CENTRE FOR MATHEMATICAL SCIENCE
UNIVERSITI MALAYSIA PAHANG AL-SULTAN ABDULLAH

BSD2333 DATA WRANGLING
SEMESTER II 2024/2025

GROUP PROJECT REPORT

MULTIDIMENSIONAL ANALYSIS OF WELLBEING, BEHAVIOR, AND PERFORMANCE

Group Photo :

LECTURERS' NAME : DR MOHD KHAIRUL BAZLI BIN MOHD AZIZ
GROUP NAME : NaN BUSTERS
SUBMISSION DATE : 8/6/2025

NO.	NAME	ID NUMBER	SECTION
1	SARA KHADIJA BINTI SAIDIN	SD23061	01G
2	ALIYA AFIFAH BINTI AL ABAS	SD23062	01G
3	FATIN ALISHA BINTI MOHAMED ZAINI	SD23015	01G
4	AIDA KAMILA FILZA BINTI ABDUL MANAF	SD23050	01G
5	WAN ZAFIRAH BINTI WAN YUSOF	SD23064	01G

TABLE OF CONTENT

1.0 Synopsis

1.1. Description of Assignment

The assignment titled 'Multidimensional Analysis of Wellbeing, Behavior, and Performance' seeks to explore how everyday habits, especially using social media, affect people's productivity. Understanding these relationships becomes extremely important. This study aims to find the relationship between everyday habits and productivity of people with different jobs and backgrounds.

One of the main goals of this assignment is to understand how social media can affect someone's productivity. Today, social media plays an important role in people's daily lives whether it is at home or at work. People use social media for different purposes. For example, someone in creative jobs may use social media platforms to increase knowledge and get inspirations from other artists while some people who are housewives might use social media to stay connected with their mutual friends. This study looks at how different jobs might lead to different outcomes when using social media. By collecting and analyzing data on how often people use social media and how productive they actually are, this project hopes to find patterns that show whether social media helps or harms productivity based on the kind of jobs they have.

Another main goal is to see how social media is linked to stress and people's sleep hours. Many people spend a lot of time on their phone or laptops based on their working needs which can affect their rest and mental health. This study will find out whether people who use social media more often feel more stressed or get less sleep at night. Not only that, it also looks at whether different habits, for example checking social media right before bed will affect the quality of sleep and makes it harder for people to relax. By learning these patterns, we can discover how online habits can influence people's health and well-being.

Moreover, it is important to find out how caffeine consumption is different for people based on their gender and job type. Since today's work life is so busy and exhausting, people tend to drink coffee daily which can affect their mental health and productivity. By studying these, we hope to find out if certain groups of people working get tired or burned out more easily than the others where they need to depend on coffee to stay focused throughout the day to get their work done.

Lastly, this assignment looks at which social media platforms are more popular or preferred by different genders to see if there are obvious patterns in how each genders uses them. This helps us understand better why social media is part of their daily routines. In conclusion, this assignment connects several important topics to show what affects people's productivity today.

1.2. Problem to be Solved

In today's world, social media plays a huge role in people's everyday lives. Many use platforms like Instagram, Tiktok, facebook, twitter not only to stay connected with friends and family but also in order to relax, be entertained, or stay informed. While social media offers many advantages, it also brings about a serious concern such as how does it affect our productivity? Many students and working adults spend long hours scrolling through content, sometimes losing track of time and delaying important tasks. This raises an important issue which is does social media usage directly reduce our ability to stay focused, organized and productive?

The "Social Media VS Productivity" dataset aims to help us explore this question. It contains self-reported survey responses from individuals about their time spent on social media, their daily routines, and their own views on how productive they feel. By analyzing this dataset, we hope to uncover patterns that show how different types of users based on age, gender, occupation, and time spent online are affected in terms of their productivity levels. This could help individuals manage their time wisely and allow educators, employers, and policymakers to offer guidance on more healthy digital habits.

However, before any meaningful insights can be drawn, there are several challenges in working with the dataset. Like many real-world datasets, this one contains missing values that can affect the accuracy of the analysis and may need to be handled through techniques like data imputation or removal of incomplete entries. Additionally, there are outliers, such as someone claiming to use social media for 24 hours a day or reporting unusually high or low productivity. These extreme values can skew the results and need to be carefully detected and treated.

Another challenge involves cleaning and organizing the data for visualisation. Some of the responses may be inconsistent or recorded in a way that makes it hard to categorize or compare. To exemplify, time spent on social media might be written in different formats, which must be standardized before analysis. Creating clear and useful visualizations also requires thoughtful grouping and labelling of data to make patterns easier to understand.

In summary, this dataset gives us the chance to explore a very relevant and relatable problem but also presents the common challenges that come with real-life data. Handling these issues carefully is necessary to make sure the final analysis is reliable and meaningful. With proper cleaning and preparation, the data can provide valuable insights that help people better balance their digital lives and personal productivity.

1.3. Research Questions

1. How does social media impact the actual productivity based on different job types?
2. Does social media have an influence towards sleep hours and stress level?
3. Do coffee consumption habits differ by gender and job type?
4. How often do people experience burnouts?
5. How do social media platform preferences vary by gender?

1.4. Objectives

1. To analyze the relationship between social media usage and productivity
2. To investigate the effect of social media use and stress on average sleep hours
3. To explore the frequency of coffee consumption across genders and job type
4. To identify the average number of burnout days across individuals
5. To determine the most and least preferred platforms for each gender

1.5. Basic description of the Data

The dataset used in this study was obtained from the Kaggle website and it is by Mahdi Mashayekhi in CSV format. This dataset is about daily digital habits of the users. It also includes social media usage, screen time, and notification exposure that will influence the individual's productivity level, stress and well-being. This dataset consists of 30,000 real-world-style records that portray the behavior pattern of people with different occupations, social customs, and lifestyle preferences make up the dataset.

