Trendy or Body Dysmorphia? Exploring the reasoning behind the increase in mental health illnesses amongst Generation Z

Project Proposal

1 Aims, objectives and background

1.1 Introduction ¶

What can be said about the generation that grew up connected to the internet? Why is Generation Z so depressed?
A quick google search would reveal the gross number of articles simply centered around this question. What are the implications of this discovery?

It is no secret that Gen Z is thought to be a troubled group. A study conducted by Mind Share Partners, SAP, and Qualtrics found that 75% of Gen Zers had left job roles in the past due to mental health issues. [1] Poor mental health is on the rise amongst this young generation, and I'll be delving into whys and hows of this terrible phenomenom in this project.

Though I'll touch on mental health in general, I will single out anxiety, body dysmorphia disorder, and depression as they are all points of contention, especially for those who fall under the bracket of Generation Z, and seems to be a ramification of growing up tied to technology.

I will begin by exploring the relationship between the internet and Generation Z. What are they using the internet for. Whether social media has replaced traditional forms of media and news. How integreted has social media become with their lives, and whether this is connected to the increase in mental illnesses.

1.2 Aims and objectives

I will be exploring the following objectives in my project proposal:

- **The impact of a media influenced and technology driven world on Gen Z.**
  - Measuring the forms of media that Gen Z take in, is it considered traditional?
  - The way in which traditional media forms have adapted.
  - Does media cause people to judge theirselves more unfairly.
  - Is there a link between actively taking in media and body dysmorphia.
- **Living with social media and its ramifications.**
  - How often do people use social media.
  - Are Gen Zers using the internet mostly for social media.
  - Is there a relationship between the increase in social media, and the increase in poor mental health?

The aims for this project are the following:

1. Find the data needed to explore my objectives.

2. Consider the methods needed to extract said data.
3. Once decided on the best approach, extract the data.
4. Clean the acquired data so that it can be used for data analysis.
5. Carry out exploratory data analysis, identifying links between data and potential flaws in the data.

1.3 Data

1.3.1 Data requirements

Finding revelant data for my topic of study was challenging. I found that searching specifically for a connection between body dysmorphia and social media within Generation Z was a bit of a wild goose chase. Everytime I finally found some piece of data that I thought would fit the bill, it either wasn't specific enough, or it was licensed. There have been a lot of articles on the use of social media, but not on the impact of growing up submerged in it, or how a person's view of theirselves could have been entired built or shifted based of images they were fed growing up.

After endless searching, I finally stumbled across some data that could be useful, though it didn't go as deep as I would have liked. I then decided that it would be better to broaden the basis of my proposal. Instead of picking a very niched topic, I decided that I would expand to mental illnesses in general, and look on the impact of technology as whole. Though I will say, finding appropriate data, still felt like pulling teeth.

I decided to extract data from multiple sources, and deduce whether there was a pattern, and whether I could formulate a conclusion from it.

1.3.2 Limitations and constraints of the data

A lot of the data that I found that could be useful was locked, and required paid memberships to use. I had originally set out to find the perfect dataset specifically for what I was looking for. I found many articles on social media having an impact on people, but there weren't many that focused on giving statistics alongside their views, nor did they go into detail about body dysmorphia. I had hoped on honing in specifically on body dysmorphia and its connection with social media among Generation Z. It is something that I had noticed, and if I had more time on my hands, I would certainly dive into that topic specifically, and perhaps conduct my own study/survey so that I could get some raw trusted data.

1.4 Ethical considerations

When it came to data acquisition I knew I had to be very careful. The way I see it is this: just because I can take it, it does not mean that I should take it. When finding data I had to ensure that I was following guidelines as best as I could. Data use guidelines varies from country to country, and there are many different licenses - therefore it is awfully easy for the lines to get blurred. In addition, the internet can only be policed to a certain extent, so it leaves the door open to moral conflict.

When starting a prosposal in which I acquire and analyse some data, I had to be mindful of its reach. Who was going to have access to my analysis, and what would they do with it. What were there intentions? Could my exploratory data analysis contain sensitive information? How would someone feel to see their personal data being used and deconstructed on the internet?

My first question was whether I was allowed to take the data. I extracted some data from The Annie E. Casey Foundation, who are a "*charitable foundation focused on improving the well-being of American children according to their ideals.*" [3]. Due to their drive to improve the livelihood of american children, I was allowed to obtain this data, as their organisation aims to educate, and by sharing data they hope that they can do just that. Is the data as thorough as the data that would come from another site, which isn't very clear about their licensing, or who impose restrictions- no. That being said, it all boils down to the principle, yes the data is a lot more extensive- but the lines around extracting it are blurry- I don't want to extract data that is restricted.

I also questioned the bias of the data that I had acquired. Was the study large enough? Was it diverse? If not, how could it represent an entire population that is millions of time larger and very much diverse? I mentioned earlier that it was hard to find data that fit my proposal topic. In addition, it was hard to find data from a large study. This poses the question, can the data I have acquired be valid? Shouldn't it be deemed analagous due to its small nature, perhaps only that small group of people felt that way, what if the study was double the size, will the results be completely different? I suggested diversity above, could a lack of diversity introduce bias, and what impact does this have on people, or on an organisation.

2 Data acquisition and Data cleaning

Below are the libraries that I will be using for this proposal.

In [23]:

```
# import libraries
import pandas as pd
from bs4 import BeautifulSoup
import html5lib
import matplotlib
import matplotlib.pyplot as plt
import seaborn as sns
```

2.1 Web scrapping and cleaning

Table of URLs for data acquisition, data scrapping, data cleaning and their corresponding index
This is a table of content for links to the places in which I obtained data. The index is very important, as I use them to refer to different sources, as the title of some of the datasets are super long, and hard to remember (I also paraphrase a number of the titles). The index also refers to and are linked to some of the created files, though that will be made clear throughout the proposal.

Index	Name	URL
0	Frequency of using selected news sources among Generation Z in the United States as of February 2022	https://www.statista.com/statistics/1124119/gen-z-news-consumption-us/ (https://www.statista.com/statistics/1124119/gen-z-news-consumption-us/)
1	Number of Gen Z users in the United States on selected social media platforms from 2020 to 2025	https://www.statista.com/statistics/1276021/instagram-snapchat-tiktok-gen-z-users/ (https://www.statista.com/statistics/1276021/instagram-snapchat-tiktok-gen-z-users/)
2	Activities performed online by Generation Z in Great Britain in 2020	https://www.statista.com/statistics/1119977/gen-z-internet-activities-in-great-britain/ (https://www.statista.com/statistics/1119977/gen-z-internet-activities-in-great-britain/)

Index	Name	URL
3	Number of Gen Z that suffer from nerves, anxiety, and more in the US in 2020	https://datacenter.kidscount.org/data/tables/11209-adults-ages-18-to-24-who-felt-nervous-anxious-or-on-edge-for-more-than-half-of-the-days-or-nearly-every-day-in-the-past-week (https://datacenter.kidscount.org/data/tables/11209-adults-ages-18-to-24-who-felt-nervous-anxious-or-on-edge-for-more-than-half-of-the-days-or-nearly-every-day-in-the-past-week)
4	Number of Gen Z that are depressed and hopeless in the US in 2021	https://datacenter.kidscount.org/data/tables/11211-adults-ages-18-to-24-who-felt-down-depressed-or-hopeless-for-more-than-half-of-the-days-or-nearly-every-day-for-the-past-two-weeks (https://datacenter.kidscount.org/data/tables/11211-adults-ages-18-to-24-who-felt-down-depressed-or-hopeless-for-more-than-half-of-the-days-or-nearly-every-day-for-the-past-two-weeks)
5	Number of Gen Z reporting poor mental health	https://datacenter.kidscount.org/data/tables/11202-young-adults-ages-18-to-24-reporting-zero-poor-mental-health-days-in-the-past-month (https://datacenter.kidscount.org/data/tables/11202-young-adults-ages-18-to-24-reporting-zero-poor-mental-health-days-in-the-past-month)
6	The link between social media and body dysmorphia	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8001450/ (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8001450/)

List of URLs to scrape

Below are links to websites that I scrapped. The index in the list below matches that of the table above for convenience.

In [3]:

```
URL = ["https://www.statista.com/statistics/1124119/gen-z-news-consumption-us/",
       "https://www.statista.com/statistics/1276021/instagram-snapchat-tiktok-gen-z-
       "https://www.statista.com/statistics/1119977/gen-z-internet-activities-in-gre
       "https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8001450/"]
```

URL[0] | Selected new sources

Frequency of using selected news sources among Generation Z in the United States as of February 2022

This source is from a survey conducted in February 2022, and it records where Generation Z news consumers frequently get their news from.

I had to create an account to access this data, and there is a limit to how many databases you have access to on this site. The next two sources are also from this site, and that was the cap on the number of databases I could access. However, I am allowed full access to this dataset, and the next two.

I was originally put off from this source, because of the idea of creating an account. However I quickly changed my mind, as I realised that beggers can't be choosers, and finding this dataset was like obtaining gold dust. The site also had other databases that I was interested in, that fit my proposal topic well, but I had to find other sources. I also felt as though it was important to show different techniques for data acquisition. If I had gotten all my data from this site, I wouldn't be able to display that, or it would get very repetitive.

This dataset is all about what sources Generation Zers use to obtain news. This was very much needed, as it helped to show just how integrated social media is with their lives, as it is their primary news source- though I will delve into that later.

I first began by scrapping the data. I first needed to see what I was to work with.

In [4]:

```
df_0 = pd.read_html(URL[0], match=".+", header = 0)
df_0
```

Out[4]:

```
[           Characteristic Daily A few times per week Once per week  \
 0            Social media   50%                  18%             9%
 1                   Radio   17%                  17%             8%
 2  Online-only news sites   13%                  18%            14%
 3                Podcasts   13%                  14%             7%
 4            Network news    9%                  13%            12%
 5     Cable news networks    8%                  14%             9%
 6               Newspapers    5%                   4%             9%

   A few times per month Once per month Less than once per month Never
 0                   10%             3%                       4%    6%
 1                    8%             6%                      13%   31%
 2                   11%             6%                       9%   29%
 3                   11%             5%                      13%   38%
 4                    9%             6%                      12%   39%
 5                    7%             6%                      12%   44%
 6                    6%             6%                      17%   53%
,
                                        Unnamed: 0 Basic Account  \
 0  Our services  Click the question mark for furt...          Free
 1                                 Basic statistics           NaN
 2                               Premium statistics           NaN
 3                                              NaN           NaN
 4                                  Graph download           NaN
 5                                    Download PDF           NaN
 6                                  Excel download           NaN
 7                                              NaN      Register

       Instant Access  Single Account  Unnamed: 3
 0  $39 per month*  (billed annually)         NaN
 1                               NaN         NaN
 2                               NaN         NaN
 3                               NaN         NaN
 4                               NaN         NaN
 5                               NaN         NaN
 6                               NaN         NaN
 7                     Purchase now          NaN  ]
```

The data looked a little messy, so I decided to take a deeper look at what it was made up of.
Checked how many tables there were.

In [5]:

```
len(df_0)
```

Out[5]:

```
2
```

Took a look at the first table.

In [6]:

```
df_0[0]
```

Out[6]:

	Characteristic	Daily	A few times per week	Once per week	A few times per month	Once per month	Less than once per month	Never
0	Social media	50%	18%	9%	10%	3%	4%	6%
1	Radio	17%	17%	8%	8%	6%	13%	31%
2	Online-only news sites	13%	18%	14%	11%	6%	9%	29%
3	Podcasts	13%	14%	7%	11%	5%	13%	38%
4	Network news	9%	13%	12%	9%	6%	12%	39%
5	Cable news networks	8%	14%	9%	7%	6%	12%	44%
6	Newspapers	5%	4%	9%	6%	6%	17%	53%

Took a look at the second table.

In [7]:

```
df_0[1]
```

Out[7]:

	Unnamed: 0	Basic Account	Instant Access Single Account	Unnamed: 3
0	Our services Click the question mark for furt...	Free	$39 per month* (billed annually)	NaN
1	Basic statistics	NaN	NaN	NaN
2	Premium statistics	NaN	NaN	NaN
3	NaN	NaN	NaN	NaN
4	Graph download	NaN	NaN	NaN
5	Download PDF	NaN	NaN	NaN
6	Excel download	NaN	NaN	NaN
7	NaN	Register	Purchase now	NaN

I decided that I would not use the second table, as it didn't contain any data valuable for analysis.
Reassigned the dataframe to only contain the first table.

In [8]:

```
df_0 = df_0[0]
```

In [9]:

```
df_0
```

Out[9]:

	Characteristic	Daily	A few times per week	Once per week	A few times per month	Once per month	Less than once per month	Never
0	Social media	50%	18%	9%	10%	3%	4%	6%
1	Radio	17%	17%	8%	8%	6%	13%	31%
2	Online-only news sites	13%	18%	14%	11%	6%	9%	29%
3	Podcasts	13%	14%	7%	11%	5%	13%	38%
4	Network news	9%	13%	12%	9%	6%	12%	39%
5	Cable news networks	8%	14%	9%	7%	6%	12%	44%
6	Newspapers	5%	4%	9%	6%	6%	17%	53%

Created the following function to make this process instantaneous.

In [10]:

```
def import_df_url(url, x):
    temp = pd.read_html(url, match=".+", header = 0)
    return temp[x]
```

Reviewing the table, led me to realise that 'Characteristic' was not the best way to the describe the column that represented the various sources from which the Gen Zers could consume media.

Therefore I renamed Characteristic to Source, as it better describes the column.

In [11]:

```
df_0 = df_0.rename(columns={'Characteristic': 'Source'})
df_0
```

Out[11]:

	Source	Daily	A few times per week	Once per week	A few times per month	Once per month	Less than once per month	Never
0	Social media	50%	18%	9%	10%	3%	4%	6%
1	Radio	17%	17%	8%	8%	6%	13%	31%
2	Online-only news sites	13%	18%	14%	11%	6%	9%	29%
3	Podcasts	13%	14%	7%	11%	5%	13%	38%
4	Network news	9%	13%	12%	9%	6%	12%	39%
5	Cable news networks	8%	14%	9%	7%	6%	12%	44%
6	Newspapers	5%	4%	9%	6%	6%	17%	53%

In [12]:

```
type(df_0['Daily'][0])
```

Out[12]:

```
str
```

The data in the dataframe were strings, but in order to further process the data, I wanted them to be in integers. I wanted to also remove the percentage sign, and instead indicate the they were percentages, by adding the percentage symbol to the headers.

