FOUNDATION OF DATA SCIENCE

HIT140

ASSESSMENT 2

Submitted by

Name	Student Number
Binaya Sedhai	382884
Easinur Rashid	374095
Riya Roy	383607
Anika Tamanna Riya	379875

INTRODUCTION

The given files consists of three interconnected datasets, each shedding light on different aspects of adolescents' lives. The first dataset provides demographic information, identifying individuals by gender, minority status, and deprivation levels. The second dataset tracks their media consumption, including time spent on computers, video games, smartphones, and television, distinguishing between weekdays and weekends. The third dataset focuses on wellbeing, with respondents rating their optimism, energy, and other mental health aspects on a 5-point Likert scale.

By linking these datasets through respondent IDs, we aim to analyze how digital screen time affects an individual's wellbeing, with a focus on how factors like deprivation and minority status may influence this relationship.

DATA OVERVIEW

This data delivery includes three related sets of data that address different aspects of adolescent's demographic data, media consumption, and well-being.

- Dataset 1: includes descriptive characteristics of 120 plus thousand adolescents who have been identified with a different ID number for each. Preached fixed variables encompass self-reported gender with one being male and zero for the remaining, ethnic minority status also with one being from ethnic minority and zero for the rest, and deprivation status with one being from the highly deprived areas and zero otherwise.

- Dataset 2: About the media consumption habits themselves, the second dataset describes how many hours these people spend at the computer, playing video games, using smartphones, as well as watching television. All these usages are distinguished by whether they occurred on a weekday or a weekend.

- Dataset 3: evaluates the respondents' welfare in all the facets in terms of optimism, usefulness, relaxation, energy, problem solving power, confidence, love and cheer. Well-being data are collected by using a 5-point Likert scale that comprises of choices from "None of the time" to "All of the time".
  The data sets are connected by a respondent ID number so that overall demographic characteristics and digital media usage impacted the adolescents' well-being could be examined.

DATA OVERVIEW

Dataset 1 contains basic demographic information of more than 120,000 adolescent respondents, each person identified by a unique ID number.

- ID: A unique number identifying a respondent
- gender: Self-reported gender (1 for male and 0 otherwise)
- minority: 0 as belonging to the majority ethnic group of the country; 1 otherwise
- deprived: 1 as residing in localities with high deprivation indices i.e. an area with high scores on unemployment, crime, poor public services, and barriers to housing etc.; 0 otherwise

Dataset 2 shows the watch time for the person by the ID from dataset 1

- ID (A unique number identifying a respondent),
- C_we (Number of hours using computers per day on weekends),
- C_wk (Number of hours using computers per day on weekdays),
- G_we (Number of hours playing video games per day on weekends),
- G_wk (Number of hours playing video games per day on weekdays),
- S_we (Number of hours using a smartphone per day on weekends),
- S_wk (Number of hours using a smartphone per day on weekdays),
- T_we (Number of hours watching TV per day on weekends),
- T_wk (Number of hours watching TV per day on weekdays);

DATA OVERVIEW

Dataset 3 provide wellbeing score for the individual in some categories.

- ID (A unique number identifying a respondent),
- Optm (I have been feeling optimistic about the future),
- Usef (I have been feeling useful),
- Relx (I have been feeling relaxed),
- Intp (I have been feeling interested in other people),
- Engs (I have had the energy to spare),
- Dealpr (I have been dealing with problems well),
- Thkclr (I have been thinking clearly),
- Goodme (I have been feeling good about myself),
- Clsep (I have been feeling close to other people),
- Conf (I have been feeling confident),
- Mkmind (I have been able to make up my own mind about things),
- Loved (I have been feeling loved),
- Intthg (I have been interested in new things),
- Cheer (I have been feeling cheerful);

It provides the score in 5 level

1.None of the time
2.Rarely
3.Some of the time
4.Often
5.All of the time

```
n [ ]:
print("Descriptive Statistics for Dataset 1:\n\n")

# Value counts for categorical variables
print("\tGender distribution:")
print(df1['gender'].value_counts())
print("\n")

print("\tMinority distribution:")
print(df1['minority'].value_counts())
print("\n")

print("\tDeprivation distribution:")
print(df1['deprived'].value_counts())
print("\n")
```

```
Descriptive Statistics for Dataset 1:


        Gender distribution:
gender
0    62962
1    57153
Name: count, dtype: int64


        Minority distribution:
minority
0    91196
1    28919
Name: count, dtype: int64


        Deprivation distribution:
deprived
0    67889
1    52226
Name: count, dtype: int64
```