The attributes of this dataset are including:

Bil	Attributes	Description
1	age	The individual's age (18-65 years old)
2	gender	Gender of individual (Male, Female, or Other)
3	job_type	The industry of employment (IT, Education, Student, etc.)
4	daily_social_media_time	Average of an individual's daily time spent on social media (hours)
5	social_platform_preference	The social media platform mostly used by an individual (Instagram, TikTok, Telegram, etc.)
6	number_of_notifications	Number of notifications an individual got per day
7	work_hours_per_day	Average of working hours each day by an individual
8	perceived_productivity_score	The productivity self rated score (scale: 0-10)
9	actual_productivity_score	Ground-truth productivity score simulation (scale: 0-10)
10	stress_level	The latest stress level (scale: 0-10)
11	sleep_hours	Average of sleeping hours per night
12	screen_time_before_sleep	Screening time spent on an individual before sleep (hours)
13	breaks_during_work	Number of breaks during working hours.
14	uses_focus_apps	The uses of digital focus application by an

		individual (boolean: True/ False)
15	has_digital_wellbeing_enabled	The activation of Digital Wellbeing (boolean: True/ False)
16	coffee_consumption_per_day	Number of cups of coffee consumed by an individual per day
17	days_feeling_burnout_per_month	Monthly number of burnouts per day reported
18	weekly_offline_hours	Total of hours offline spent in a week (excluding sleep)
19	job_satisfaction_score	The job/ life satisfaction of responsibilities (scale: 0-10)

This dataset also consists of few missing values in critical attributes that should undergo the cleaning process.

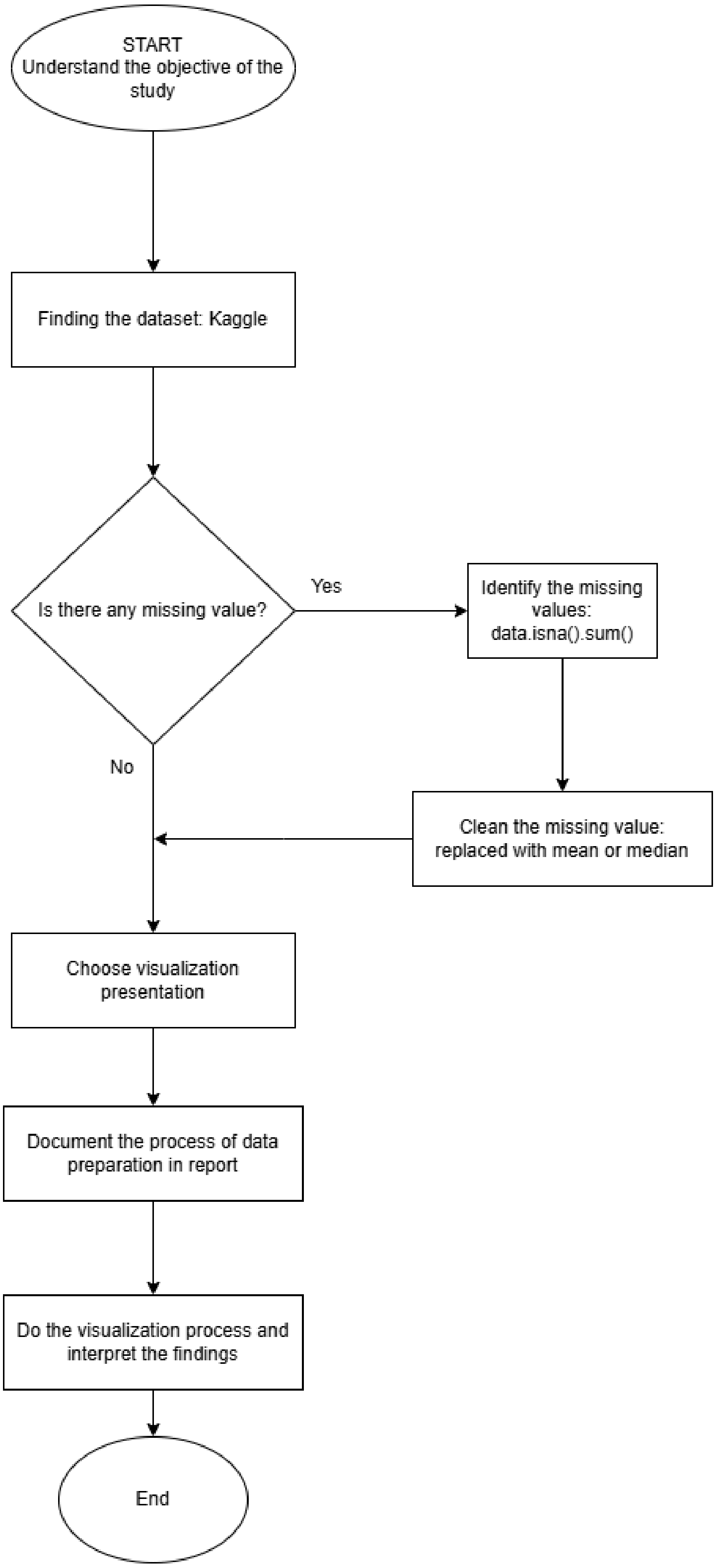
Bil	Attributes	Number of NaN
1	daily_social_media_time	2765
2	perceived_productivity_score	1614
3	actual_productivity_score	2365
4	stress_level	1904
5	sleep_hours	2598
6	screen_time_before_sleep	2211
7	job_satisfaction_score	2730