First I created a remove character function.
This was so that if I needed to do another process like this again, I could reuse my code.

In [13]:

```
def remove_characters(values, char):
    stripped = []

    for i in range(0, len(values)):
        stripped.append(int(values[i].strip(char)))
    return stripped
```

Next I created an insert columns function.
Again, this was so that I could reuse the code, and to make the code a little shorter.

In [14]:

```
def insert_columns(index, column, values, dataf):
    dataf.insert(index, column, values, True)
```

Finally I created a delete columns function.

For the same reasons as above.

In [15]:

```
def delete_columns(columns, dataf):
    dataf.drop(columns, inplace=True, axis=1)
```

There were originally 8 columns.

In [16]:

```
df_0.shape
```

Out[16]:

```
(7, 8)
```

First I created the new columns.

I called both the **insert_columns()** function and the **remove_characters()** function.

In [17]:

```
before = ['Daily', 'A few times per week', 'Once per week', 'A few times per month',
          'Once per month', 'Less than once per month', 'Never']

after = ['Daily (%)', 'A few times per week (%)', 'Once per week (%)', 'A few times
          'Once per month (%)', 'Less than once per month (%)', 'Never (%)']

for i in range((len(before) - 1), -1, -1):
    insert_columns(1, after[i], remove_characters(df_0[before[i]], '%'), df_0)

df_0
```

Out[17]:

	Source	Daily (%)	A few times per week (%)	Once per week (%)	A few times per month (%)	Once per month (%)	Less than once per month (%)	Never (%)	Daily	A few times per week	Once per week	A few times per month
0	Social media	50	18	9	10	3	4	6	50%	18%	9%	10%
1	Radio	17	17	8	8	6	13	31	17%	17%	8%	8%
2	Online-only news sites	13	18	14	11	6	9	29	13%	18%	14%	11%
3	Podcasts	13	14	7	11	5	13	38	13%	14%	7%	11%
4	Network news	9	13	12	9	6	12	39	9%	13%	12%	9%
5	Cable news networks	8	14	9	7	6	12	44	8%	14%	9%	7%
6	Newspapers	5	4	9	6	6	17	53	5%	4%	9%	6%

In [18]:

```
df_0.shape
```

Out[18]:

```
(7, 15)
```

Now there was 15 columns.

Next I deleted the columns containing the old data.

I am called the **delete_columns()** function.

In [19]:

```
for i in range(0, len(before)):
    delete_columns([before[i]], df_0)
df_0
```

Out[19]:

	Source	Daily (%)	A few times per week (%)	Once per week (%)	A few times per month (%)	Once per month (%)	Less than once per month (%)	Never (%)
0	Social media	50	18	9	10	3	4	6
1	Radio	17	17	8	8	6	13	31
2	Online-only news sites	13	18	14	11	6	9	29
3	Podcasts	13	14	7	11	5	13	38
4	Network news	9	13	12	9	6	12	39
5	Cable news networks	8	14	9	7	6	12	44
6	Newspapers	5	4	9	6	6	17	53

In [20]:

```
df_0.shape
```

Out[20]:

```
(7, 8)
```

Now there was 8 columns again.

I then wrote the above data into a csv file, which is saved in the data folder. This was so that there was a concrete copy of the altered data.
Saving it in the data folder meant that all the data acquired is in one place and is easy to locate.

Wrote the table to a csv file

In [21]:

```
# Setting to false stops the index from being written out
df_0.to_csv('./data/df_0.csv', index=False)
```

Checked whether the file was created and written to successfully

In [52]:

```
cd data
```

```
/Users/worthy/Desktop/PWD/midterm/data
```

In [53]:

```
!ls
```

```
df_0.csv df_1.csv df_2.csv
```

In [54]:

```
!cat df_0.csv
```

```
Source,Daily (%),A few times per week (%),Once per week (%),A few time
s per month (%),Once per month (%),Less than once per month (%),Never
(%)
Social media,50,18,9,10,3,4,6
Radio,17,17,8,8,6,13,31
Online-only news sites,13,18,14,11,6,9,29
Podcasts,13,14,7,11,5,13,38
Network news,9,13,12,9,6,12,39
Cable news networks,8,14,9,7,6,12,44
Newspapers,5,4,9,6,6,17,53
```

In [55]:

```
cd -
```

```
/Users/worthy/Desktop/PWD/midterm
```

The file was successfully created!

URL[1] | Popular social platforms

Number of Gen Z users in the United States on selected social media platforms from 2020 to 2025

This source records which social media platforms were most popular amongst Generation Z. (The data is in millions).

This was the second dataframe that I had access to from the I account that I had to create. This data was all about what social apps people used the most. This was a good dataset as the study size was huge, adding a greater amount of accuracy. This source was needed to draw a conclusion from the most popular social apps.

by concluding why they were the most popular, and what that revealed about Generation Z.

I again began by scrapping the data.

I set the database equal to the first table instantaneously using the import_df_url() function

In [39]:

```
df_1 = import_df_url(URL[1], 0)
```

In [40]:

```
df_1
```

Out[40]:

	Characteristic	Snapchat	Instagram	TikTok	Facebook	Pinterest	Twitter	Reddit
0	2025*	49.5	48.3	48.8	40.7	26.1	19.2	17.1
1	2024*	49.6	45.7	48.2	38.3	25.1	18.8	15.4
2	2023*	47.0	41.4	45.0	35.0	23.0	17.8	13.6
3	2022*	44.5	37.3	41.4	31.8	20.9	16.7	11.9
4	2021*	42.0	33.3	37.3	28.7	18.9	15.8	10.3
5	2020	38.1	30.0	29.5	26.8	16.7	14.4	8.4

I renamed 'Characteristic' to 'Year' to properly represent the column.

In [41]:

```
df_1 = df_1.rename(columns={'Characteristic': 'Year'})
df_1
```

Out[41]:

	Year	Snapchat	Instagram	TikTok	Facebook	Pinterest	Twitter	Reddit
0	2025*	49.5	48.3	48.8	40.7	26.1	19.2	17.1
1	2024*	49.6	45.7	48.2	38.3	25.1	18.8	15.4
2	2023*	47.0	41.4	45.0	35.0	23.0	17.8	13.6
3	2022*	44.5	37.3	41.4	31.8	20.9	16.7	11.9
4	2021*	42.0	33.3	37.3	28.7	18.9	15.8	10.3
5	2020	38.1	30.0	29.5	26.8	16.7	14.4	8.4

Looking at the data above, I saw that the 'Year' column had some weird data type.
I then ran the following code to learn the type.

In [42]:

```
type(df_1['Year'][0])
```

Out[42]:

```
str
```

The data was in the form of a string, but I wanted it in the form of an integer.
Therefore I ran the following code:

In [43]:

```
update_df_1 = pd.DataFrame({'Year': [2025, 2024, 2023, 2022, 2021, 2020]})

df_1.update(update_df_1)
df_1
```

Out[43]:

	Year	Snapchat	Instagram	TikTok	Facebook	Pinterest	Twitter	Reddit
0	2025	49.5	48.3	48.8	40.7	26.1	19.2	17.1
1	2024	49.6	45.7	48.2	38.3	25.1	18.8	15.4
2	2023	47.0	41.4	45.0	35.0	23.0	17.8	13.6
3	2022	44.5	37.3	41.4	31.8	20.9	16.7	11.9
4	2021	42.0	33.3	37.3	28.7	18.9	15.8	10.3
5	2020	38.1	30.0	29.5	26.8	16.7	14.4	8.4

I reordered the data so that the year was in ascending order rather than descending order.

In [44]:

```
df_1_reorder = df_1.sort_values('Year', inplace = True)
df_1.update(df_1_reorder)
df_1
```

Out[44]:

	Year	Snapchat	Instagram	TikTok	Facebook	Pinterest	Twitter	Reddit
5	2020	38.1	30.0	29.5	26.8	16.7	14.4	8.4
4	2021	42.0	33.3	37.3	28.7	18.9	15.8	10.3
3	2022	44.5	37.3	41.4	31.8	20.9	16.7	11.9
2	2023	47.0	41.4	45.0	35.0	23.0	17.8	13.6
1	2024	49.6	45.7	48.2	38.3	25.1	18.8	15.4
0	2025	49.5	48.3	48.8	40.7	26.1	19.2	17.1

I wrote the table to a csv file

In [45]:

```
# Setting to false stops the index from being written out
df_1.to_csv('./data/df_1.csv', index=False)
```

I read the data to ensure that the changes were successfully met

In [46]:

```
df_1 = pd.read_csv('./data/df_1.csv')
```

In [47]:

```
df_1
```

Out[47]:

	Year	Snapchat	Instagram	TikTok	Facebook	Pinterest	Twitter	Reddit
0	2020	38.1	30.0	29.5	26.8	16.7	14.4	8.4
1	2021	42.0	33.3	37.3	28.7	18.9	15.8	10.3
2	2022	44.5	37.3	41.4	31.8	20.9	16.7	11.9
3	2023	47.0	41.4	45.0	35.0	23.0	17.8	13.6
4	2024	49.6	45.7	48.2	38.3	25.1	18.8	15.4
5	2025	49.5	48.3	48.8	40.7	26.1	19.2	17.1

URL[2] | Online activities

Activities performed online by Generation Z in Great Britain in 2020

This source records the activities carried out online by Generation Z. It was drawn from a survey conducted in 2020. This was the final dataset I had access to from the source I had to login to access.

I felt as though I wasn't able to draw the strongest conclusion from this data, as it was really ambigious. However, I go into more detail on this later.

I first began by using my created function to scrap the data.

In [50]:

```
df_2 = import_df_url(URL[2], 0)
df_2
```

Out[50]:

	Characteristic	Share of respondents
0	Social networking	97%
1	Using instant messaging services	92%
2	Listening to/downloading music	93%
3	Sending/receiving emails	94%
4	Internet banking	90%
5	Finding information about goods and services	84%
6	Reading online news, newspapers or magazines	78%
7	Making video or voice calls over the internet	70%
8	Uploading content created by you to a website ...	56%
9	Selling goods or services over the internet	22%
10	Watching video content from services such as Y...	95%

I renamed 'Characteristic' to 'Activity' to properly represent the column.

In [51]:

```
df_2 = df_2.rename(columns={'Characteristic': 'Activity'})
df_2
```

Out[51]:

	Activity	Share of respondents
0	Social networking	97%
1	Using instant messaging services	92%
2	Listening to/downloading music	93%
3	Sending/receiving emails	94%
4	Internet banking	90%
5	Finding information about goods and services	84%
6	Reading online news, newspapers or magazines	78%
7	Making video or voice calls over the internet	70%
8	Uploading content created by you to a website ...	56%
9	Selling goods or services over the internet	22%
10	Watching video content from services such as Y...	95%

I wanted to create 2 more columns in the above dataframe (I also removed one). I wanted the float version of

the percentages, as I believed that I can work more with floats. I also wanted a new column for the percentages where the values are not strings but integers, so that I could work with the data.

First I created the values that would go in the columns.

I utilised the **remove_characters()** function for this.

In [52]:

```
# sor stands for share of respondents
sor_percentage = remove_characters(df_2['Share of respondents'], '%')
sor_float = []

for i in range(0, len(sor_percentage)):
    sor_float.append(float(format(sor_percentage[i]/100, '.2f')))

print("Percentages  =", sor_percentage)
print("Floats       =", sor_float)
```

```
Percentages  = [97, 92, 93, 94, 90, 84, 78, 70, 56, 22, 95]
Floats       = [0.97, 0.92, 0.93, 0.94, 0.9, 0.84, 0.78, 0.7, 0.56, 0.
22, 0.95]
```

I then removed the 'Share of respondents' column.

I utilised the **delete_columns()** function for this.

In [53]:

```
delete_columns('Share of respondents', df_2)
df_2
```

Out[53]:

	Activity
0	Social networking
1	Using instant messaging services
2	Listening to/downloading music
3	Sending/receiving emails
4	Internet banking
5	Finding information about goods and services
6	Reading online news, newspapers or magazines
7	Making video or voice calls over the internet
8	Uploading content created by you to a website ...
9	Selling goods or services over the internet
10	Watching video content from services such as Y...

I then created the two new columns.

I utilised the **insert_columns()** function for this.

In [54]:

```
insert_columns(1, "Share of respondents (f)", sor_float, df_2)
insert_columns(2, "Share of respondents (%)", sor_percentage, df_2)
df_2
```

Out[54]:

	Activity	Share of respondents (f)	Share of respondents (%)
0	Social networking	0.97	97
1	Using instant messaging services	0.92	92
2	Listening to/downloading music	0.93	93
3	Sending/receiving emails	0.94	94
4	Internet banking	0.90	90
5	Finding information about goods and services	0.84	84
6	Reading online news, newspapers or magazines	0.78	78
7	Making video or voice calls over the internet	0.70	70
8	Uploading content created by you to a website ...	0.56	56
9	Selling goods or services over the internet	0.22	22
10	Watching video content from services such as Y...	0.95	95

I wrote the table to a csv file

In [55]:

```
# Setting to false stops the index from being written out
df_2.to_csv('./data/df_2.csv', index=False)
```

I read the data to ensure that the changes were successfully met

In [56]:

```
df_2 = pd.read_csv('./data/df_2.csv')
```

In [57]:

```
df_2
```

Out[57]:

	Activity	Share of respondents (f)	Share of respondents (%)
0	Social networking	0.97	97
1	Using instant messaging services	0.92	92
2	Listening to/downloading music	0.93	93
3	Sending/receiving emails	0.94	94
4	Internet banking	0.90	90
5	Finding information about goods and services	0.84	84
6	Reading online news, newspapers or magazines	0.78	78
7	Making video or voice calls over the internet	0.70	70
8	Uploading content created by you to a website ...	0.56	56
9	Selling goods or services over the internet	0.22	22
10	Watching video content from services such as Y...	0.95	95

URL[6] | Social media and body dysmorphia

This source was very helpful, however it required a lot sanitisation. I was able to create 4 dataframes from this one dataset.