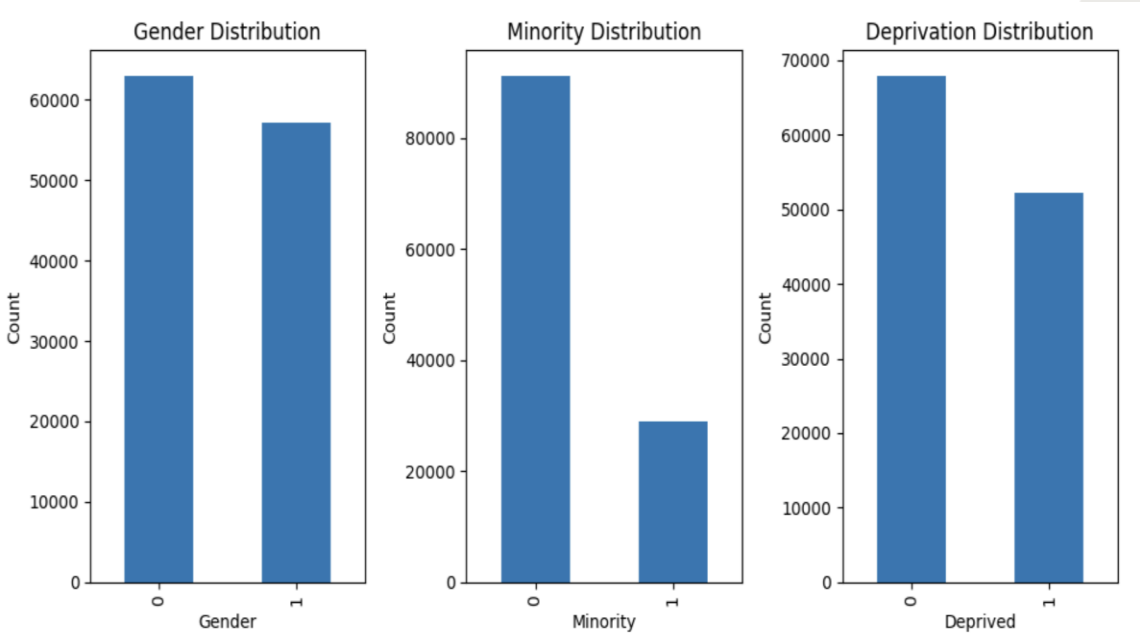
DESCRIPTIVE STATISTIC- DATASET 1

- Gender Distribution
- Categories:
  - Gender 0 (e.g., Male/Female): 62,962 individuals
  - Gender 1 (e.g., Female/Male): 57,153 individuals
- Explanation:
  - The dataset is relatively balanced between the two gender categories, with a slight majority of individuals in the "Gender 0" category.
  - This suggests that there is no overwhelming skew in terms of gender representation in the dataset.
- Minority Distribution
- Categories:
  - Minority 0 (Non-minority): 91,196 individuals
  - Minority 1 (Minority): 28,919 individuals
- Explanation:
  - The majority of individuals in the dataset (roughly 76%) do not belong to a minority group.
  - A smaller portion (approximately 24%) of the population falls under the minority category.
  - This distribution could reflect a common societal demographic where non-minority groups are larger than minority groups.
- Deprivation Distribution
- Categories:
  - Deprived 0 (Non-deprived): 67,889 individuals
  - Deprived 1 (Deprived): 52,226 individuals
- Explanation:
  - The dataset has a significant portion of individuals classified as "deprived" (43%).
  - Although there is a slight majority of non-deprived individuals (57%), the proportion of deprived individuals is substantial, which might indicate potential areas of concern for welfare and economic support.

Descriptive Statistics Dataset 1

The code generates a figure with three bar charts to visualize the distribution of three categorical variables- gender, minority, and deprived. The x-axis will show the categories of gender (e.g.; male, female) and the y-axis will show the count of occurrences for each gender. Whereas the x-axis will display the categories(e.g.; yes/no for minority status) and the y-axis will display the corresponding count. Similarly, the x-axis will show the categories (e.g.; deprived/ not deprived) and y-axis will show how many times each category appears. Each chart helps to visually understand the distribution of these categorical variables in the dataset. From the bar chart we can see that there are more females as compared to male. In second bar it shows that there are more majority than minority. In last bar we can see that non-deprived category is more than deprived.