2.0 Packages Required

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.graph_objects as go
import plotly.express as px
import numpy as np
```

3.0 Data Preparation

3.1. Flowchart process of Data Preparation



3.2. Data Exploration

Import Data :

```
import pandas as pd
data = pd.read_csv('/content/dataprojectwrlg.csv')
data
```

	age	gender	job_type	daily_social_media_time	social_platform_preference	number_of_notifications	work_hours_per_day	perceived_
0	56	Male	Unemployed	4.180940	Facebook	61	6.753558	
1	46	Male	Health	3.249603	Twitter	59	9.169296	
2	32	Male	Finance	NaN	Twitter	57	7.910952	
3	60	Female	Unemployed	NaN	Facebook	59	6.355027	
4	25	Male	IT	NaN	Telegram	66	6.214096	
...	...	...	...	...	...	...	...	
29995	34	Female	Health	1.877297	Facebook	59	10.226358	
29996	39	Male	Health	4.437784	Instagram	46	4.692862	
29997	42	Male	Education	17.724981	TikTok	64	10.915036	
29998	20	Female	Education	3.796634	Instagram	56	6.937410	
29999	44	Male	Unemployed	NaN	Twitter	70	8.069883	

30000 rows × 19 columns

Check how many null :

```
print(data.isnull().sum())
```

```
age                                0
gender                             0
job_type                           0
daily_social_media_time         2765
social_platform_preference         0
number_of_notifications            0
work_hours_per_day                 0
perceived_productivity_score    1614
actual_productivity_score       2365
stress_level                    1904
sleep_hours                     2598
screen_time_before_sleep        2211
breaks_during_work                 0
uses_focus_apps                    0
has_digital_wellbeing_enabled      0
coffee_consumption_per_day         0
days_feeling_burnout_per_month     0
weekly_offline_hours               0
job_satisfaction_score          2730
dtype: int64
```

Print last 20 data :

```
data.tail(20)
```

	age	gender	job_type	daily_social_media_time	social_platform_preference	number_of_notifications	work_hours_per_day	perceive
29980	46	Other	IT	3.984904	TikTok	63	7.824013	
29981	26	Female	Education	3.171159	Twitter	68	5.984242	
29982	31	Female	Finance	NaN	Instagram	62	6.415302	
29983	53	Male	Student	4.642333	Facebook	68	4.863586	
29984	21	Male	Education	0.793902	Telegram	61	6.457148	
29985	55	Female	Education	3.489716	TikTok	52	10.009356	
29986	54	Female	Finance	3.834516	TikTok	64	8.630886	
29987	41	Female	Student	5.119140	TikTok	61	5.587729	
29988	34	Male	IT	3.562316	Twitter	59	7.685305	
29989	30	Female	Education	3.838301	Instagram	65	9.606261	
29990	37	Male	IT	8.379660	TikTok	46	7.601374	
29991	34	Male	Health	4.780144	Twitter	50	6.823827	
29992	44	Female	Unemployed	1.648825	TikTok	39	6.547262	
29993	50	Female	Education	3.070669	Instagram	56	5.870277	
29994	38	Male	Student	4.833425	Twitter	64	6.235001	
29995	34	Female	Health	1.877297	Facebook	59	10.226358	
29996	39	Male	Health	4.437784	Instagram	46	4.692862	
29997	42	Male	Education	17.724981	TikTok	64	10.915036	
29998	20	Female	Education	3.796634	Instagram	56	6.937410	
29999	44	Male	Unemployed	NaN	Twitter	70	8.069883	

3.3. Data Cleaning

```
data_clean = data.copy()

mean_columns = [
    'job_satisfaction_score',
    'perceived_productivity_score'
]

median_columns = [
    "daily_social_media_time",
    "actual_productivity_score",
    "sleep_hours",
    "screen_time_before_sleep",
    "weekly_offline_hours"
]

for col in mean_columns:
    col_mean = data_clean[col].mean()
    print(f"{col} mean: {col_mean:.4f}")
    data_clean[col] = data_clean[col].fillna(col_mean)

for col2 in median_columns:
    col_median = data_clean[col2].median()
    print(f"{col2} median: {col_median:.4f}")
    data_clean[col2] = data_clean[col2].fillna(col_median)

mod = data_clean['stress_level'].mode()[0]
print(f"stress_level mode: {mod}")
data_clean['stress_level'] =
data_clean['stress_level'].fillna(mod)

data_clean2 = data_clean[data_clean['gender'] != 'Other']

data_clean2.tail(20).round(4)
```

Output :