The link between social media and body dysmorphia

I began by scrapping the data the traditional way, so that I could get a look at it as a whole.

In [62]:

```
temp = pd.read_html(URL[6], match=".+", header = 0)
temp
```

...

The data looked very messy, but I could deduce that I would be using the third table, so index 2.
I then used the import_df_url() function to attain the data in the third table.

In [63]:

```
records = import_df_url(URL[6], 2)
records
```

Out[63]:

	Unnamed: 0	Final Sample (n = 1331)	Final Sample (n = 1331).1	SCOFF– (n = 378)	SCOFF– (n = 378).1	SCOFF+ (n = 953)	SCOFF+ (= 953).
0	NaN	Mean or Number of Participants	Standard Deviation or Percentage	Mean or Number of Participants	Standard Deviation or Percentage	Mean or Number of Participants	Standar Deviatio c Percentag
1	SOCIODEMOGRAPHIC CHARACTERISTICS	NaN	NaN	NaN	NaN	NaN	Nal
2	Age	24.2	4.2	25.1	4.2	23.9	4.
3	NaN	NaN	NaN	NaN	NaN	NaN	Nal
4	Sex	NaN	NaN	NaN	NaN	NaN	Nal
5	Female	1300	97.7%	363	96.0%	937	98.3%
6	Male	31	2.3%	15	4.0%	16	1.7%
7	Studies level	NaN	NaN	NaN	NaN	NaN	Nal
8	Less than Level 12	71	5.3%	16	4%	55	6%
9	Level 12	229	17.2%	62	16%	167	18%
10	Level 12 + 2 years	208	15.6%	50	13%	158	17%
11	Level 12 + 3 (Bachelor's degree)	320	24.0%	89	24%	231	24%
12	Level 12 + 5 (Master's degree)	380	0.285	96	25%	284	30%
13	Degree over Level 12 + 5	123	0.092	65	17%	58	6%
14	SOCIAL MEDIA USE	NaN	NaN	NaN	NaN	NaN	Nal
15	Frequency	NaN	NaN	NaN	NaN	NaN	Nal
16	Max. once a day	64	5%	17	4%	47	5%
17	2 to 10 times a day	578	43%	194	51%	384	40%
18	10 to 20 times a day	439	33%	115	30%	324	34%
19	Over 20 times a day	250	19%	52	14%	198	21%
20	Time spent	NaN	NaN	NaN	NaN	NaN	Nal
21	Less than 1 h	232	17%	81	21%	151	16%
22	Between 1 and 5 h	1048	79%	289	76%	759	80%

	Unnamed: 0	Final Sample (n = 1331)	Final Sample (n = 1331).1	SCOFF– (n = 378)	SCOFF– (n = 378).1	SCOFF+ (n = 953)	SCOFF+ (= 953).
23	Over 5 h	51	4%	8	2%	43	5%
24	Body comparison	NaN	NaN	NaN	NaN	NaN	Nal
25	Never	33	2%	18	5%	15	2%
26	Seldom	114	9%	56	15%	58	6%
27	Sometimes	317	24%	130	34%	187	20%
28	Often	523	39%	133	35%	390	41%
29	Always	344	26%	41	11%	303	32%
30	Posting selfies	NaN	NaN	NaN	NaN	NaN	Nal
31	Never	457	34%	146	39%	311	33%
32	1 or 2 times a month	756	57%	199	53%	557	58%
33	Once a week	93	7%	24	6%	69	7%
34	3 to 4 times a week	18	1%	7	2%	11	1%
35	Daily	7	1%	2	1%	5	1%
36	EATING DISORDERS	NaN	NaN	NaN	NaN	NaN	Nal
37	EDI-BD	12.4	7.5	7.9	6.6	14.2	
38	NaN	NaN	NaN	NaN	NaN	NaN	Nal
39	EDI-DT	8.9	6	4.1	4.2	10.8	5.
40	NaN	NaN	NaN	NaN	NaN	NaN	Nal
41	Average BMI	22.3	4.2	22.2	3.5	22.3	4.
42	NaN	NaN	NaN	NaN	NaN	NaN	Nal
43	Categories of BMI	NaN	NaN	NaN	NaN	NaN	Nal
44	<17.5	96	7.2%	9	2.4%	87	9.1%
45	[17.5–25]	981	73.7%	306	81.0%	675	70.8%
46	≥25	254	19.1%	63	16.7%	191	20.0%

The data was still messy, but there was now less data to work with.
After scanning through the dataframe, I decided that I would be take different parts from the table and group them, to make different dataframes.

In [64]:

```
records.shape
```

Out[64]:

```
(47, 8)
```

There were 47 rows, and 8 columns. Though I worked with only two of the columns.
The first data that I extracted and attempted to group was the social media use. I split the data into two dataframes, one on the frequency of social media use. The other on the spent on social media.

I first extracted the data from the two columns that I was using. I then put them into two separate lists.

Column 1: Categories

In [65]:

```
categories = records['Unnamed: 0']
categories
```

Out[65]:

```
0                                    NaN
1     SOCIODEMOGRAPHIC CHARACTERISTICS
2                                    Age
3                                    NaN
4                                    Sex
5                                 Female
6                                   Male
7                          Studies level
8                   Less than Level 12
9                              Level 12
10                  Level 12 + 2 years
11    Level 12 + 3 (Bachelor's degree)
12      Level 12 + 5 (Master's degree)
13            Degree over Level 12 + 5
14                     SOCIAL MEDIA USE
15                            Frequency
16                     Max. once a day
17                  2 to 10 times a day
18                 10 to 20 times a day
19                  Over 20 times a day
20                            Time spent
21                         Less than 1 h
22                    Between 1 and 5 h
23                             Over 5 h
24                     Body comparison
25                               Never
26                              Seldom
27                           Sometimes
28                               Often
29                              Always
30                     Posting selfies
31                               Never
32                1 or 2 times a month
33                          Once a week
34                 3 to 4 times a week
35                               Daily
36                    EATING DISORDERS
37                              EDI-BD
38                                 NaN
39                              EDI-DT
40                                 NaN
41                         Average BMI
42                                 NaN
43                    Categories of BMI
44                               <17.5
45                            [17.5–25]
46                                ≥25
Name: Unnamed: 0, dtype: object
```

Column 2: Feedback

In [66]:

```
feedback = records['Final Sample (n = 1331)']
feedback
```

Out[66]:

```
0     Mean or Number of Participants
1                                NaN
2                               24.2
3                                NaN
4                                NaN
5                               1300
6                                 31
7                                NaN
8                                 71
9                                229
10                               208
11                               320
12                               380
13                               123
14                               NaN
15                               NaN
16                                64
17                               578
18                               439
19                               250
20                               NaN
21                               232
22                              1048
23                                51
24                               NaN
25                                33
26                               114
27                               317
28                               523
29                               344
30                               NaN
31                               457
32                               756
33                                93
34                                18
35                                 7
36                               NaN
37                              12.4
38                               NaN
39                               8.9
40                               NaN
41                              22.3
42                               NaN
43                               NaN
44                                96
45                               981
46                               254
Name: Final Sample (n = 1331), dtype: object
```

You can see that the data in categories is more so describing what the data in column 2 represents.

I then checked the length of the lists.
This was so that I could ensure that there was no errors in the copying of data, and so that I could see whether there was any missing data.

In [67]:
```
print(len(categories), len(feedback))
```

```
47 47
```

Since the lists were equal in length, I moved on.

I began by creating a function that would extract data from either the categories list, or the column list, based of the inputs it was given. This saved me time, as I didn't need to continuosly write the same thing.

The create data function

In [68]:
```
def create_data(start, end, char='c'):
    if char == 'c':
        temp = [categories[x] for x in range(start, end)]
    elif char == 'f':
        temp = [feedback[x] for x in range(start, end)]
    return temp
```

The above function will automatically extract from the categories list unless told otherwise.

Social media use (frequency)

Then I created the dataframe that I would be putting the data into.

In [69]:
```
social_media_use_f = pd.DataFrame()
```

Then I created the data, and inserted it into the dataframe, using the functions that I created earlier.

In [70]:
```
social = create_data(14, 24, 'c')
insert_columns(0, 'Frequency', social[2:6], social_media_use_f)
```

In [71]:

```
# Let's look at the dataframe so far.
social_media_use_f
```

Out[71]:

	Frequency
0	Max. once a day
1	2 to 10 times a day
2	10 to 20 times a day
3	Over 20 times a day

The first column was now complete. I then transformed the data needed for the second column.

To begin with I extracted the data into a list.

In [72]:

```
social_data = create_data(14, 24, 'f')
social_data
```

Out[72]:

```
[nan, nan, '64', '578', '439', '250', nan, '232', '1048', '51']
```

I only used a part of the list, so I could ignore the NaNs, as they were not part of the data that I required.
I then collected the part of the list that I wanted. Converted them into integers, and then calculated the percentage. I then calculated the percentage by suming the number of responses, and calculated the fraction for each response by dividing them by the sum.

In [73]:

```
# splicing the list
social_data_splice = social_data[2:6]
social_data_splice
```

Out[73]:

```
['64', '578', '439', '250']
```

In [74]:

```
# converting into integers
social_data_int = [int(social_data_splice[x]) for x in range(0, len(social_data_spli
social_data_int
```

Out[74]:

```
[64, 578, 439, 250]
```

In [75]:

```
# calculating total
social_data_total = sum(social_data_int)
social_data_total
```

Out[75]:

```
1331
```

In [76]:

```
# calculating the float
social_data_f = [float(format(social_data_int[x]/social_data_total, '.3f')) for x in
social_data_f
```

Out[76]:

```
[0.048, 0.434, 0.33, 0.188]
```

In [77]:

```
# calculating the percentage
social_data_p = [float(format(social_data_f[x]*100, '.3f')) for x in range(0, len(so
social_data_p
```

Out[77]:

```
[4.8, 43.4, 33.0, 18.8]
```

I then added the two new columns to the dataframe using the insert_columns() function I created before.

In [78]:

```
insert_columns(1, 'Percentage (f)', social_data_f, social_media_use_f)
insert_columns(1, 'Percentage (%)', social_data_p, social_media_use_f)
```

In [79]:

```
# Let's take a look at the dataframe
social_media_use_f
```

Out[79]:

	Frequency	Percentage (%)	Percentage (f)
0	Max. once a day	4.8	0.048
1	2 to 10 times a day	43.4	0.434
2	10 to 20 times a day	33.0	0.330
3	Over 20 times a day	18.8	0.188

The first dataframe was then completed, it represented the frequency of social media use.
I then wrote the data to a csv file.

In [80]:

```
# Setting to false stops the index from being written out
social_media_use_f.to_csv('./data/social_media_use_f.csv', index=False)
```

Creating some functions to simplify the task.

I used a similar method for creating data and adding data to the dataframe as I did above. Therefore I decided to create a set of functions that I could use, to collapse the process into 2 lines.

Below are functions that covert the data into integers, sum the data, and calculate the percentages/floats.

In [87]:

```
def convert_int(data):
    temp = [int(data[x]) for x in range(0, len(data))]
    print("Ints:       ", temp)
    return temp

def calc_total(data):
    temp = sum(data)
    print("Total:      ", temp)
    return temp

def calc_float(data, total):
    temp = [float(format(data[x]/total, '.3f')) for x in range(0, len(data))]
    print("Floats:     ", temp)
    return temp

def calc_per(data):
    temp = [float(format(data[x]*100, '.3f')) for x in range(0, len(data))]
    print("Percentages:", temp)
    return temp
```

Social media use frequency (time spent)

I used the same process that I did for the frequency, with time spent.
I began by creating the dataframe and inserting the first column.

In [88]:

```
social_media_use_t = pd.DataFrame()
insert_columns(0, 'Time Spent', social[7:10], social_media_use_t)
social_media_use_t
```

Out[88]:

	Time Spent
0	Less than 1 h
1	Between 1 and 5 h
2	Over 5 h

I then created the data for the other columns, and inserted them into the dataframe.
I used the functions I created earlier.

In [89]:

```
# splicing the list
social_data_splice_t = social_data[7:10]
print("Spliced list:", social_data_splice_t)
```

```
Spliced list: ['232', '1048', '51']
```

I used all four functions to create the data for the percentages/floats.

In [90]:

```
social_data_f_t = calc_float(convert_int(social_data_splice_t), calc_total(convert_i
social_data_p_t = calc_per(calc_float(convert_int(social_data_splice_t), calc_total(
```

```
Ints:        [232, 1048, 51]
Ints:        [232, 1048, 51]
Total:       1331
Floats:      [0.174, 0.787, 0.038]
Ints:        [232, 1048, 51]
Ints:        [232, 1048, 51]
Total:       1331
Floats:      [0.174, 0.787, 0.038]
Percentages: [17.4, 78.7, 3.8]
```

I inserted the columns:

In [91]:

```
insert_columns(1, 'Percentage (f)', social_data_f_t, social_media_use_t)
insert_columns(1, 'Percentage (%)', social_data_p_t, social_media_use_t)
social_media_use_t
```

Out[91]:

	Time Spent	Percentage (%)	Percentage (f)
0	Less than 1 h	17.4	0.174
1	Between 1 and 5 h	78.7	0.787
2	Over 5 h	3.8	0.038

The second dataframe was now completed, it represented the time spent on social media.

I then wrote the data to a csv file.

In [92]:

```
# Setting to false stops the index from being written out
social_media_use_t.to_csv('./data/social_media_use_t.csv', index=False)
```

Body Comparison

I then extracted the data I needed to create a body comparison dataframe. The process of extracting and cleaning the data was the same as before.

First I created the dataframe.

In [96]:

```
body_comparison = pd.DataFrame()
```

I then created the first column.
I utilised the same **create_data()** function I created for the purpose of extracting data from this source. Just like in the previous two dataframes. I am used the **insert_columns()** function I created earlier. The functions made the process a lot faster and easier.