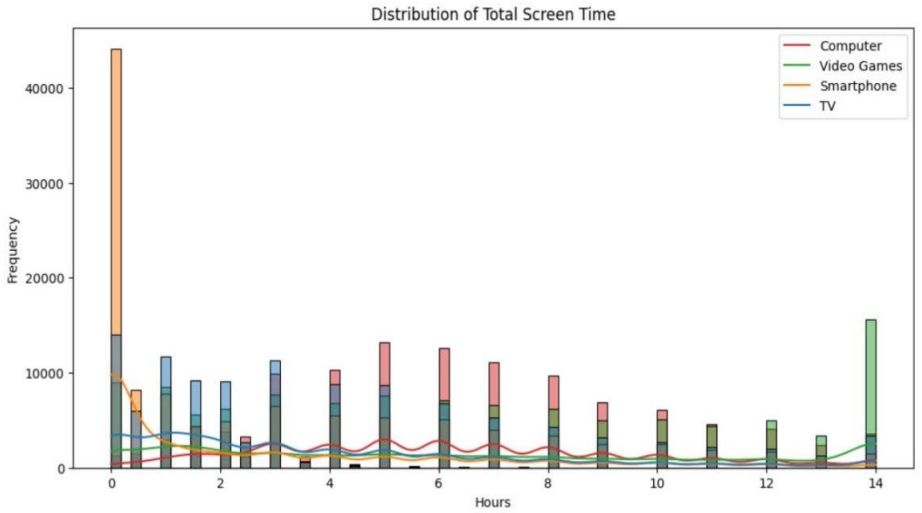
```
import matplotlib.pyplot as plt
# Bar charts for categorical variables
plt.figure(figsize=(10, 5))

plt.subplot(1, 3, 1)
df1['gender'].value_counts().plot(kind='bar')
plt.title('Gender Distribution')
plt.xlabel('Gender')
plt.ylabel('Count')

plt.subplot(1, 3, 2)
df1['minority'].value_counts().plot(kind='bar')
plt.title('Minority Distribution')
plt.xlabel('Minority')
plt.ylabel('Count')

plt.subplot(1, 3, 3)
df1['deprived'].value_counts().plot(kind='bar')
plt.title('Deprivation Distribution')
plt.xlabel('Deprived')
plt.ylabel('Count')

plt.tight_layout()
plt.show()
```



Descriptive Statistics Dataset 2

The overall screen time for a variety of devices, such as computers, video games, smartphones, and televisions, is shown on the graph. In the stacked bar chart, these four variables are shown in various colors (red for PCs, green for video games, orange for cellphones, and blue for televisions). The y-axis shows frequency, while the x-axis shows amount of time spent in front of the screen. For example, red is used for computers and green for video games on each screen time form. The data is displayed as stacked bars with lines that represent smoothed trends for each kind of device. Every screen variable's total observations. Mean: the typical amount of time spent on screens across all categories. The recorded screen time, both minimum and maximum, for every device and time frame (ranging from 0 to 7 hours). 25% in the first quartile, 50% in the median, and 75% in the third. Weekend screen time surpasses weekday screen time, especially when it comes to computers and video games. This data explores correlations between these variables to see if people spend much time to spend more time on other devices example- smartphones. Cumulative hours spent across all four devices for each observation and hours of screen time range from 0 to 14.

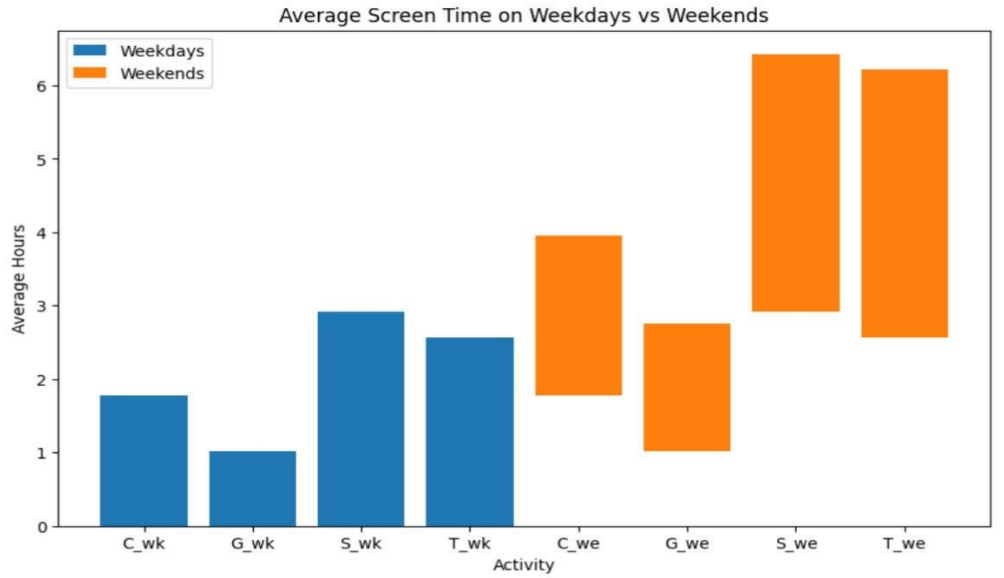
```
# Visualize total screen time
plt.figure(figsize=(12, 6))
sns.histplot(df2[['Total_Computer_We', 'To
plt.title('Distribution of Total Screen Ti
plt.xlabel('Hours')
plt.ylabel('Frequency')
plt.legend(['Computer', 'Video Games', 'Sm
plt.show()
```



Dataset 2- Continued.