```
job_satisfaction_score mean: 4.9649
perceived_productivity_score mean: 5.5105
daily_social_media_time median: 3.0259
actual_productivity_score median: 4.9517
sleep_hours median: 6.4983
screen_time_before_sleep median: 1.0062
weekly_offline_hours median: 10.0137
stress_level mode: 3.0
```

	age	gender	job_type	daily_social_media_time	social_platform_preference	number_of_notifications	work_hours_per_day	perceived_
29979	51	Female	Unemployed	0.2399	Telegram	53	4.4734	
29981	26	Female	Education	3.1712	Twitter	68	5.9842	
29982	31	Female	Finance	3.0259	Instagram	62	6.4153	
29983	53	Male	Student	4.6423	Facebook	68	4.8636	
29984	21	Male	Education	0.7939	Telegram	61	6.4571	
29985	55	Female	Education	3.4897	TikTok	52	10.0094	
29986	54	Female	Finance	3.8345	TikTok	64	8.6309	
29987	41	Female	Student	5.1191	TikTok	61	5.5877	
29988	34	Male	IT	3.5623	Twitter	59	7.6853	
29989	30	Female	Education	3.8383	Instagram	65	9.6063	
29990	37	Male	IT	8.3797	TikTok	46	7.6014	
29991	34	Male	Health	4.7801	Twitter	50	6.8238	
29992	44	Female	Unemployed	1.6488	TikTok	39	6.5473	
29993	50	Female	Education	3.0707	Instagram	56	5.8703	
29994	38	Male	Student	4.8334	Twitter	64	6.2350	
29995	34	Female	Health	1.8773	Facebook	59	10.2264	
29996	39	Male	Health	4.4378	Instagram	46	4.6929	
29997	42	Male	Education	17.7250	TikTok	64	10.9150	
29998	20	Female	Education	3.7966	Instagram	56	6.9374	
29999	44	Male	Unemployed	3.0259	Twitter	70	8.0699	

4.0 Exploratory Data Analysis

4.1. Bar Chart (Actual Productivity Score by Job Type)

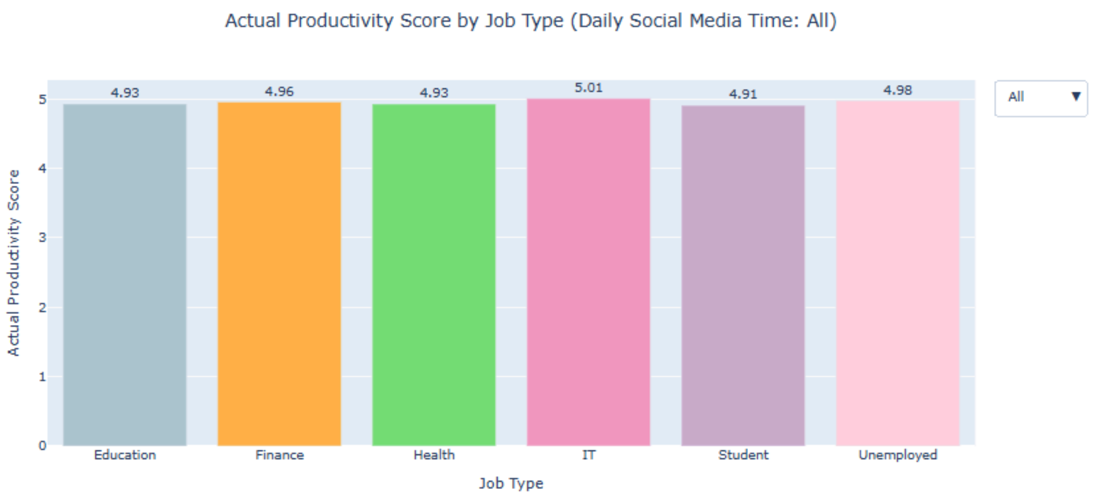
Codes :

```
import pandas as pd
import plotly.express as px

pastel_colors = [
    '#AEC6CF',  '#FFB347',  '#77DD77',
    '#F49AC2',  '#CBAACB', '#FFD1DC'
]

bins = [0, 1, 3, 5, 24]
labels = ['0-1 hrs', '1-3 hrs', '3-5 hrs', '>5 hrs']
data_clean2['smt_bin'] = pd.cut(
    data_clean2['daily_social_media_time'],
    bins=bins,
    labels=labels,
    include_lowest=True
)

smt_bins = data_clean2['smt_bin'].cat.categories.tolist()
smt_bins = [str(label) for label in smt_bins]
smt_bins.insert(0, 'All')

aggregated_data = {}
for smt in smt_bins:
    if smt == 'All':
        subset = data_clean2
    else:
        subset = data_clean2[data_clean2['smt_bin'] == smt]
    grouped =
subset.groupby('job_type')['actual_productivity_score'].mean().reset_i
ndex()
    grouped['actual_productivity_score'] =
grouped['actual_productivity_score'].round(2)
    grouped['smt_bin'] = smt
    aggregated_data[smt] = grouped

fig = px.bar(
    aggregated_data['All'],
    x='job_type',
    y='actual_productivity_score',
    labels={
        'job_type': 'Job Type',
        'actual_productivity_score': 'Actual Productivity Score'
    },
```

```
    title='Actual Productivity Score by Job Type filtered by Daily
Social Media Time',
    text=aggregated_data['All']['actual_productivity_score']
)

fig.update_traces(textposition='outside', marker_line_width=0)

dropdown_buttons = []
for smt in smt_bins:
    data = aggregated_data[smt]
    dropdown_buttons.append(
        dict(
            label=smt,
            method='update',
            args=[
                {
                    'x': [data['job_type']],
                    'y': [data['actual_productivity_score']],
                    'text': [data['actual_productivity_score']],
                    'type': 'bar',
                    'marker': {'color': pastel_colors[:len(data)]}
                },
                {
                    'title': f'Actual Productivity Score by Job Type
(Daily Social Media Time: {smt})'
                }
            ]
        )
    )

fig.update_layout(
    title=dict(x=0.5, xanchor='center'),
    updatemenus=[dict(
        buttons=dropdown_buttons,
        direction='down',
        showactive=True,
        x=1.02,
        xanchor='left',
        y=1,
        yanchor='top'
    )],
    yaxis=dict(range=[0,
aggregated_data['All']['actual_productivity_score'].max() + 1]),
    showlegend=False
)