In [97]:

```
bc = create_data(24, 30, 'c')
insert_columns(0, 'Body Comparison', bc[1:], body_comparison)
```

In [98]:

```
# Let's see the dataframe so far
body_comparison
```

Out[98]:

	Body Comparison
0	Never
1	Seldom
2	Sometimes
3	Often
4	Always

I then moved on to the next two columns by creating the data using the functions created above.

In [99]:

```
bc_data = create_data(25, 30, 'f')
bc_data_f = calc_float(convert_int(bc_data), calc_total(convert_int(bc_data)))
bc_data_p = calc_per(calc_float(convert_int(bc_data), calc_total(convert_int(bc_data
```

```
Ints:        [33, 114, 317, 523, 344]
Ints:        [33, 114, 317, 523, 344]
Total:       1331
Floats:      [0.025, 0.086, 0.238, 0.393, 0.258]
Ints:        [33, 114, 317, 523, 344]
Ints:        [33, 114, 317, 523, 344]
Total:       1331
Floats:      [0.025, 0.086, 0.238, 0.393, 0.258]
Percentages: [2.5, 8.6, 23.8, 39.3, 25.8]
```

I then added the data to the dataframe.

In [100]:

```
insert_columns(1, 'Percentage (f)', bc_data_f, body_comparison)
insert_columns(1, 'Percentage (%)', bc_data_p, body_comparison)
```

In [101]:

```
body_comparison
```

Out[101]:

	Body Comparison	Percentage (%)	Percentage (f)
0	Never	2.5	0.025
1	Seldom	8.6	0.086
2	Sometimes	23.8	0.238
3	Often	39.3	0.393
4	Always	25.8	0.258

The third dataframe from this set of data was then completed, it represented how often people compare their bodies to others.

I then wrote the data to a csv file.

In [102]:

```
# Setting to false stops the index from being written out
body_comparison.to_csv('./data/body_comparison.csv', index=False)
```

Posting Selfies

This was the last group of data that I created from the records dataframe.
I extracted the data I needed to create a posting selfies dataframe. The process of extracting and cleaning the data was identical to before.

First I created the dataframe.

In [105]:

```
posting_selfies = pd.DataFrame()
```

I then created the first column.
I used the **create_data()** and **insert_columns()** functions I created before.

In [106]:

```
ps = create_data(31, 35)
insert_columns(0, 'Posting Selfies', ps, posting_selfies)
# Let's see the dataframe so far
posting_selfies
```

Out[106]:

	Posting Selfies
0	Never
1	1 or 2 times a month
2	Once a week
3	3 to 4 times a week

I then created the next two columns, used the five functions, and then I added them to the dataframe.

In [107]:

```
# creating the data
ps_data = create_data(31, 35, 'f')

# calculating the float and percentages
ps_data_f = calc_float(convert_int(ps_data), calc_total(convert_int(ps_data)))
ps_data_p = calc_per(calc_float(convert_int(ps_data), calc_total(convert_int(ps_data

# inserting them into the dataframe
insert_columns(1, 'Percentage (f)', ps_data_f, posting_selfies)
insert_columns(1, 'Percentage (%)', ps_data_p, posting_selfies)
```

```
Ints:        [457, 756, 93, 18]
Ints:        [457, 756, 93, 18]
Total:       1324
Floats:      [0.345, 0.571, 0.07, 0.014]
Ints:        [457, 756, 93, 18]
Ints:        [457, 756, 93, 18]
Total:       1324
Floats:      [0.345, 0.571, 0.07, 0.014]
Percentages: [34.5, 57.1, 7.0, 1.4]
```

In [108]:

```
posting_selfies
```

Out[108]:

	Posting Selfies	Percentage (%)	Percentage (f)
0	Never	34.5	0.345
1	1 or 2 times a month	57.1	0.571
2	Once a week	7.0	0.070
3	3 to 4 times a week	1.4	0.014

The last dataframe from this set of data was now completed, it represented how often people post selfies

online.

I then wrote the data to a csv file.

In [109]:

```
# Setting to false stops the index from being written out
posting_selfies.to_csv('./data/posting_selfies.csv', index=False)
```

2.2 Transforming excel files

URL[3] | Suffering from anxiety

Number of Gen Z that suffer from nerves, anxiety, and more in the US in 2020

This source contains data on the percentage of people that feel nervous, anxious, or on edge more than half of the days or daily.
The data is in the form of an excel spreadsheet. I started by creating a function, so that I could repeat this process with ease.

Importing excel files

In [127]:

```
def import_excel(fn):
    temp = pd.read_excel(f'./data/{fn}.xlsx')
    return temp
```

I first imported the file.

In [128]:

```
df_3 = import_excel('suffering_from_anxiety')
```

I then displayed the table.

In [129]:

```
df_3
```

Out[129]:

	LocationType	Location	TimeFrame	DataFormat	Data
0	Nation	United States	Apr 23-May 12, 2020	Percent	0.39
1	Nation	United States	May 7-May 19, 2020	Percent	0.38
2	Nation	United States	May 14-May 26, 2020	Percent	0.37
3	Nation	United States	May 21-Jun 2, 2020	Percent	0.39
4	Nation	United States	May 28-Jun 9, 2020	Percent	0.39
...	...	...	...	...	...
1503	State	Wyoming	Jun 4-Jun 16, 2020	Percent	0.52
1504	State	Wyoming	Apr 23-May 12, 2020	Percent	S
1505	State	Wyoming	May 7-May 19, 2020	Percent	S
1506	State	Wyoming	May 14-May 26, 2020	Percent	S
1507	State	Wyoming	May 21-Jun 2, 2020	Percent	S

1508 rows × 5 columns

After looking through the data, I saw that there was a lot of NaN data, so I got rid of them.

In [130]:

```
df_3[df_3['Data'] == 'S']
```

Out[130]:

	LocationType	Location	TimeFrame	DataFormat	Data
29	State	Alabama	Jun 09-Jul 05, 2021	Percent	S
30	State	Alabama	May 26-Jun 21, 2021	Percent	S
31	State	Alabama	May 12-Jun 07, 2021	Percent	S
32	State	Alabama	Apr 28-May 24, 2021	Percent	S
33	State	Alabama	Apr 14-May 10, 2021	Percent	S
...	...	...	...	...	...
1502	State	Wyoming	May 28-Jun 9, 2020	Percent	S
1504	State	Wyoming	Apr 23-May 12, 2020	Percent	S
1505	State	Wyoming	May 7-May 19, 2020	Percent	S
1506	State	Wyoming	May 14-May 26, 2020	Percent	S
1507	State	Wyoming	May 21-Jun 2, 2020	Percent	S

567 rows × 5 columns

In [131]:

```
df_3 = df_3.drop(df_3[df_3['Data'] == 'S'].index)
```

In [132]:

```
df_3.shape
```

Out[132]:

```
(941, 5)
```

In [133]:

```
df_3[df_3['Location'] != 'United States']
```

Out[133]:

	LocationType	Location	TimeFrame	DataFormat	Data
34	State	Alabama	Mar 3-Mar 29, 2021	Percent	0.63
39	State	Alabama	Nov 25-Dec 21, 2020	Percent	0.31
40	State	Alabama	Oct 14-Nov 9, 2020	Percent	0.51
43	State	Alabama	Sep 16-Oct 12, 2020	Percent	0.4
44	State	Alabama	Sep 30-Oct 26, 2020	Percent	0.5
...	...	...	...	...	...
1493	State	Wyoming	Oct 14-Nov 9, 2020	Percent	0.13
1494	State	Wyoming	Sep 2-Sep 28, 2020	Percent	0.46
1496	State	Wyoming	Jul 9-Jul 21, 2020	Percent	0.42
1497	State	Wyoming	Aug 19-Sep 14, 2020	Percent	0.4
1503	State	Wyoming	Jun 4-Jun 16, 2020	Percent	0.52

912 rows × 5 columns

Once I got rid of all the empty data, I then proceeded to remove the data pertaining to individual states, as I felt that it wasn't necessary to analyse individual states for this particular proposal.

In [134]:

```
df_3 = df_3.drop(df_3[df_3['Location'] != 'United States'].index)
```

In [135]:

```
df_3
```

Out[135]:

	LocationType	Location	TimeFrame	DataFormat	Data
0	Nation	United States	Apr 23-May 12, 2020	Percent	0.39
1	Nation	United States	May 7-May 19, 2020	Percent	0.38
2	Nation	United States	May 14-May 26, 2020	Percent	0.37
3	Nation	United States	May 21-Jun 2, 2020	Percent	0.39
4	Nation	United States	May 28-Jun 9, 2020	Percent	0.39
5	Nation	United States	Jun 4-Jun 16, 2020	Percent	0.38
6	Nation	United States	Jun 11-Jun 23, 2020	Percent	0.4
7	Nation	United States	Jun 18-Jun 30, 2020	Percent	0.4
8	Nation	United States	Jun 25-Jul 7, 2020	Percent	0.41
9	Nation	United States	Jul 2-Jul 14, 2020	Percent	0.42
10	Nation	United States	Jul 9-Jul 21, 2020	Percent	0.41
11	Nation	United States	Sep 2-Sep 28, 2020	Percent	0.39
12	Nation	United States	Aug 19-Sep 14, 2020	Percent	0.39
13	Nation	United States	Sep 16-Oct 12, 2020	Percent	0.4
14	Nation	United States	Sep 30-Oct 26, 2020	Percent	0.43
15	Nation	United States	Oct 14-Nov 9, 2020	Percent	0.47
16	Nation	United States	Oct 28-Nov 23, 2020	Percent	0.5
17	Nation	United States	Nov 11-Dec 7, 2020	Percent	0.5
18	Nation	United States	Nov 25-Dec 21, 2020	Percent	0.48
19	Nation	United States	Jan 6-Feb 1, 2021	Percent	0.45
20	Nation	United States	Jan 20-Feb 15, 2021	Percent	0.47
21	Nation	United States	Feb 3-Mar 1, 2021	Percent	0.48
22	Nation	United States	Feb 17-Mar 15, 2021	Percent	0.46
23	Nation	United States	Mar 3-Mar 29, 2021	Percent	0.44
24	Nation	United States	Apr 14-May 10, 2021	Percent	0.44
25	Nation	United States	Apr 28-May 24, 2021	Percent	0.42
26	Nation	United States	May 12-Jun 07, 2021	Percent	0.39
27	Nation	United States	May 26-Jun 21, 2021	Percent	0.38
28	Nation	United States	Jun 09-Jul 05, 2021	Percent	0.39

In [136]:

```
delete_columns('LocationType', df_3)
delete_columns('Location', df_3)
delete_columns('DataFormat', df_3)
df_3
```

Out[136]:

	TimeFrame	Data
0	Apr 23-May 12, 2020	0.39
1	May 7-May 19, 2020	0.38
2	May 14-May 26, 2020	0.37
3	May 21-Jun 2, 2020	0.39
4	May 28-Jun 9, 2020	0.39
5	Jun 4-Jun 16, 2020	0.38
6	Jun 11-Jun 23, 2020	0.4
7	Jun 18-Jun 30, 2020	0.4
8	Jun 25-Jul 7, 2020	0.41
9	Jul 2-Jul 14, 2020	0.42
10	Jul 9-Jul 21, 2020	0.41
11	Sep 2-Sep 28, 2020	0.39
12	Aug 19-Sep 14, 2020	0.39
13	Sep 16-Oct 12, 2020	0.4
14	Sep 30-Oct 26, 2020	0.43
15	Oct 14-Nov 9, 2020	0.47
16	Oct 28-Nov 23, 2020	0.5
17	Nov 11-Dec 7, 2020	0.5
18	Nov 25-Dec 21, 2020	0.48
19	Jan 6-Feb 1, 2021	0.45
20	Jan 20-Feb 15, 2021	0.47
21	Feb 3-Mar 1, 2021	0.48
22	Feb 17-Mar 15, 2021	0.46
23	Mar 3-Mar 29, 2021	0.44
24	Apr 14-May 10, 2021	0.44
25	Apr 28-May 24, 2021	0.42
26	May 12-Jun 07, 2021	0.39
27	May 26-Jun 21, 2021	0.38
28	Jun 09-Jul 05, 2021	0.39

I then decided to convert the strings into numerical data.

In [137]:

```
df_3['Data']
```

Out[137]:

```
0     0.39
1     0.38
2     0.37
3     0.39
4     0.39
5     0.38
6      0.4
7      0.4
8     0.41
9     0.42
10    0.41
11    0.39
12    0.39
13     0.4
14    0.43
15    0.47
16     0.5
17     0.5
18    0.48
19    0.45
20    0.47
21    0.48
22    0.46
23    0.44
24    0.44
25    0.42
26    0.39
27    0.38
28    0.39
Name: Data, dtype: object
```

I used my calc_per() function to turn the floats into percentages, once I no longer had strings.

In [142]:

```
# convert to float
temp_3 = [float(df_3['Data'][x]) for x in range(0, len(df_3['Data']))]
data_p_3 = calc_per(temp_3)
```

```
Percentages: [39.0, 38.0, 37.0, 39.0, 39.0, 38.0, 40.0, 40.0, 41.0, 4
2.0, 41.0, 39.0, 39.0, 40.0, 43.0, 47.0, 50.0, 50.0, 48.0, 45.0, 47.0,
48.0, 46.0, 44.0, 44.0, 42.0, 39.0, 38.0, 39.0]
```

I then inserted the newly created data into the table.