The number of rows indicates how many people or how long the observations were made. Eight columns make up the dataset, and each one shows the average number of hours spent using various screens throughout the week and on the weekends. This is compared to weekend usage in the graph. The four screen activities—computer, video games, smartphones, and television—are represented by the X-axis. The average number of hours that people spend on these activities is shown on the y-axis. The graph makes it possible to compare visually how much time is spent on different activities during the week versus the weekend. The columns show correlation and behavior changes. In highlights, people spend more time on screen on weekdays, particularly when watching TV or using smartphones. The dataset's comparison between weekdays and weekends shows relationships between screen time and daily activity patterns.

```
plt.bar(weekday_avg.index, weekday_avg.valu
plt.bar(weekend_avg.index, weekend_avg.valu
plt.title('Average Screen Time on Weekdays
plt.xlabel('Activity')
plt.ylabel('Average Hours')
plt.legend()
plt.show()
```



Descriptive Statistics - Dataset 3

The dataset provides wellbeing values, with the mean showing average mood, max highlighting peak feelings, and the standard deviation showing variability. By comparing these to digital screen time, we can assess if screen usage impacts how people feel. Extreme values (max/min) help define the range of emotions and screen time, offering a clearer picture of potential correlations between them.

	ID	Optm	Usef	Relx	\
count	1.025800e+05	102580.000000	102580.000000	102580.000000	
mean	1.059921e+06	3.276087	3.107428	3.097826	
std	3.478290e+04	0.997897	0.953128	1.015441	
min	1.000001e+06	1.000000	1.000000	1.000000	
25%	1.029723e+06	3.000000	3.000000	2.000000	
50%	1.059760e+06	3.000000	3.000000	3.000000	
75%	1.090162e+06	4.000000	4.000000	4.000000	
max	1.120115e+06	5.000000	5.000000	5.000000	

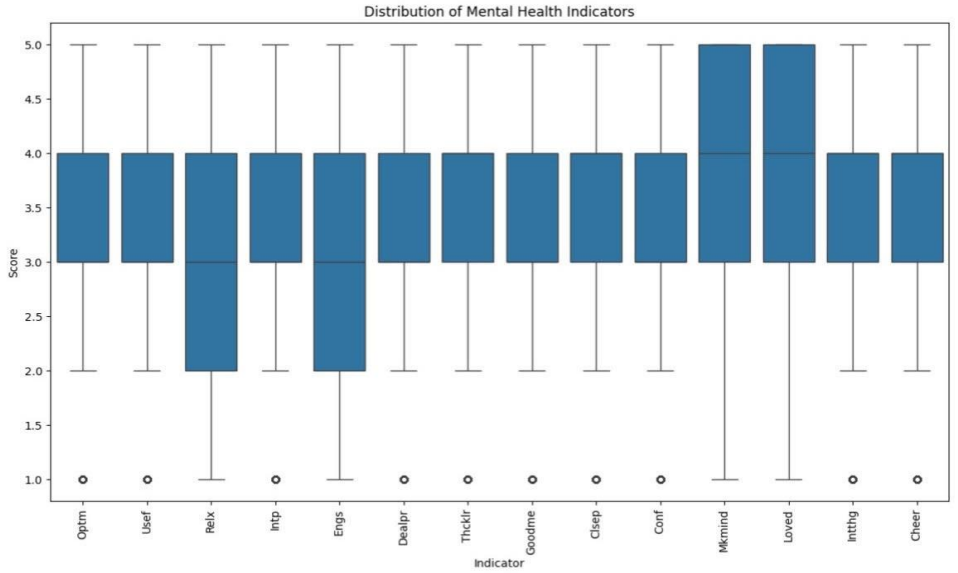
	Intp	Engs	Dealpr	Thcklr	\
count	102580.000000	102580.000000	102580.000000	102580.000000	
mean	3.272314	3.048050	3.369448	3.488039	
std	1.018537	1.076483	1.049365	1.018274	
min	1.000000	1.000000	1.000000	1.000000	
25%	3.000000	2.000000	3.000000	3.000000	
50%	3.000000	3.000000	3.000000	4.000000	
75%	4.000000	4.000000	4.000000	4.000000	
max	5.000000	5.000000	5.000000	5.000000	

	Goodme	Clsep	Conf	Mkmind	\
count	102580.000000	102580.000000	102580.000000	102580.000000	
mean	3.273250	3.557116	3.308491	3.851267	
std	1.126084	1.031297	1.115874	0.974468	
min	1.000000	1.000000	1.000000	1.000000	
25%	3.000000	3.000000	3.000000	3.000000	
50%	3.000000	4.000000	3.000000	4.000000	
75%	4.000000	4.000000	4.000000	5.000000	
max	5.000000	5.000000	5.000000	5.000000	