fig.show()
```

Output :

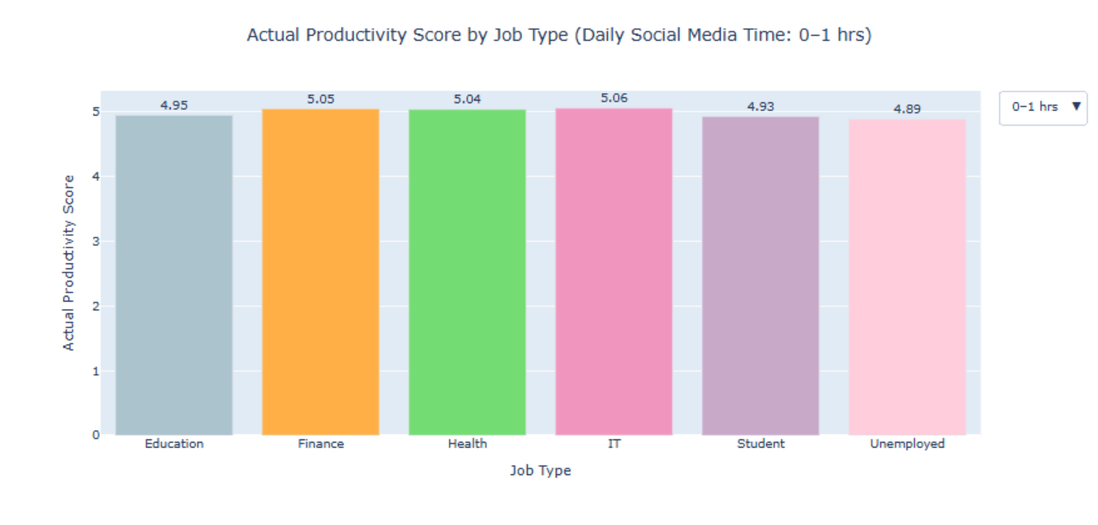
Figure 4.1.1: Actual Productivity Score by Job Type (Daily Social Media : All)



Interpretation:

The interactive bar chart above illustrates the relationship between actual productivity score and job types, categorised by daily social media time. The daily social media time is divided into 5 categories: All, 0-1 hours, 1-3 hours, 3-5 hours, and more than 5 hours. This analysis aims to determine which job types have the highest and lowest actual productivity, based on how much time individuals spend on social media daily. In this chart, we focus on the general time individuals spend on social media daily by setting the filter to "All". Under this general setting, individuals working in IT reported the highest actual productivity score, with a score of 5.01, while students had the lowest actual productivity score, with a score of 4.91. This chart demonstrates a noticeable difference in productivity across job types even under general settings. The result reflects that IT professionals tend to be more productive compared to students because IT professionals usually operate in structured, goal-oriented environments and have better time management, while students often rely on self-discipline.

Figure 4.1.2: Actual Productivity Score by Job Type (Daily Social Media : 0-1hour)

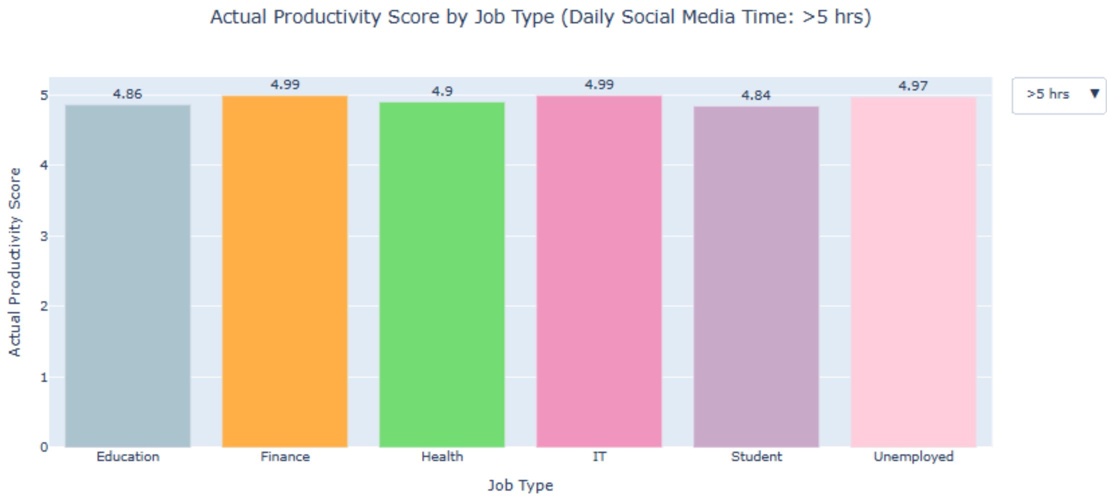


Interpretation:

In this bar chart, we focus on individuals who spend less than an hour daily on social media by setting the filter to "0-1 hours". This bar chart shows that the highest actual productivity scores within this job type are reported by individuals who work in IT with a productivity score (5.06), followed by individuals who work in Finance with a productivity score (5.05). This can be seen when an IT professional often spends long hours focused on coding and debugging software, as it is crucial for them to maintain concentration, as a single mistake or error can be difficult to identify and fix. For that reason, IT professionals tend to avoid checking their phone (social media) frequently, which contributes to their high productivity scores.

In contrast, the lowest productivity scores are reported by the unemployed with a 4.89 productivity score, followed by students with a 4.96 productivity score. Unemployed and students have the lowest productivity score because they often feel less responsible, which leads to lower productivity. For example, most of the students tend to spend more time on social media, even when they have numerous assignments to complete. This will lead to a lack of focus and lower productivity, as they lack time management skills and responsibility toward their assignments.

Figure 4.1.3: Actual Productivity Score by Job Type (Daily Social Media : > 5 hours)



Interpretation:

In this chart, we focus on individuals who spend more than 5 hours daily on social media. Shockingly, the highest actual productivity scores in this chart are reported by individuals who work in IT and Finance, both actual productivity scores are 4.99, while the lowest productivity score is 4.84 by students.

From this observation, we can see that even with high social media use, IT and Finance job types maintain strong productivity due to the job responsibilities, which require focus, accountability and time sensitivity. To illustrate, individuals who work in IT, such as developer professionals, must work with precision, as a small error in code can disrupt the whole system. Meanwhile, individuals who work in Finance must handle sensitive tasks that require focus, such as checking account balancing, since one mistake can result in serious financial consequences.

This happened because IT and Finance often work in structured, high-pressure and real-based environments where deadlines and their performance are being monitored. This encourages focus and avoids unnecessary distractions such as social media. In contrast, students typically operate in more relaxed and flexible settings, where the urgency to complete the task or assignments depends on their self-discipline and self-motivation.

From this analysis, we can see why students tend to have low productivity scores compared to other job types. Even when the daily time spent on social media varies, the results of the actual productivity score remain consistent. This is likely because students often lack external pressure and environmental discipline that push them to stay on task.

To conclude, IT and Finance maintain the highest actual productivity score even though they spend more than 5 hours on social media, because in some cases, these jobs use social media in their work. They might be using social media as a platform for communication, marketing and data monitoring. Thus, it is not just the amount of time individuals spend on social media that affects productivity score, but how and why social media is used in the context of their jobs.

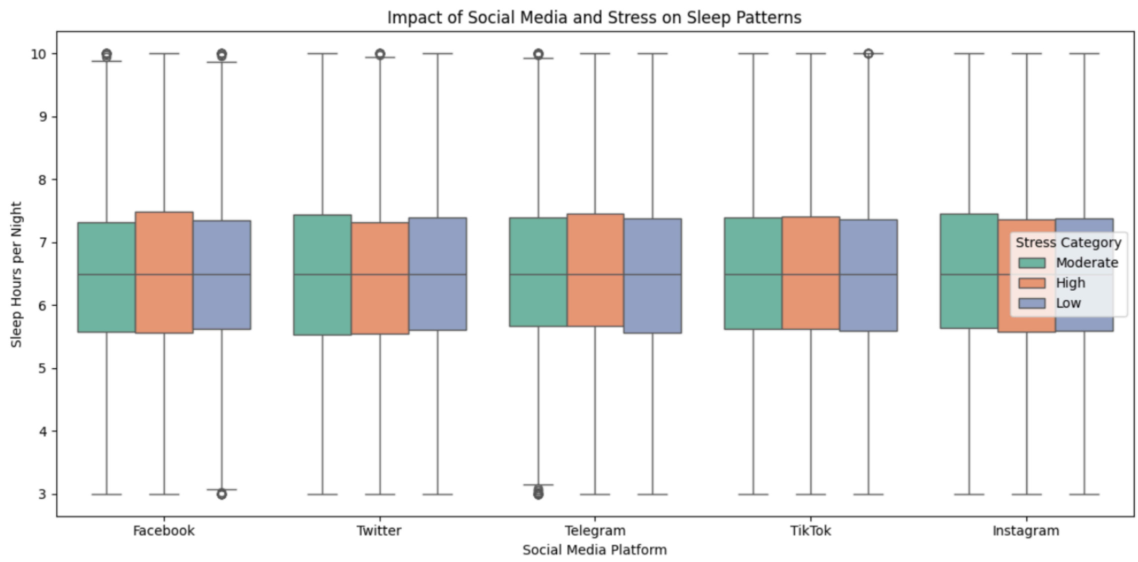
4.2. Box plot (Impact of Social Media and Stress on Sleep Patterns)

Code:

```
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd

def categorize_stress(level):
    if level <= 3:
        return 'Low'
    elif level <= 7:
        return 'Moderate'
    elif level <= 10:
        return 'High'
    else:
        return 'Extreme'

data_clean2['stress_category'] =
     data_clean2['stress_level'].apply(categorize_stress)


plt.figure(figsize=(12, 6))
sns.boxplot(
    data=data_clean2,
    x='social_platform_preference',
    y='sleep_hours',
    hue='stress_category',
    palette='Set2'
)
plt.title('Impact of Social Media and Stress on Sleep Patterns')
plt.ylabel('Sleep Hours per Night')
plt.xlabel('Social Media Platform')
plt.legend(title='Stress Category')
plt.tight_layout()
plt.show()
```

Output:

Figure 4.2.1: Impact of Social Media and Stress on Sleep Patterns



Interpretation:

The boxplot above visualizes the relationship between sleep hours and the preferred social media platform while categorising it by stress levels. The stress levels are categorized by three levels which are low, moderate and high where low level is from 1 to 3, moderate level is 4 to 7 and high level is from 8 to 10. The different colours indicate the different levels.

Based on the visualizations, it can be seen that the median for sleep hours across all social media platforms which includes Facebook, Twitter, Telegram, Tiktok and Instagram, have a consistent value of 6.5 hours of sleep. This means that on average, individuals get the same amount of sleep regardless of the social media platform they use.

Firstly, the boxplot for Facebook users shows a consistent sleep pattern across all stress categories with the interquartile range of 5.5 to more than 7 hours and a median around 6.5. The whiskers spread out evenly and show an approximately normal distribution. There also seem to be outliers in the low and moderate stress category in the upper range, indicating that the individuals who have a low-moderate stress level are getting more sleep than average. Furthermore, there are outliers in the lower range in the low category showing that even with a low amount of sleep, the individuals involved are less stressed. This interpretation suggests that this social media has users who have consistent sleep durations regardless of stress levels.

Next, Twitter users also show a normal distribution in sleep hours across stress levels. The IQR is also from 5.5 to more than 7 hours with the median located centrally in each boxplot of 6.5 hours. There are evenly spread whiskers for each boxplot. There are also outliers located in the high stress category, indicating there are some individuals who have high stress levels using Twitter even though they have a 10 hour sleep. This shows that there is a possible high stress level outcome even if one has more than enough sleep.

The telegram boxplot shows a right-skewed distribution in the moderate and high stress category and a normal distribution for the low category. The median is slightly closer to the lower range of the boxplot and for the low category it remains at the centre. The upper whisker in the moderate category seems to be longer than the other boxplots, indicating that some users have more sleep than average and the boxplot also shows that there are outliers in both the lower and upper range. This visualization shows that there are inconsistent sleep behaviours among Telegram users.

The distribution of sleep hours among TikTok users also appears to be slightly right-skewed especially in the moderate and high stress categories. The median is also lower within the IQR range of 5.5 to more than 7 hours of sleep hours. The upper whiskers seem to be longer than the other boxplots in the other social media platforms. There are also outliers at the high end of the low stress category. This suggests that while most users have shorter sleep hours, there are some to have more sleep than average, which pulls the distribution towards the higher end. Thus, the sleep pattern among TikTok users is less consistent and more varied compared to other platforms.

Lastly, Instagram users present a slightly right-skewed distribution in the moderate and high stress level in an IQR range the same as other platforms with a median that is slightly lower within the boxes. The upper whiskers also extend more than the lower ones. There seem to be no outliers present in the Instagram boxplots and the sleep pattern seems to show a stable sleep pattern.

In conclusion, while the median sleep duration across all social media platforms is consistently around 6.5 hours, the distribution patterns vary. Facebook and Twitter show a normal distributed and stable sleep pattern, while the others have a right-skewed distribution especially in the moderate to high stress categories. This suggests that although average sleep hours remain similar, sleep consistency may differ depending on the platform and stress level.

4.3. Grouped Bar Chart (Coffee Habits at Work: A Gender and Job Type Perspective)

Codes:

```
import plotly.express as px

grouped_data = data_clean2.groupby(['job_type', 'gender',
'coffee_consumption_per_day']).size().reset_index(name='coun
t')

grouped_data = grouped_data.sort_values(by='job_type')

first_job_type = grouped_data['job_type'].iloc[0]

fig = px.bar(
    grouped_data,
    x="coffee_consumption_per_day",
    y="count",
    color="gender",
    color_discrete_map={
        "Female": "#660000",
        "Male": "#0066cc"
    },
    barmode="group",
    animation_frame="job_type",
    text="count",
    labels={
        "coffee_consumption_per_day": "Cups of Coffee per
Day",
        "count": "Number of People",
        "gender": "Gender",
        "job_type": "Job Type"
    }
)

fig.update_layout(
    title={
        'text': f'Coffee Consumption by Gender for Job Type:
{first_job_type}',
        'x': 0.5
    },
    xaxis_title="Cups of Coffee per Day",
    yaxis_title="Number of People",
```

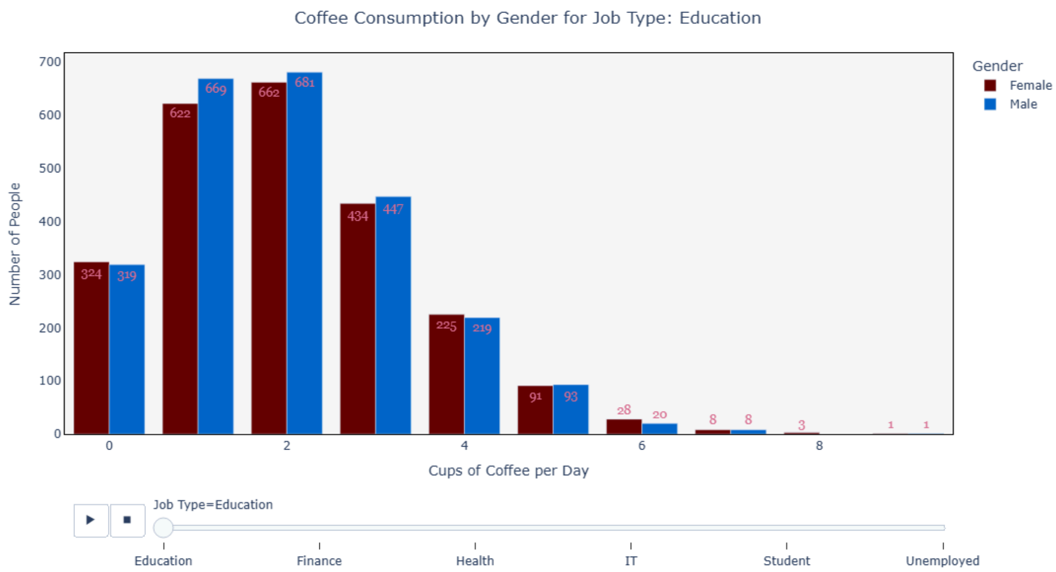
```
    legend_title="Gender",
    width=1100,
    height=600,
    paper_bgcolor='white',
    plot_bgcolor='whitesmoke',
    xaxis=dict(
        showline=True,
        linewidth=1.1,
        linecolor='black',
        mirror=True,
        showgrid=False
    ),
    yaxis=dict(
        showline=True,
        linewidth=1.1,
        linecolor='black',
        mirror=True,
        showgrid=False
    )
)

fig.update_traces(
    textposition='auto',
    textfont=dict(color='palevioletred', family='Georgia',
size=12.5)
)

fig.layout.updatemenus[0].buttons[0].args[1]['frame']['durat
ion'] = 2400
fig.layout.updatemenus[0].buttons[0].args[1]['transition']['
duration'] = 300

for frame in fig.frames:
    job = frame.name
    frame.layout = {
        'title': {
            'text': f'Coffee Consumption by Gender for Job
Type: {job}',
            'x': 0.5
        }
    }