In [144]:

```
insert_columns(2, 'Data (%)', data_p_3, df_3)
```

In [145]:

```
df_3
```

Out[145]:

	TimeFrame	Data	Data (%)
0	Apr 23-May 12, 2020	0.39	39.0
1	May 7-May 19, 2020	0.38	38.0
2	May 14-May 26, 2020	0.37	37.0
3	May 21-Jun 2, 2020	0.39	39.0
4	May 28-Jun 9, 2020	0.39	39.0
5	Jun 4-Jun 16, 2020	0.38	38.0
6	Jun 11-Jun 23, 2020	0.4	40.0
7	Jun 18-Jun 30, 2020	0.4	40.0
8	Jun 25-Jul 7, 2020	0.41	41.0
9	Jul 2-Jul 14, 2020	0.42	42.0
10	Jul 9-Jul 21, 2020	0.41	41.0
11	Sep 2-Sep 28, 2020	0.39	39.0
12	Aug 19-Sep 14, 2020	0.39	39.0
13	Sep 16-Oct 12, 2020	0.4	40.0
14	Sep 30-Oct 26, 2020	0.43	43.0
15	Oct 14-Nov 9, 2020	0.47	47.0
16	Oct 28-Nov 23, 2020	0.5	50.0
17	Nov 11-Dec 7, 2020	0.5	50.0
18	Nov 25-Dec 21, 2020	0.48	48.0
19	Jan 6-Feb 1, 2021	0.45	45.0
20	Jan 20-Feb 15, 2021	0.47	47.0
21	Feb 3-Mar 1, 2021	0.48	48.0
22	Feb 17-Mar 15, 2021	0.46	46.0
23	Mar 3-Mar 29, 2021	0.44	44.0
24	Apr 14-May 10, 2021	0.44	44.0
25	Apr 28-May 24, 2021	0.42	42.0
26	May 12-Jun 07, 2021	0.39	39.0
27	May 26-Jun 21, 2021	0.38	38.0
28	Jun 09-Jul 05, 2021	0.39	39.0

I deleted the other column of the same name as I knew that I wouldn't need it.

In [146]:

```
delete_columns(['Data'], df_3)
df_3
```

Out[146]:

	TimeFrame	Data (%)
0	Apr 23-May 12, 2020	39.0
1	May 7-May 19, 2020	38.0
2	May 14-May 26, 2020	37.0
3	May 21-Jun 2, 2020	39.0
4	May 28-Jun 9, 2020	39.0
5	Jun 4-Jun 16, 2020	38.0
6	Jun 11-Jun 23, 2020	40.0
7	Jun 18-Jun 30, 2020	40.0
8	Jun 25-Jul 7, 2020	41.0
9	Jul 2-Jul 14, 2020	42.0
10	Jul 9-Jul 21, 2020	41.0
11	Sep 2-Sep 28, 2020	39.0
12	Aug 19-Sep 14, 2020	39.0
13	Sep 16-Oct 12, 2020	40.0
14	Sep 30-Oct 26, 2020	43.0
15	Oct 14-Nov 9, 2020	47.0
16	Oct 28-Nov 23, 2020	50.0
17	Nov 11-Dec 7, 2020	50.0
18	Nov 25-Dec 21, 2020	48.0
19	Jan 6-Feb 1, 2021	45.0
20	Jan 20-Feb 15, 2021	47.0
21	Feb 3-Mar 1, 2021	48.0
22	Feb 17-Mar 15, 2021	46.0
23	Mar 3-Mar 29, 2021	44.0
24	Apr 14-May 10, 2021	44.0
25	Apr 28-May 24, 2021	42.0
26	May 12-Jun 07, 2021	39.0
27	May 26-Jun 21, 2021	38.0
28	Jun 09-Jul 05, 2021	39.0

I then wrote the data to a csv file.

In [148]:

```
# Setting to false stops the index from being written out
df_3.to_csv('./data/df_3.csv', index=False)
```

URL[4] | Suffering from depression

Number of Gen Z that are depressed and hopeless in the US in 2021

This source contains data on the percentage of people that feel depressed and hopeless, more than half of the days or daily.

Importing the data

In [149]:

```
df_4 = import_excel('suffering_from_depression')
df_4
```

Out[149]:

	LocationType	Location	TimeFrame	DataFormat	Data
0	Nation	United States	Apr 23-May 12, 2020	Percent	0.31
1	Nation	United States	May 7-May 19, 2020	Percent	0.3
2	Nation	United States	May 14-May 26, 2020	Percent	0.31
3	Nation	United States	May 21-Jun 2, 2020	Percent	0.32
4	Nation	United States	May 28-Jun 9, 2020	Percent	0.32
...	...	...	...	...	...
1503	State	Wyoming	Jun 4-Jun 16, 2020	Percent	S
1504	State	Wyoming	Apr 23-May 12, 2020	Percent	S
1505	State	Wyoming	May 7-May 19, 2020	Percent	S
1506	State	Wyoming	May 14-May 26, 2020	Percent	S
1507	State	Wyoming	May 21-Jun 2, 2020	Percent	S

1508 rows × 5 columns

I then repeated the same method that I did for the previous data to get rid of unwanted data.

In [150]:

```
df_4 = df_4.drop(df_4[df_4['Data'] == 'S'].index)
df_4 = df_4.drop(df_4[df_4['Location'] != 'United States'].index)

delete_columns('LocationType', df_4)
delete_columns('Location', df_4)
delete_columns('DataFormat', df_4)
df_4
```

Out[150]:

	TimeFrame	Data
0	Apr 23-May 12, 2020	0.31
1	May 7-May 19, 2020	0.3
2	May 14-May 26, 2020	0.31
3	May 21-Jun 2, 2020	0.32
4	May 28-Jun 9, 2020	0.32
5	Jun 4-Jun 16, 2020	0.32
6	Jun 11-Jun 23, 2020	0.33
7	Jun 18-Jun 30, 2020	0.32
8	Jun 25-Jul 7, 2020	0.31
9	Jul 2-Jul 14, 2020	0.3
10	Jul 9-Jul 21, 2020	0.32
11	Sep 2-Sep 28, 2020	0.31
12	Aug 19-Sep 14, 2020	0.32
13	Sep 16-Oct 12, 2020	0.32
14	Sep 30-Oct 26, 2020	0.36
15	Oct 14-Nov 9, 2020	0.37
16	Oct 28-Nov 23, 2020	0.39
17	Nov 11-Dec 7, 2020	0.39
18	Nov 25-Dec 21, 2020	0.38
19	Jan 6-Feb 1, 2021	0.36
20	Jan 20-Feb 15, 2021	0.4
21	Feb 3-Mar 1, 2021	0.42
22	Feb 17-Mar 15, 2021	0.4
23	Mar 3-Mar 29, 2021	0.36
24	Apr 14-May 10, 2021	0.35
25	Apr 28-May 24, 2021	0.32
26	May 12-Jun 07, 2021	0.31
27	May 26-Jun 21, 2021	0.32
28	Jun 09-Jul 05, 2021	0.35

I again repeated the same method so that I could create percentages and clean the data.

In [151]:

```
# convert to float
temp_4 = [float(df_4['Data'][x]) for x in range(0, len(df_4['Data']))]
data_p_4 = calc_per(temp_4)
```

```
Percentages: [31.0, 30.0, 31.0, 32.0, 32.0, 32.0, 33.0, 32.0, 31.0, 3
0.0, 32.0, 31.0, 32.0, 32.0, 36.0, 37.0, 39.0, 39.0, 38.0, 36.0, 40.0,
42.0, 40.0, 36.0, 35.0, 32.0, 31.0, 32.0, 35.0]
```

I then inserted the new column and deleted the old.

In [152]:

```
insert_columns(2, 'Data (%)', data_p_4, df_4)
delete_columns(['Data'], df_4)
df_4
```

Out[152]:

	TimeFrame	Data (%)
0	Apr 23-May 12, 2020	31.0
1	May 7-May 19, 2020	30.0
2	May 14-May 26, 2020	31.0
3	May 21-Jun 2, 2020	32.0
4	May 28-Jun 9, 2020	32.0
5	Jun 4-Jun 16, 2020	32.0
6	Jun 11-Jun 23, 2020	33.0
7	Jun 18-Jun 30, 2020	32.0
8	Jun 25-Jul 7, 2020	31.0
9	Jul 2-Jul 14, 2020	30.0
10	Jul 9-Jul 21, 2020	32.0
11	Sep 2-Sep 28, 2020	31.0
12	Aug 19-Sep 14, 2020	32.0
13	Sep 16-Oct 12, 2020	32.0
14	Sep 30-Oct 26, 2020	36.0
15	Oct 14-Nov 9, 2020	37.0
16	Oct 28-Nov 23, 2020	39.0
17	Nov 11-Dec 7, 2020	39.0
18	Nov 25-Dec 21, 2020	38.0
19	Jan 6-Feb 1, 2021	36.0
20	Jan 20-Feb 15, 2021	40.0
21	Feb 3-Mar 1, 2021	42.0
22	Feb 17-Mar 15, 2021	40.0
23	Mar 3-Mar 29, 2021	36.0
24	Apr 14-May 10, 2021	35.0
25	Apr 28-May 24, 2021	32.0
26	May 12-Jun 07, 2021	31.0
27	May 26-Jun 21, 2021	32.0
28	Jun 09-Jul 05, 2021	35.0

I then wrote the data to a csv file.

In [153]:

```
# Setting to false stops the index from being written out
df_4.to_csv('./data/df_4.csv', index=False)
```

URL[5] | Poor mental health

Number of Gen Z reporting poor mental health

This source contains data on the percentage of people who reported having zero days in the past 30 days, where there mental health was poor.

Importing the data

In [156]:

```
df_5 = import_excel('poor_mental_health')
df_5
```

Out[156]:

	LocationType	Location	TimeFrame	DataFormat	Data
0	Nation	United States	2011-2013	Percent	0.54
1	Nation	United States	2012-2014	Percent	0.54
2	Nation	United States	2013-2015	Percent	0.54
3	Nation	United States	2014-2016	Percent	0.53
4	Nation	United States	2015-2017	Percent	0.51
...	...	...	...	...	...
366	Territory	Puerto Rico	2013-2015	Percent	0.80
367	Territory	Puerto Rico	2014-2016	Percent	0.81
368	Territory	Puerto Rico	2015-2017	Percent	0.81
369	Territory	Puerto Rico	2016-2018	Percent	0.79
370	Territory	Puerto Rico	2017-2019	Percent	0.78

371 rows × 5 columns

This time there was no missing or erroneous data.

In [162]:

```
df_5[df_5['Data'] == 'S']
```

Out[162]:

	LocationType	Location	TimeFrame	DataFormat	Data

Though I still had to get rid of the data pertaining to individual states, and the unneeded columns.

In [163]:

```
df_5 = df_5.drop(df_5[df_5['Location'] != 'United States'].index)

delete_columns('LocationType', df_5)
delete_columns('Location', df_5)
delete_columns('DataFormat', df_5)
df_5
```

Out[163]:

	TimeFrame	Data
0	2011-2013	0.54
1	2012-2014	0.54
2	2013-2015	0.54
3	2014-2016	0.53
4	2015-2017	0.51
5	2016-2018	0.49
6	2017-2019	0.46

I repeated the exact same method as before to turn the data into percentages.

In [164]:

```
# convert to float
temp_5 = [float(df_5['Data'][x]) for x in range(0, len(df_5['Data']))]
data_p_5 = calc_per(temp_5)
```

```
Percentages: [54.0, 54.0, 54.0, 53.0, 51.0, 49.0, 46.0]
```

I then added the new column for data and removed the old one.

In [165]:

```
insert_columns(2, 'Data (%)', data_p_5, df_5)
delete_columns(['Data'], df_5)
df_5
```

Out[165]:

	TimeFrame	Data (%)
0	2011-2013	54.0
1	2012-2014	54.0
2	2013-2015	54.0
3	2014-2016	53.0
4	2015-2017	51.0
5	2016-2018	49.0
6	2017-2019	46.0

I then wrote the data to a csv file.

In [166]:

```
# Setting to false stops the index from being written out
df_5.to_csv('./data/df_5.csv', index=False)
```

---

3 Exploratory data analysis

3.1 The impact of a media influenced and technology driven world on Gen Z

3.1.1 Measuring the forms of media that Gen Z take in, is it considered traditional?

It is no secret the Generation Z are said to be glued to their phones, but to what extent does this relationship go? Just how dependent are Generation Zers to technology?

To go about answering this question, I decided that I would look into the news. This sounds terribly random, but the connection is uncanny. We live in a world that is constantly changing, with new headlines released each second, and news spreading like wildfire, people want to stay in the loop. There is a major FOMO (fear of missing out) amongst this young generation, and the easiest way to stay connected is to intake the news in some way or form.

In [22]:

```
df_0
```

Out[22]:

	Source	Daily (%)	A few times per week (%)	Once per week (%)	A few times per month (%)	Once per month (%)	Less than once per month (%)	Never (%)
0	Social media	50	18	9	10	3	4	6
1	Radio	17	17	8	8	6	13	31
2	Online-only news sites	13	18	14	11	6	9	29
3	Podcasts	13	14	7	11	5	13	38
4	Network news	9	13	12	9	6	12	39
5	Cable news networks	8	14	9	7	6	12	44
6	Newspapers	5	4	9	6	6	17	53

The table above displays the information that I found.
I created a bar chart below to better display the data, and make it more digestible.

In [ ]:

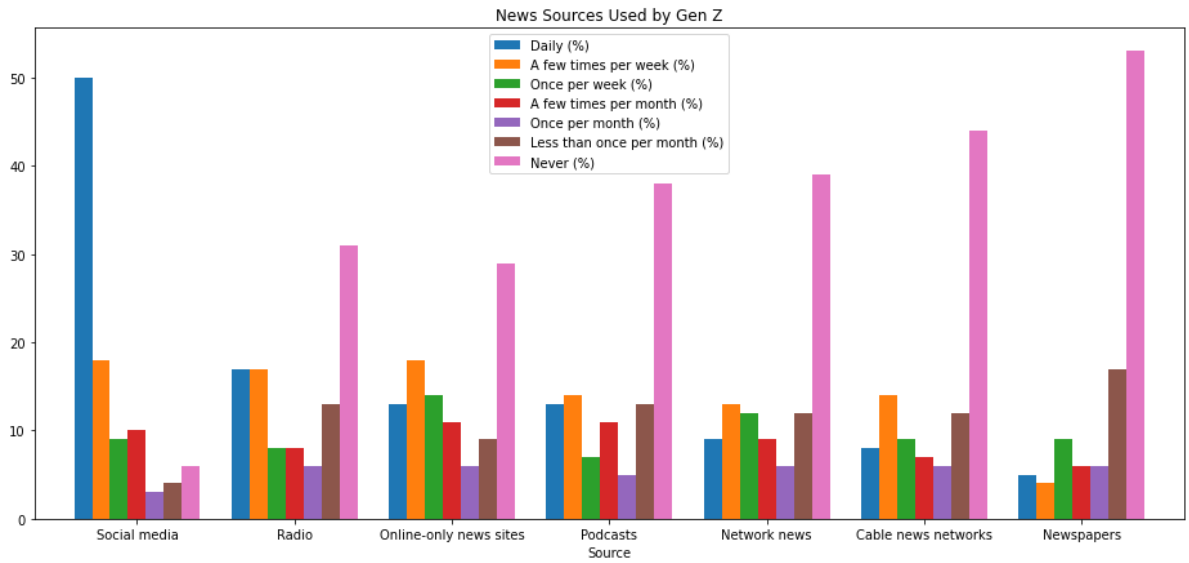
```
df_0_copy = df_0
df_0_copy.set_index('Source', inplace=True)
```

In [185]:

```
df_0_bar_chart = df_0_copy.plot.bar(rot=0, figsize = (16, 7), width = 0.8, title="Ne
df_0_bar_chart
```

Out[185]:

```
<AxesSubplot:title={'center':'News Sources Used by Gen Z'}, xlabel='So
urce'>
```



In the bar chart above, I saw that the never bar was the highest for all other sources of news but social media. This meant that Generation Z's primary source of news was from social media. Taking a further look at the social media section, I came to realise that 50% of those who use social media as a news source, actually check the news daily. As opposed to the 5% of Newspapers users who check daily.