Descriptive Statistics - Dataset 3

Dataset 3 represents mental health indicators for individuals, focusing on various emotional and psychological well-being metrics. The rows in the dataset represent the number of individuals being assessed; each row corresponds to an individual's self-reported scores on different mental indicators. The dataset comprises 16 columns, including ID and 15 mental health indicators. Variables have a score ranging from low to high (1-5). Central tendency is the mean; Dispersion is each indicator's standard deviation and range (min, max). Most indicators have an average score between 3 and 4, indicating a moderate to high sense of well-being across the population. A standard deviation of around 1 shows a moderate variability in how people feel in these areas. Here, the X-axis represents a specific mental health condition (relaxation and confidence). Relationships between these variables could help analyze how different aspects of mental health are correlated. People who score high on love may also score higher on confidence or feeling good about themselves. Dataset 2 explores the direct impact and indirect relationship. This dataset provides how digital screen time may impact mental health and well-being dimensions.

```
# Visualize the distribution of each indic
df3_melted = df3.melt(id_vars='ID', var_na
plt.figure(figsize=(14, 8))
sns.boxplot(x='Indicator', y='Score', data
plt.title('Distribution of Mental Health I
plt.xticks(rotation=90)
plt.show()
```



Relationship Between Datasets

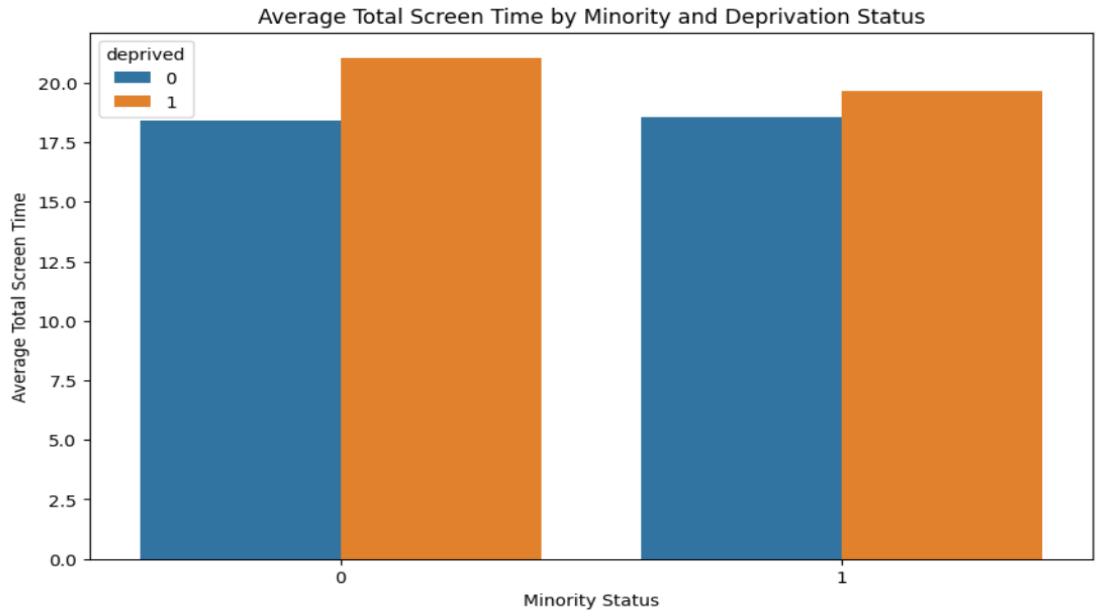
The Python method collects data on minority and deprivation status with total screen time, which is determined by adding several screen time columns. These categories classify the data, calculate the average total screen time for each group, and use color to create a bar chart showing each group's deprivation status.

The average overall screen time for minority and disadvantaged groups is shown in the bar chart. The x-axis represents minority status, while the y-axis represents average screen time. Bar colors show the deprivation status and highlight the variations in screen time according to these criteria, making comparing and analyzing the relationships between them simple.

The average total screen time is depicted in the figure, and the blue bars (deprivation status 0) represent about eighteen hours for minority status 0. Individuals with orange bars (deprivation class 1) have increased screen time.

The average total screen time is depicted in the figure, and the blue bars (deprivation status 0) represent about eighteen hours for minority status 0. Individuals with orange bars (deprivation class 1) have increased screen time.

```
# Calculate total screen time for each individual
df2['Total_Screen_Time'] = df2[['C_we', 'C_wk', 'G_we', 'G_wk', 'S_we', 'S_wk', 'T_we', 'T_wk']].su

# Merge relevant columns from df1 and df2
df_merged = pd.merge(df1[['ID', 'minority', 'deprived']], df2[['ID', 'Total_Screen_Time']], on='ID'

# Group by minority and deprived status and calculate average total screen time
grouped = df_merged.groupby(['minority', 'deprived'])['Total_Screen_Time'].mean().reset_index()

# Plot the results
plt.figure(figsize=(10, 6))
sns.barplot(x='minority', y='Total_Screen_Time', hue='deprived', data=grouped)
plt.title('Average Total Screen Time by Minority and Deprivation Status')
plt.xlabel('Minority Status')
plt.ylabel('Average Total Screen Time')
plt.show()
```



