its Connection to Mental Health: A Systematic Review”. In: *Cureus* 12.6 (2020).