The lowest never bar aside from social media (6%) was online new sources which was at 29%. Based of the results from this study, I saw that the most common news sources were internet based, and social media was the most used as a news source.

What did this tell me? This brought about a lot of questions as social media is not very reliable because it is not monitored. At least with some of the other news sources, there has to be some checks that they go through. This is a demonstration of the way in which traditional news forms have adapted. A lot of newpapers have moved online, or have found other ways distributing the news, and so have their clientele.

What does this have to do with the influence of the media? This displayed just how dependent Generation Z is on social media. It is the main source of news for a high portion of them, with the other half mostly being internet based too. Without social media many would lose their connection to the news. It also illustrated the level of trust that they had for social media, as to take in news regularly from a particular source, you have trust that the information is somewhat factual.

3.1.2 What impact has the media had on mental health?

We live in a world that thrives of competititon, people find themselves competiting for things that have no meaning at all. This is a concept that has been around for a while, technology just amplified it. People spend their time browsing through social media comparing what they see, to who they are. It is a vicious cycle,

everyone posts theirselves at there best, no one wants to share their downfalls, unless they can make a joke out of it. Having grown up in ever tech growing world Generation Z have faced the brunt of it. The generation grew up sharing information online. What impact has that left on their mental health?

I transformed the data that I got from the body comparison source into a pie chart.

In [103]:

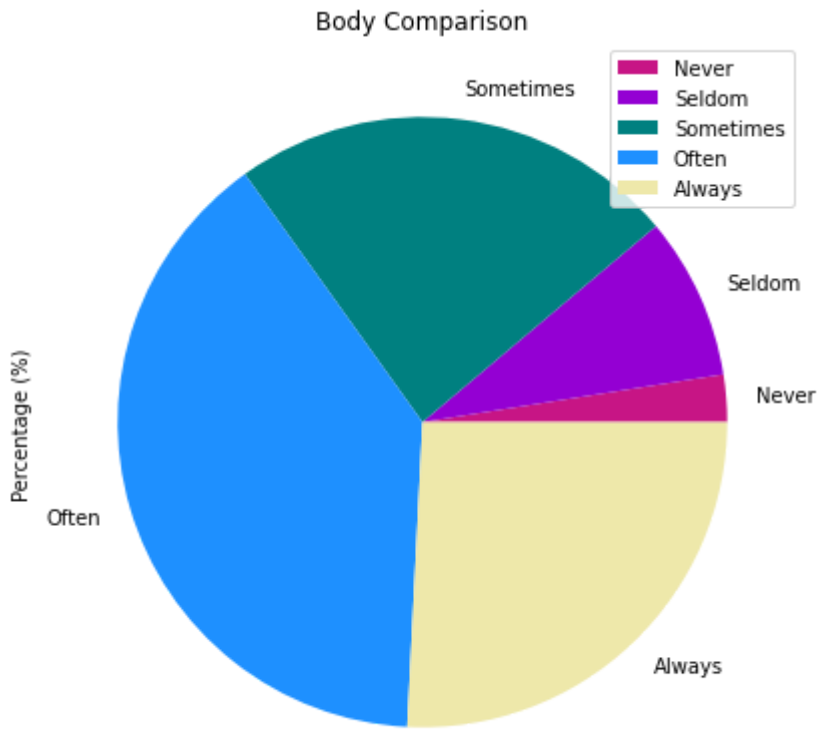
```
body_comparison_copy = body_comparison
body_comparison_copy.set_index('Body Comparison', inplace=True)
delete_columns('Percentage (f)', body_comparison_copy)

body_comparison_copy
```

Out[103]:

Body Comparison	Percentage (%)
Never	2.5
Seldom	8.6
Sometimes	23.8
Often	39.3
Always	25.8

In [180]:

```
body_comparison_pie = body_comparison_copy.plot.pie(y='Percentage (%)', figsize=(7,
                                                    colors=["#C71585", "#9400D3", "#
                                                    , title="Body Comparison")
```



Above you can see that a large proportion of Generation Z compare their bodies to what they see online. When I put the never and seldom category together, and then lumped the rest together- I inferred that 88.9% of Genereration Z sometimes or more than sometimes compare their bodies to others. That puts a lot of stress on a person, but where is this need to better coming from? Why do young people increasingly feel the need to look perfect?

According to the site where I got this data. In order to attain these finding, those taking part had to fill out a questionnaire. The contents of the questionnaire is below.

Drive for Thinness	Always (=3)	Usually (=2)	Often (=1)	Sometimes (=0)	Seldom (=0)	Never (=0)
1—I eat sweets and carbohydrates without feeling nervous	□	□	□	□	□	□
2—I think about dieting	□	□	□	□	□	□
3—I feel extremely guilty after overeating	□	□	□	□	□	□
4—I am terrified of gaining weight	□	□	□	□	□	□
5—I exaggerate or magnify the importance of weight	□	□	□	□	□	□
6—I am preoccupied with the desire to be thinner	□	□	□	□	□	□
7—If I gain a pound, I worry that I will keep gaining	□	□	□	□	□	□
Body Dissatisfaction	**Always**	**Usually**	**Often**	**Sometimes**	**Seldom**	**Never**
1—I think that my stomach is too big (+)	□	□	□	□	□	□
2—I think that my thighs are too large (+)	□	□	□	□	□	□
3—I think that my stomach is just the right size (−)	□	□	□	□	□	□
4—I feel satisfied with the shape of my body (−)	□	□	□	□	□	□
5—I like the shape of my buttocks (−)	□	□	□	□	□	□
6—I think my hips are too big (+)	□	□	□	□	□	□
7—I think that my thighs are just the right size (−)	□	□	□	□	□	□
8—I think my buttocks are too large (+)	□	□	□	□	□	□
9—I think that my hips are just the right size (−)	□	□	□	□	□	□

Questionnaire taken from ncbi (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8001450/)

Looking at the figure above you are able to see just how deep this can be. It is not simply comparing hair colour, people are really critically analysing theirselves. This comes as no surprise. As our world becomes more digital, things appear perfect, and now people also feel the need to be perfect.

3.2 Living with social media and its ramifications

3.2.1 How often are people on social media?

I have written a lot about the impact of social media. That being said, for something to have that big of an impact on you, you have to be influenced by it. In order to be influenced, you need to be around the influencer- therefore you must be using social media frequently. To get closer to the bottom of this, I decided to have a closer look at the social media use frequency data that I attained.

In [95]:

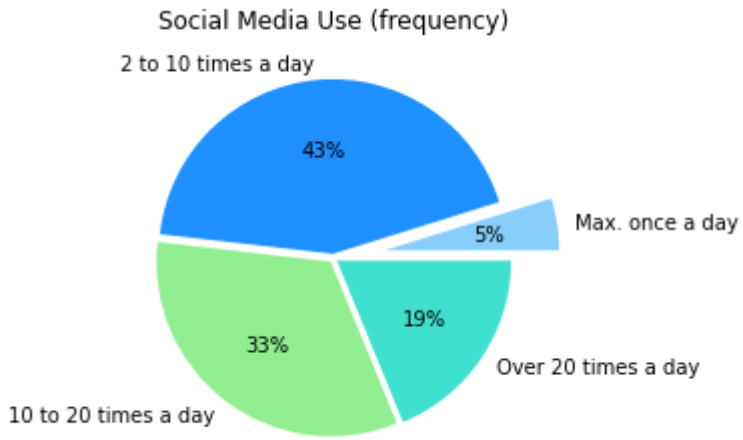
```
social_media_use_f = pd.read_csv('./data/social_media_use_f.csv')
social_media_use_f
```

Out[95]:

	Frequency	Percentage (%)	Percentage (f)
0	Max. once a day	4.8	0.048
1	2 to 10 times a day	43.4	0.434
2	10 to 20 times a day	33.0	0.330
3	Over 20 times a day	18.8	0.188

In [96]:

```
# data
sm_use_f_data = [x for x in social_media_use_f['Percentage (%)']]

# keys
sm_use_f_keys = [x for x in social_media_use_f['Frequency']]

# exploding pie
sm_use_f_explode = [0.3, 0.025, 0.025, 0.025]

# colours
#LightSkyBlue, DodgerBlue, LightGreen, Turquoise
sm_use_f_palette = ['#87CEFA', '#1E90FF', '#90EE90', '#40E0D0']
sm_use_f_colours = sns.color_palette(sm_use_f_palette)

#create pie chart
plt.pie(sm_use_f_data, labels=sm_use_f_keys, colors=sm_use_f_colours, explode=sm_use
plt.title("Social Media Use (frequency)")
plt.show()
```



Above is the pie chart that I created that clearly shows the frequency of social media usage. In the figure above, you can see that only 5% percent of the people that were surveyed used social media a maximum of once a day. According to the data above, a whopping 95% of people used social media twice or more times a

day. That is going onto their chosen platform browsing for a while and getting off, then repeating the same thing all over again.

52% of people used social media over 10 times a day, and 19% used social media over 20 times a day.

From this data, I can deduce that people are more than likely constantly checking their phones, and waiting for a notification. With the amount dedicated to checking social media, this begs the question, just how many hours are people on social media for?

In [81]:

```
social_media_use_f_copy = social_media_use_f
social_media_use_f_copy.set_index('Frequency', inplace=True)
delete_columns('Percentage (f)', social_media_use_f_copy)

social_media_use_f_copy
```

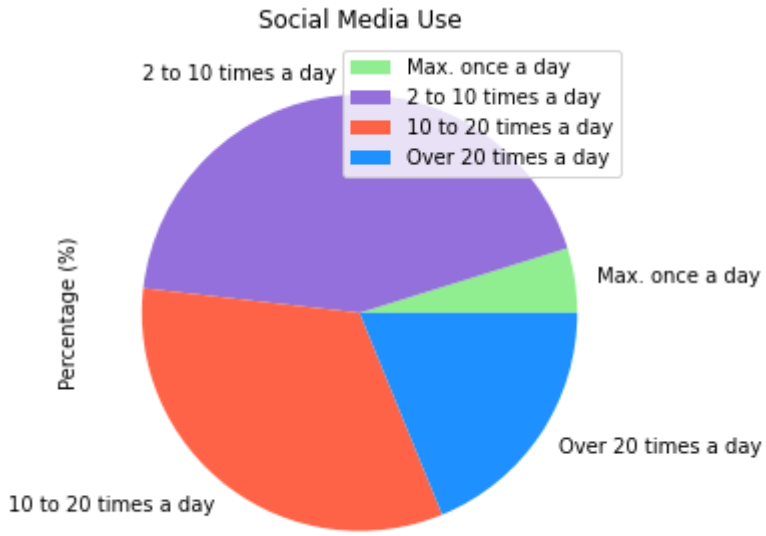
Out[81]:

Frequency	Percentage (%)
Max. once a day	4.8
2 to 10 times a day	43.4
10 to 20 times a day	33.0
Over 20 times a day	18.8

Another version of the pie chart above.

In [186]:

```
social_media_use_f_pie = social_media_use_f_copy.plot.pie(y='Percentage (%)', figsiz
                                                          colors=["#90EE90", "#9370D
                                                          title="Social Media Use")
```



3.2.2 How long are people on social media for?

That being said, the issue with the data above is that, yes people are using social media frequently, but if they are not using it for hours upon hours, those it really matter?

I have explored how often people are on social media, but what about how long they are on it for? If they are just on it for a total of 10 minutes, then them using it 10 times, doesn't really mean much, they only are on it for a minute a piece.

I created another pie chart to display the data that I refined on the time spent on social media.

In [93]:

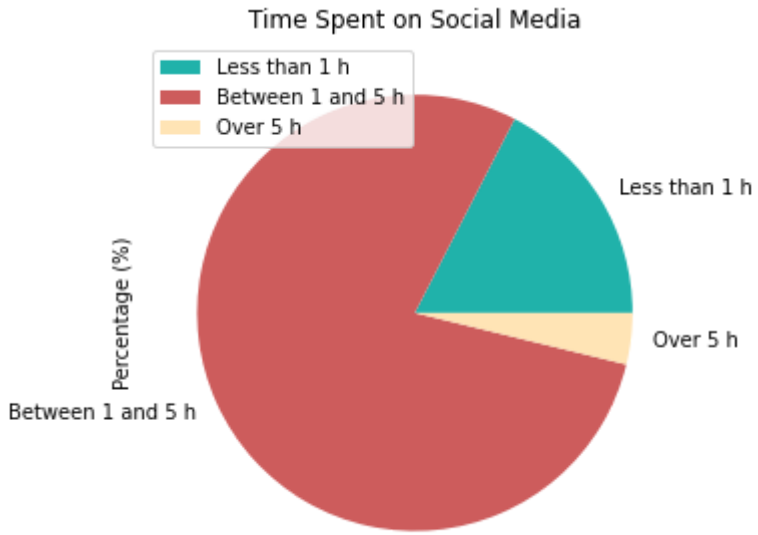
```
social_media_use_t_copy = social_media_use_t
social_media_use_t_copy.set_index('Time Spent', inplace=True)
delete_columns('Percentage (f)', social_media_use_t_copy)

social_media_use_t_copy
```

Out[93]:

Time Spent	Percentage (%)
Less than 1 h	17.4
Between 1 and 5 h	78.7
Over 5 h	3.8

In [188]:

```
social_media_use_t_pie = social_media_use_t_copy.plot.pie(y='Percentage (%)', figsiz
                                                          colors=["#20B2AA", "#CD5C5
                                                          title="Time Spent on Socia
```



The pie chart above is a visual representation of the table just above it. According to the figure about 2/3 of Generation Zers spend between 1 and 5 hours on social media, with 3.8% of them spending over 5 hours, and

17.4% spending less than 1 hour.