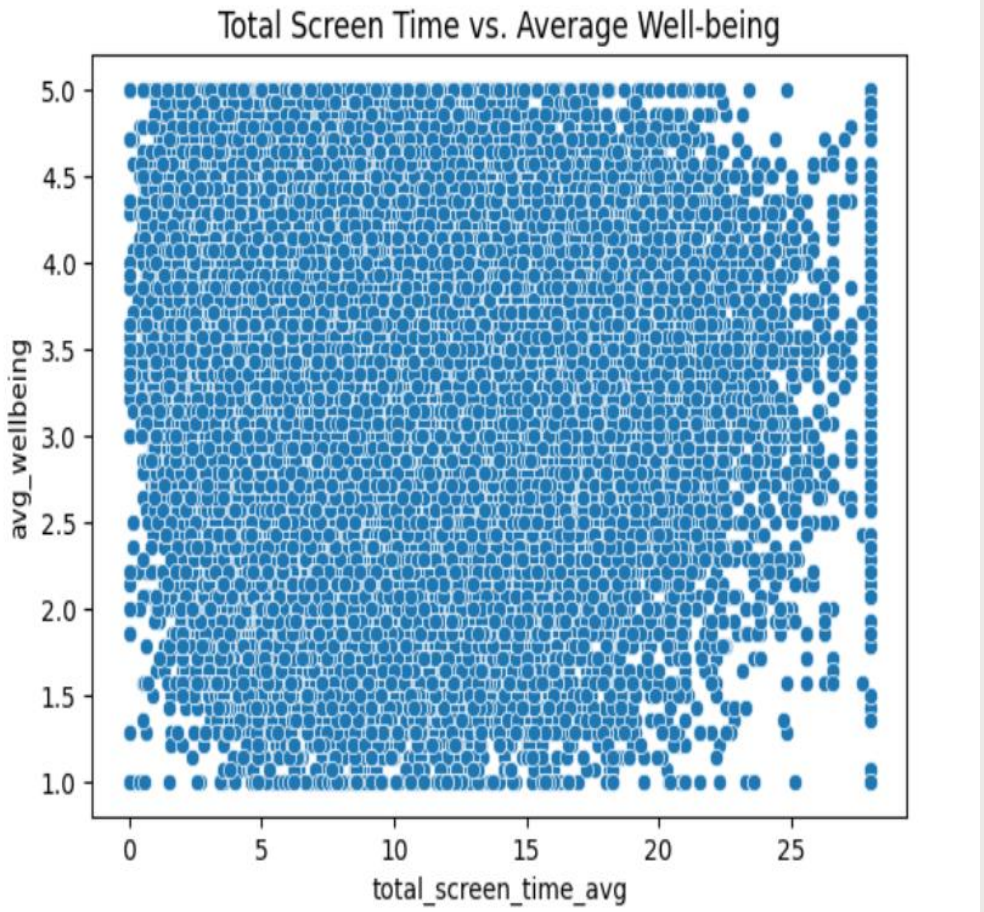
Total Screen Time vs Average Well-Being

The given Python code uses the Seaborn module to produce a scatter plot. The graphic illustrates the connection between average well-being and total screen time. The needed data is selected from the DataFrame 'df', the required libraries are imported, and the scatter plot is created. The average overall screen time is shown on the x-axis, while the average well-being score is on the y-axis.

This code produces a scatter plot as an image, which displays the distribution of well-being scores for different screen time amounts. Each data point on the plot represents individuals from the DataFrame 'df'. We can determine whether there is a pattern or association between screen time and well-being scores by visually examining the positioning of the dots.

For instance, if there is a positive link, we may observe a concentration of dots where better well-being scores correlate with more screen time. On the other hand, in the event of a negative link, we may observe a concentration of dots corresponding to poorer well-being scores and increased screen time. We can visually examine this relationship and see any potential patterns or trends with the help of the scatter plot.

```
sns.scatterplot(x='total_screen_time_avg', y='avg_wellbeing', data=df)
plt.title('Total Screen Time vs. Average Well-being')
plt.show()
```