From the above data we can see that about 66% of Generation Z spend 1 hours to 5 hours on social. This doesn't seem like such a large number. Though I would like to mention that there is a big difference between 1 hour and 5 hours, 5 hours is exactly 5 times the amount of time- which is a huge difference. To improve on the validity of the data, and to give more detailed data, I believe that the study should have been conducted with smaller categories. I feel as though the jump from 1 to 5 is simply to large.

What I took from this data, was that Generation Z spent a moderate time on social media. I know from reading the source of the data, that those who took part are on the older half of the Generation (18-24). This means that they were possibly working full time, or perhaps studying full time. This suggests that they didn't have as much time to spare on spending hours upon hours on social media. That being said, I saw earlier that they use social media very frequently (majoritity more than 10 times a day). This led me to believe that were constantly checking social media, either during their work breaks, or perhaps during work itself- due to how often the were checking.

Overall, this data shows just how much social media is integrated in their system, due to how frequently they use it, and how long they are on it for.

That being said, I would be interested to see how the younger half of the generation would fare in this study, seeing as they have more free time, and were born with Instagram and YouTube.

3.3 What are Generation Z doing online?

3.3.1 What activities are Generation Z doing online?

I established that Generation Z spent a lot of time on social media, but what other activities are they using the internet for? Does social media encompass their time spent online, or are they equally sharing their time?

In order to find an answer to this question I used the data from df_2 which had information about various activities, and the number of correspondents who take part in them. I further refined the data to create a horizontal bar chart that displayed the data visually.

In [ ]:

```
df_2_copy = df_2
df_2_copy.set_index('Activity', inplace=True)

df_2_copy
```

Out[58]:

Activity	Share of respondents (f)	Share of respondents (%)
Social networking	0.97	97
Using instant messaging services	0.92	92
Listening to/downloading music	0.93	93
Sending/receiving emails	0.94	94
Internet banking	0.90	90
Finding information about goods and services	0.84	84
Reading online news, newspapers or magazines	0.78	78
Making video or voice calls over the internet	0.70	70
Uploading content created by you to a website to be shared	0.56	56
Selling goods or services over the internet	0.22	22
Watching video content from services such as YouTube	0.95	95

In [59]:

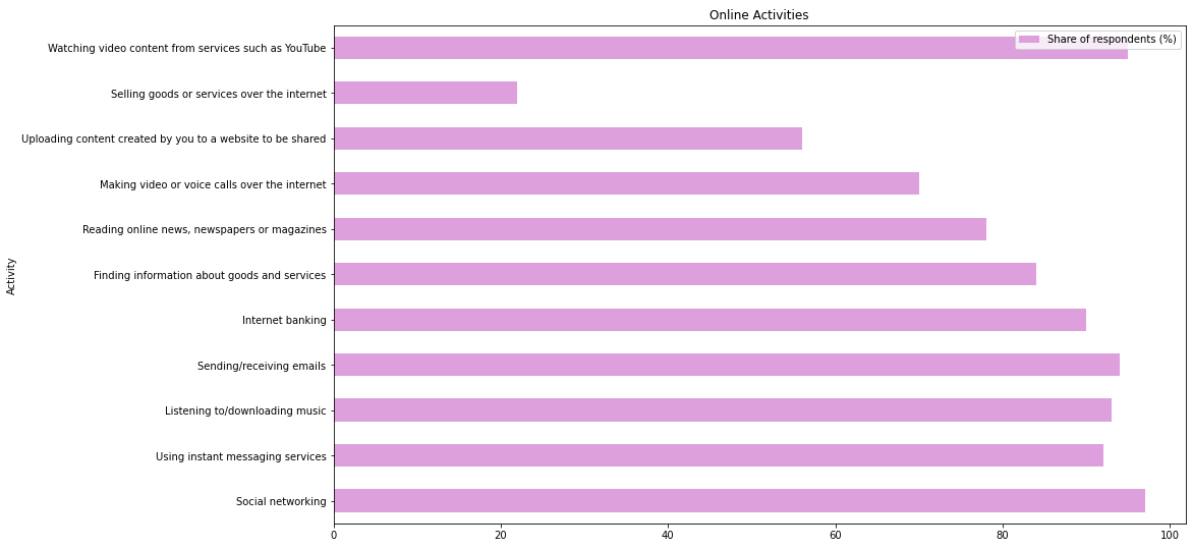
```
delete_columns('Share of respondents (f)', df_2_copy)
df_2_copy
```

Out[59]:

Activity	Share of respondents (%)
Social networking	97
Using instant messaging services	92
Listening to/downloading music	93
Sending/receiving emails	94
Internet banking	90
Finding information about goods and services	84
Reading online news, newspapers or magazines	78
Making video or voice calls over the internet	70
Uploading content created by you to a website to be shared	56
Selling goods or services over the internet	22
Watching video content from services such as YouTube	95

In [183]:

```
df_2_barh = df_2_copy.plot.barh(figsize=(15, 9), color='#DDA0DD', title="Online Acti
```



The bar chart above displays the percentage of correspondents that do each activity. According to the data, 97% of the correspondents used social networking sites, 95% watched video content online and 94% checked/sent email.

The data shows that there is a fair bit of activities that Generation Z do online, as the majority of the data was over 80%. On the lower end 22% of the correspondents reported selling goods or services over the internet. This led me to believe that the internet was used more for fun activities, or as a way to take a break. As most of the activities were not heavily connected to work.

On the other hand, 94% of correspondents checked/sent emails, and who's to say that they weren't work emails. However, it seemed that there was a lot of ambiguity from the data. There was not necessarily more that could have been done, it was just the nature of the categories- they were not very specific. Although if they had been too specific the data wouldn't have been accurate or representative anymore. I believed that formulating an analysis on this would lead to a lot of bias, as there would have been a lot of assumptions from the one analysing, as the data wasn't very niched.

All in all, it was pretty hard to come to a solid conclusion from this data, there were more things that I would have needed to know to reach a valid conclusion. But what was quite clear was that most Generation Zers spend time on social networking sites.

3.3.2 What social apps do Generation Z favour, and what does that say about them?

I spoke a lot about Generation Z using social media, but what applications have they favoured, and what do they say about them? What impact has it had on them as a whole?

I decided to explore df_1 to approach this. It is a dataframe that represents the number of users (from Generation Z) that were using each social platforms in millions. It also gives data for the specific year, and predicts data for years to come. I transformed the data into a line chart, so that it could display the growth (whether positive or negative) of each social platform over the years.

In [48]:

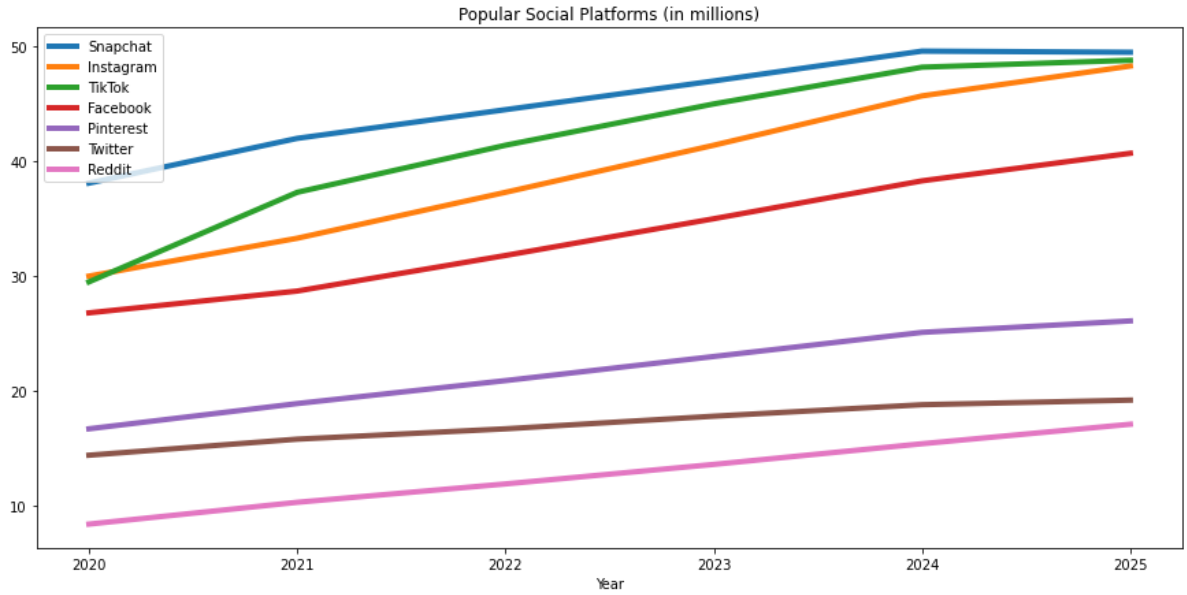
```
df_1_copy = df_1
df_1_copy.set_index('Year', inplace=True)

df_1_copy
```

Out[48]:

Year	Snapchat	Instagram	TikTok	Facebook	Pinterest	Twitter	Reddit
2020	38.1	30.0	29.5	26.8	16.7	14.4	8.4
2021	42.0	33.3	37.3	28.7	18.9	15.8	10.3
2022	44.5	37.3	41.4	31.8	20.9	16.7	11.9
2023	47.0	41.4	45.0	35.0	23.0	17.8	13.6
2024	49.6	45.7	48.2	38.3	25.1	18.8	15.4
2025	49.5	48.3	48.8	40.7	26.1	19.2	17.1

In [189]:

```
df_1_line_chart = df_1_copy.plot.line(figsize = (15, 7), lw=4, title="Popular Social
```



In the chart above it was clear that there was an increase in the number of users on the various social platforms. This was very representative of the world that we are living in today, being that it is a rapidly growing technological world. However, I decided to look at the data more closely, seeing as they were all increasing, and set to continue increasing.

I noticed that there was a significant difference in users (around 10 million) between the purple (Pinterest), brown (Twitter), pink (Reddit) group, and the blue (Snapchat), green (TikTok), orange (Instagram), red (Facebook) group. This led me to consider what these groups each had in common. I came to the realisation that the group that had the more users (blue group) were the social platforms that had an emphasis on short videos, as opposed to being more text based, with pictures and videos here and there.

*Point to mention- the highest of the lower group was Pinterest, which is set to continue increasing in users steadily. Pinterest is more visual based than the other two, though it lacks as much social interaction- the concept of short videos has slowly been introduced onto the app.

Why is that the case? Why are Generation Zers choosing videos and pictures, over more text-based platforms. I had an inkling that it was due to two main things. One that videos are more personal, and two because humans are visual beings.

We have gotten very far in technology, and that has expanded to how we share online. More and more people are using filters or are editing their pictures before posting them. While videos can still be edited, and people can use filters, videos still come across as more authentic. There are also the perfect way for people to share short clips, vlog about their lives. In addition, humans are very visual and seeing things helps the experience feel more real. Many people who follow strangers online feel as though they know those strangers, due to images and video content that they have shared online.

What can we infer from this? Upon analysing the data I realised that there could be a connection between the visual nature of humans and body dysmorphia or body comparison. Body comparison is inevitable when people use such visual platforms where people post their best angles.

The chart below displays how often Generation Zers post selfies. The data came from the data I acquired from URL[6] data.

In [110]:

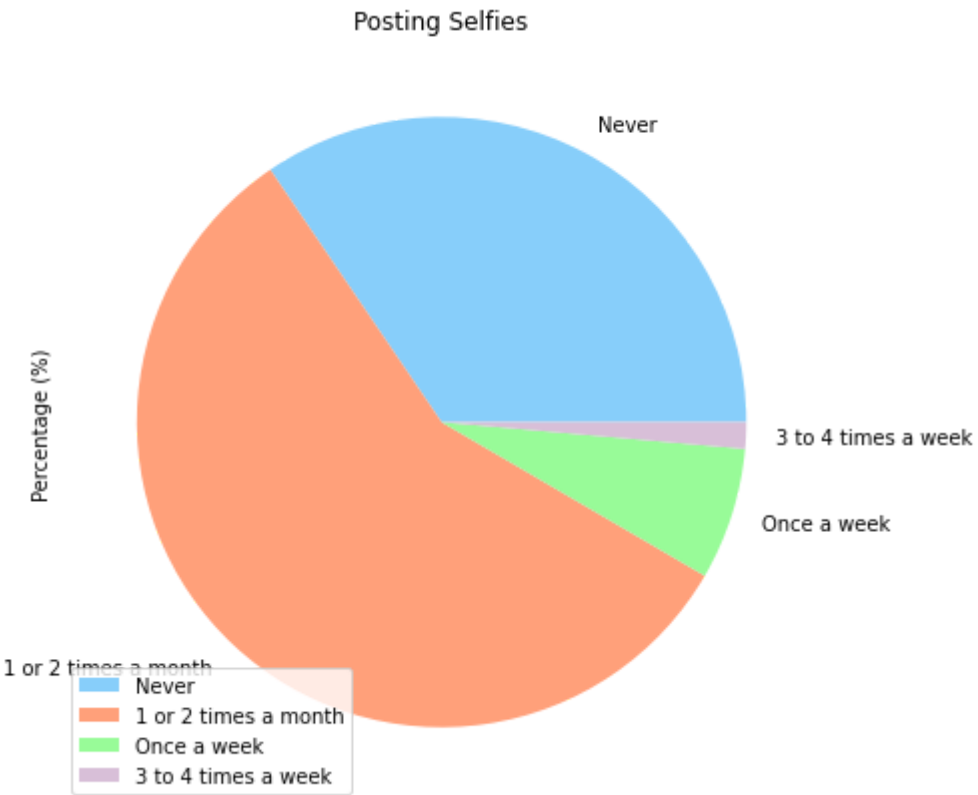
```
posting_selfies_copy = posting_selfies
posting_selfies_copy.set_index('Posting Selfies', inplace=True)
delete_columns('Percentage (f)', posting_selfies_copy)

posting_selfies_copy
```

Out[110]:

Posting Selfies	Percentage (%)
Never	34.5
1 or 2 times a month	57.1
Once a week	7.0
3 to 4 times a week	1.4

In [178]:

```
posting_selfies_pie = posting_selfies_copy.plot.pie(y='Percentage (%)', figsize=(7,
                                                    colors=["#87CEFA", "#FFA07A", "#
                                                    ,title="Posting Selfies")
```



The pie chart showed that people are not posting frequently. 57.1% of people post once or twice a month, and 34.5% never post at all. The makes up a total of 91.6% who don't post too frequently. This was a lot to unpack. Why aren't they posting if they are constantly checking social media?

Well, this may have come as a surprise since we saw earlier that Generation Zers use social platforms a lot. All in all, I concluded that most users were ghosting, a colloquial term in social media that meant that they were simply just observing but not contributing to what they observed. They kept their social platforms so they could get insight on what was happenning around the world- since we saw earlier that, that was their primary news source. In addition, they keep it so that they can be kept up to date with what is trending, where they get fed a ton of perfection through media, leading to a vicious cycle of body comparison.

3.4 What are the mental health impacts of all this?

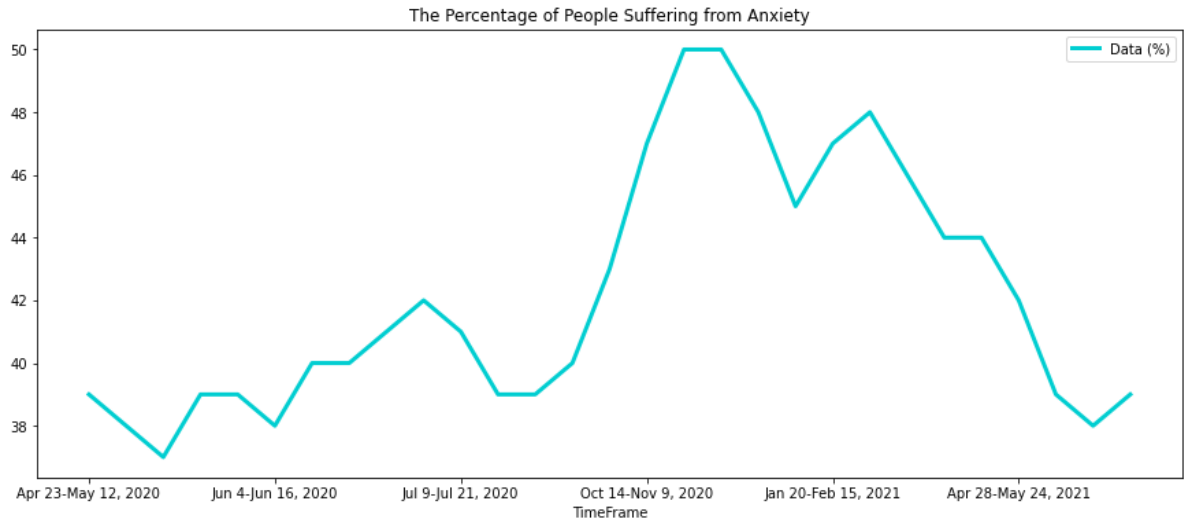
We already know that social media leads to body dysmorphia, and I alluded to other issues earlier, but what other mental health impacts are there?

3.4.1 Are people suffering from anxiety?

Anxiety is descibed as "*feeling of unease, such as worry or fear, that can be mild or severe*" [4], and is seen at work across a variety of age groups. The graph below displays the percentage of Generation Zers that experienced anxiety from 2020 to 2021.

In [174]:

```
df_3_line = df_3.plot.line(figsize=(15, 6), lw=3, color="#00CED1", x='TimeFrame', y=
                           title="The Percentage of People Suffering from Anxiety")
```



The graph above shows that the number of Generation Zers that suffer from anxiety has not been very constant. There was a peak around Oct to Nov 2020, which could be attributed to the pandemic. Analysing this brought me to the conclusion that there were two ways that I would want to look at the data. First, by ignoring the peaks caused by a global pandemic, as people are bound to become more anxious, and then by looking at the whole data including the peaks.

By ignoring the peaks, I concluded that on average 40% of Generation Z suffered from anxiety- which is incredibly high. Including the peaks brought the average up to about 45%, with the max at one point having been 50%. This meant that there was a time when 1 in 2 Generation Zers suffered from anxiety, which is a collosal amount of people.

What is the cause of their anxiety? We have already seen that they have spent a lot of their time online, and that many compare theirselves to others, all these things are catalysts for anxiety. Though I wasn't able to say for certain that technology was the main cause of their anxiety, I was able to conclude that it plays a large role in causing it, or perhaps adding to it.

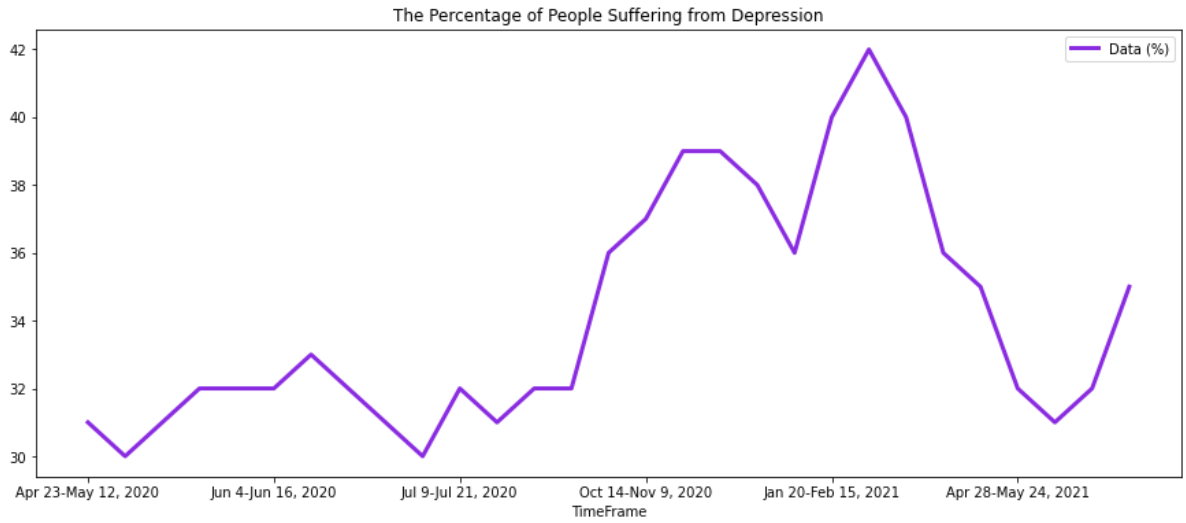
3.4.2 Are people suffering from depression?

Depression is defined as "*a low mood that lasts for weeks or months and affects your daily life*" [4]. At the begininning of this proposal I stated that Generation Z were seen as a depressed generation. Let's see how true of a statement that was.

Below I created another line graph that represented the percentage of Generation Zers that suffered from depression. It like the anxiety one, was measured over the course of 2020 and 2021.

In [175]:

```
df_4_line = df_4.plot.line(figsize=(15, 6), lw=3, color="#8A2BE2", x='TimeFrame', y=
                           title="The Percentage of People Suffering from Depression"
```



From the chart above I inferred that the number of those that had depression had been increasing over the month until a certain point. I assumed that this was due to an after affect of the pandemic, and that at the point where it dropped, the level of depression was slowly going back to the norm- though it looked to be rising again.

Due to this, it was very hard to draw a conclusion from this data. This proposal was about the impact of technology/internet/social media on mental health not a pandemic. I concluded that the data was accurate however not useful for what I was trying to find an answer too. I then decided to just focus on the part before the rapid jump, and came to this final conclusion. On average about 32% of Generation Z suffer from depression, which is almost 1 in 3, which is another pretty high stat. Therefore to some degree, there was plenty of truth behind my initial statement.

---

4 Conclusion and Critical Evaluation

In terms of my objectives, I had originally set out to analyse the correlation between body dymorphia and social media with Generation Z, and eventually decided to broaden the proposal to media and technology as a whole, and their impact on the mental health of Generation Z- which I mentioned before.

I was able to analyse what I finally set out to, and I came to the conclusion that there was in fact a correlation. Though I will say that the strength of that correlation most definitely lies heavily on the data that you are able to acquire. I was able to attain a good amount of data that helped me reach my conclusion. Though I found that some of the data could have been impacted by other events that happened over the world, or other factors in general. This is because the data I attained wasn't as niched as I would have liked it to have been, in order to feel more confident with what I discovered- though I still believe that I reached a solid conclusion.

I explored the forms of media that Generation Z take in, and also whether they were the most reliable forms of media to use. I also explored whether people that frequented social platforms more often, were more inclined to develop poor mental health.

If I were to have more time, I would go with my original objective, and spend a significant amount of time finding data, or perhaps even conducting surveys to acquire my own data. I believe that this was an important topic as many of this young generation are facing these things, as a result of their technologically influenced upbringing, and they are the hope for tomorrow, so it is in our best interest to ensure that everyone has the opportunity to improve on their mental health.

In terms of my aims, I set out to first find the data needed to explore my objects. I was able to successfully do this, though I do acknowledge some bias present in some of the data, due to smaller study sizes in some cases. There may also have been a lack in diversity in the data I collected too. However, I was still able to obtain some solid data, that was able to help me draw a conclusion.

Another of my aims was to consider the methods needed to extract said data and then extact it, which I did. I spent the majority of the extracting time scrapping data, and if I didn't scrap I extracted from excel spreadsheets. With that I achieved another of my aims, which was to clean the data so that I would be able to analyse it- which I did, and I was also able to pick up on strange patterns due to that. My final aim was to carry out exploratory data analysis by identifying links between data, and also potential flaws. I believe that I met this aim, though if I had more time, or I was to approach this again, there are more correlations that I would touch on, and I would also go into more detail, and use more tools.

5 References and Resources

5.1 References

[1] Harvard Business Review. (2019, October 7) *Research: People Want Their Employers to Talk About Mental Health* [online]. Available: https://hbr.org/2019/10/research-people-want-their-employers-to-talk-about-mental-health (https://hbr.org/2019/10/research-people-want-their-employers-to-talk-about-mental-health)
[2] NHS. *Body dysmorphic disorder (BDD)* [Online]. Available: https://www.nhs.uk/mental-health/conditions/body-dysmorphia/#overview (https://www.nhs.uk/mental-health/conditions/body-dysmorphia/#overview)
[3] The Annie E. Casey Foundation. (2021, March 3) *Generation Z and Mental Health* [online]. Available: https://www.aecf.org/blog/generation-z-and-mental-health (https://www.aecf.org/blog/generation-z-and-mental-health)
[4] NHS. *Generalised Anxiety Disorder in Adults* [Online]. Available: https://www.nhs.uk/mental-

health/conditions/generalised-anxiety-disorder/overview/#:~:text=Anxiety%20is%20a%20feeling%20of,medical%20test%20or%20job%20interview (https://www.nhs.uk/mental-health/conditions/generalised-anxiety-disorder/overview/#:~:text=Anxiety%20is%20a%20feeling%20of,medical%20test%20or%20job%20interview)
[5] NHS. *Clinical Depression* [Online]. Available: https://www.nhs.uk/mental-health/conditions/clinical-depression/overview/ (https://www.nhs.uk/mental-health/conditions/clinical-depression/overview/)

5.2 Resources used

Webscraping

- Frequency of using selected news sources among Generation Z in the United States as of February 2022 | https://www.statista.com/statistics/1124119/gen-z-news-consumption-us/ (https://www.statista.com/statistics/1124119/gen-z-news-consumption-us/)
- Number of Gen Z users in the United States on selected social media platforms from 2020 to 2025 | https://www.statista.com/statistics/1276021/instagram-snapchat-tiktok-gen-z-users/ (https://www.statista.com/statistics/1276021/instagram-snapchat-tiktok-gen-z-users/)
- Activities performed online by Generation Z in Great Britain in 2020 | https://www.statista.com/statistics/1119977/gen-z-internet-activities-in-great-britain/ (https://www.statista.com/statistics/1119977/gen-z-internet-activities-in-great-britain/)
- Number of Gen Z that suffer from nerves, anxiety, and more in the US in 2020 | https://datacenter.kidscount.org/data/tables/11209-adults-ages-18-to-24-who-felt-nervous-anxious-or-on-edge-for-more-than-half-of-the-days-or-nearly-every-day-in-the-past-week (https://datacenter.kidscount.org/data/tables/11209-adults-ages-18-to-24-who-felt-nervous-anxious-or-on-edge-for-more-than-half-of-the-days-or-nearly-every-day-in-the-past-week)
- Number of Gen Z that are depressed and hopeless in the US in 2021 | https://datacenter.kidscount.org/data/tables/11211-adults-ages-18-to-24-who-felt-down-depressed-or-hopeless-for-more-than-half-of-the-days-or-nearly-every-day-for-the-past-two-weeks (https://datacenter.kidscount.org/data/tables/11211-adults-ages-18-to-24-who-felt-down-depressed-or-hopeless-for-more-than-half-of-the-days-or-nearly-every-day-for-the-past-two-weeks)
- Number of Gen Z reporting poor mental health | https://datacenter.kidscount.org/data/tables/11202-young-adults-ages-18-to-24-reporting-zero-poor-mental-health-days-in-the-past-month (https://datacenter.kidscount.org/data/tables/11202-young-adults-ages-18-to-24-reporting-zero-poor-mental-health-days-in-the-past-month)
- The link between social media and body dysmorphia | https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8001450/ (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8001450/)