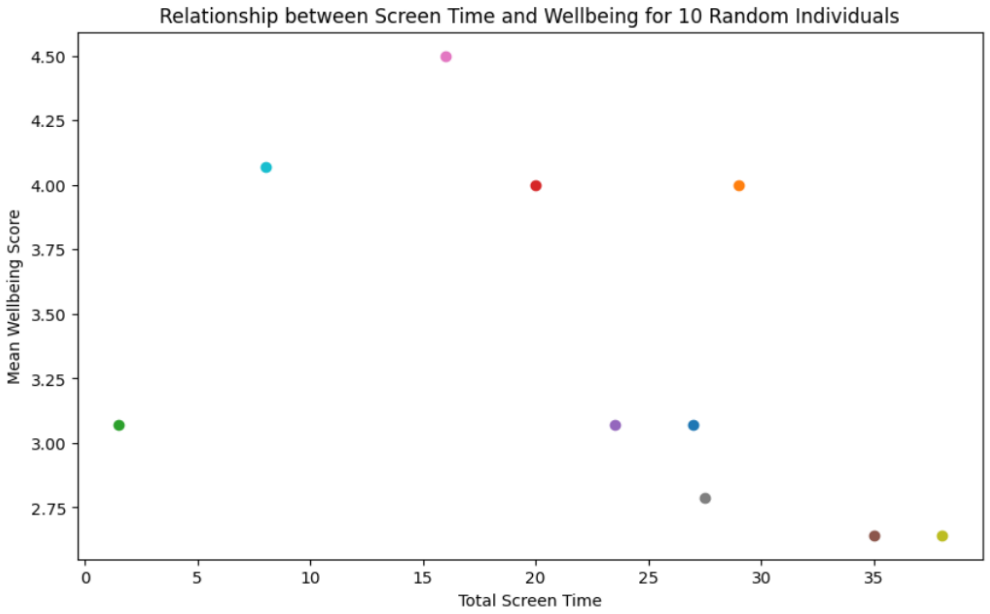


Relationship between screen time and well-being of 10 random individuals

The supplied Python code computes the mean well-being score and total screen time for each participant combines these data frames and then chooses ten participants at random for analysis. Using a scatter plot, it then shows how these ten people's mean well-being and total screen time relate to one another.

This code creates a scatter plot image that illustrates the correlation between the mean well-being of ten randomly chosen individuals and their total screen time. With the mean well-being score (on a scale of 1 to 5) on the y-axis and total screen time (measured in hours) on the x-axis, each data point on the plot represents everyone. In the sample population, this graphic aids in finding trends or connections between various variables. For instance, a negative connection could be shown in the plot, suggesting that the sample's well-being scores are negatively correlated with more screen time.

```
# Calculate total screen time for each individual in df2
df2['Total_Screen_Time'] = df2[['C_we', 'C_wk', 'G_we', 'G_wk', 'S_we', 'S_wk', 'T_we', 'T_wk']].sum(axis=1)

# Calculate mean wellbeing score for each individual in df3
df3['Mean_Wellbeing'] = df3[['Optm', 'Usef', 'Relx', 'Intp', 'Engs', 'Dealpr', 'Thcklr', 'Goodme', 'Clsep', 'Conf', 'Mkmi

# Merge the relevant data from df2 and df3
df_merged = pd.merge(df2[['ID', 'Total_Screen_Time']], df3[['ID', 'Mean_Wellbeing']], on='ID')

# Select 10 random individuals
import random
random_ids = random.sample(list(df_merged['ID']), 10)
df_sample = df_merged[df_merged['ID'].isin(random_ids)]

# Plot the relationship between total screen time and mean wellbeing for the sampled individuals
plt.figure(figsize=(10, 6))
for _, row in df_sample.iterrows():
  plt.plot(row['Total_Screen_Time'], row['Mean_Wellbeing'], marker='o')
plt.xlabel('Total Screen Time')
plt.ylabel('Mean Wellbeing Score')
plt.title('Relationship between Screen Time and Wellbeing for 10 Random Individuals')
plt.show()
```



T-test for Each Type of Screen Time

Using independent t-tests, the offered Python code contrasts screen time between deprived groups and those that are not. The pertinent data from "df2" and "df1" are first combined, and the combined data is then divided into "deprived" and "non-deprived" groups according to the "deprived" column. Then, for each category of screen time—"Total Computer We," "Total Video Games," "Total Smartphone," and "Total TV"—t-tests are run. With p-values of 0.000 for each test, the t-test results demonstrate substantial differences in screen time for each of the four categories between the two groups. This suggests a significant statistical difference in the amount of screen time that deprived people and those who are not.

```
# Merge datasets for analysis
merged_df = pd.merge(df2, df1[['ID', 'deprived']], on='ID')

# Compare screen time between deprived and non-deprived groups
deprived = merged_df[merged_df['deprived'] == 1]
non_deprived = merged_df[merged_df['deprived'] == 0]

# Import the necessary library
import scipy.stats as stats # Import the stats module from scipy

# Perform t-tests for each type of screen time
for column in ['Total_Computer_We', 'Total_Video_Games', 'Total_Smartphone', 'Total_TV']:
    t_stat, p_val = stats.ttest_ind(deprived[column], non_deprived[column], nan_policy='omit')
    print(f'T-test for {column}: t-statistic = {t_stat:.2f}, p-value = {p_val:.3f}')
```

```
T-test for Total_Computer_We: t-statistic = 13.19, p-value = 0.000
T-test for Total_Video_Games: t-statistic = 15.53, p-value = 0.000
T-test for Total_Smartphone: t-statistic = 29.87, p-value = 0.000
T-test for Total_TV: t-statistic = 33.55, p-value = 0.000
```

Correlations

The Pearson correlation coefficients between various forms of screen time and different mental health metrics are computed using the given Python code. After combining the pertinent data frames, iteratively calculate the association and p-value for each screen time and mental health column using the stats—Pearson function.

The code output is a series of statements indicating the correlation coefficient and p-value for each pair of screen time and mental health measures. For example, one statement might say, "Correlation between Total_Computer_We and Optm: correlation = -0.02, p-value = 0.000". This means a weak negative correlation between total computer use and optimism is statistically significant at a p-value of 0.000.

By examining these correlation coefficients, we may comprehend the connections between screen usage and several facets of mental health, like optimism, usefulness, relaxation, and more.

```python
# Calculate correlations
mental_health_columns = df3.columns[1:]  # Exclude 'ID'
screen_time_columns = ['Total_Computer_We', 'Total_Video_Games', 'Total_Smartphone', 'Total_TV']

# Merge datasets for correlation analysis
merged_df2 = pd.merge(df2, df1[['ID', 'deprived']], on='ID')
merged_df3 = pd.merge(merged_df2, df3[['ID'] + list(mental_health_columns)], on='ID')

# Compute correlations
for column in screen_time_columns:
    for mh_column in mental_health_columns:
        corr, p_val = stats.pearsonr(merged_df3[column], merged_df3[mh_column])
        print(f'Correlation between {column} and {mh_column}: correlation = {corr:.2f}, p-value = {p_val:.3f}')
```

```
Correlation between Total_Computer_We and Optm: correlation = -0.02, p-value = 0.000
Correlation between Total_Computer_We and Usef: correlation = -0.04, p-value = 0.000
Correlation between Total_Computer_We and Relx: correlation = -0.07, p-value = 0.000
Correlation between Total_Computer_We and Intp: correlation = 0.00, p-value = 0.712
Correlation between Total_Computer_We and Engs: correlation = -0.05, p-value = 0.000
Correlation between Total_Computer_We and Dealpr: correlation = -0.06, p-value = 0.000
Correlation between Total_Computer_We and Thcklr: correlation = -0.07, p-value = 0.000
Correlation between Total_Computer_We and Goodme: correlation = -0.08, p-value = 0.000
Correlation between Total_Computer_We and Clsep: correlation = -0.03, p-value = 0.000
Correlation between Total_Computer_We and Conf: correlation = -0.06, p-value = 0.000
Correlation between Total_Computer_We and Mkmind: correlation = -0.05, p-value = 0.000
Correlation between Total_Computer_We and Loved: correlation = -0.06, p-value = 0.000
Correlation between Total_Computer_We and Intthg: correlation = -0.01, p-value = 0.095
Correlation between Total_Computer_We and Cheer: correlation = -0.06, p-value = 0.000
Correlation between Total_Video_Games and Optm: correlation = -0.03, p-value = 0.000
Correlation between Total_Video_Games and Usef: correlation = 0.02, p-value = 0.000
Correlation between Total_Video_Games and Relx: correlation = 0.18, p-value = 0.000
Correlation between Total_Video_Games and Intp: correlation = -0.01, p-value = 0.089
Correlation between Total_Video_Games and Engs: correlation = 0.15, p-value = 0.000
Correlation between Total_Video_Games and Dealpr: correlation = 0.04, p-value = 0.000
Correlation between Total_Video_Games and Thcklr: correlation = 0.07, p-value = 0.000
Correlation between Total_Video_Games and Goodme: correlation = 0.14, p-value = 0.000
Correlation between Total_Video_Games and Clsep: correlation = -0.01, p-value = 0.002
Correlation between Total_Video_Games and Conf: correlation = 0.11, p-value = 0.000
Correlation between Total_Video_Games and Mkmind: correlation = 0.06, p-value = 0.000
Correlation between Total_Video_Games and Loved: correlation = -0.00, p-value = 0.999
Correlation between Total_Video_Games and Intthg: correlation = 0.02, p-value = 0.000
Correlation between Total_Video_Games and Cheer: correlation = 0.05, p-value = 0.000
Correlation between Total_Smartphone and Optm: correlation = -0.07, p-value = 0.000
Correlation between Total_Smartphone and Usef: correlation = -0.15, p-value = 0.000
Correlation between Total_Smartphone and Relx: correlation = -0.13, p-value = 0.000
```

Conclusion

- Gender Distribution is quite same.
- Minor and deprived people tends to spend more time on screen.
- People spend more time on smartphone and TV. However, we will calculate average time for weekdays and weekend.
- Average time directly impact on the wellbeing of the individual. The lower the value of screen time is, the more wellbeing value the individual have.
- Comparing the screen time and wellbeing, we can measure a person's situation
- By Applying regression model we can say the accuracy for the dataset

Thank you