fig.show()
```

Output and interpretations:

The visualization below shows coffee consumption by people per day separated by gender and filtered by 6 different types of jobs presented in a grouped bar chart. The color used indicates different gender, where red is for females while blue stands for male. Other than that, the filter in an animation mode is used for filtering the job type. The labels inside the plotting area indicate the number of people who take coffee per day for the selected type of employment. The title of the chart is in an interactive mode, where the type of job will change based on users preferred.

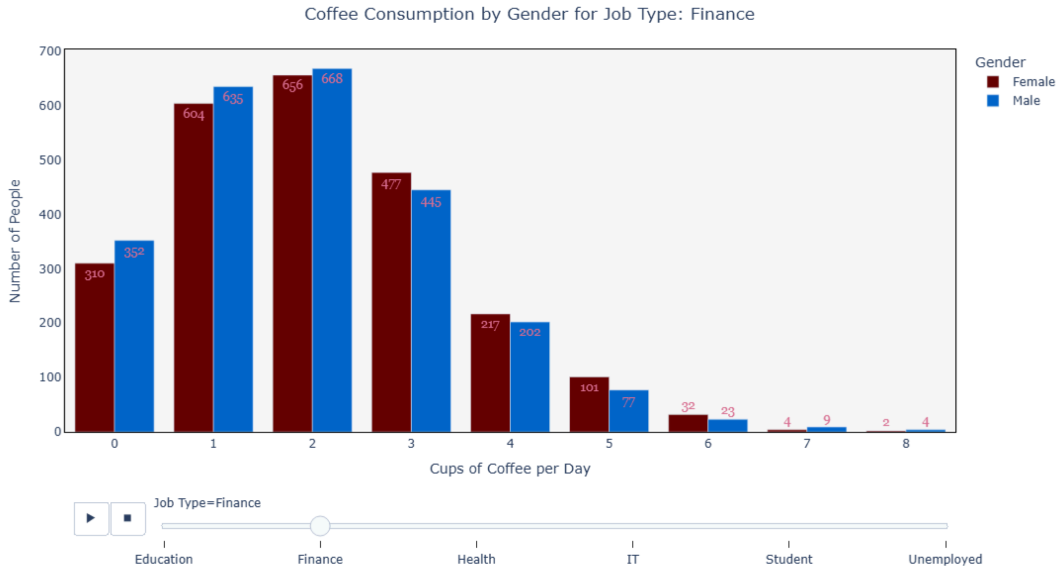
Generally, all types of jobs show two cups of coffee has the highest consumer for both gender, male and female. Male lead the number of cups of coffee for all types of jobs except for “Student”. “IT” employers are the only employment that consume the highest cups of coffee as much as 10 cups compared to other types of jobs. Most people preferred around 1 to 3 cups of coffee per day where both genders, male and female show a similar pattern of coffee intake distributions. Additionally, there are fewer people consuming 5 or more cups in a day resulting in the most common cup intake across all types of jobs being 2 cups per day. However, all types of jobs record zero cup intake indicating that people still can focus on work even without caffeine intake.

Figure 4.3.1 Coffee Consumption by Gender for Job Type: Education



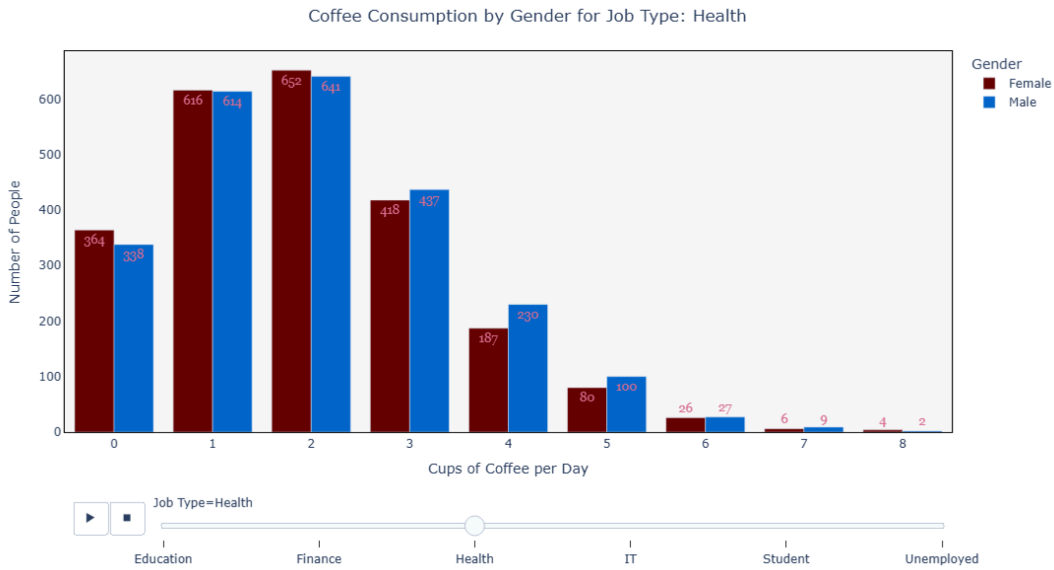
In the education sector, the highest coffee consumption is 2 cups per day for both genders, male as much as 681 while female displayed 662 individuals. Next, males slightly outnumber females in highest consumption categories between 3 to 5 cups per day. Overall, the highest cups consumed was 9 cups, one person from both genders as a representative. However, the number of non-coffee drinkers which is 0 cups consumed is slightly higher among females as 324 compared to male, 319.

Figure 4.3.2 Coffee Consumption by Gender for Job Type: Finance



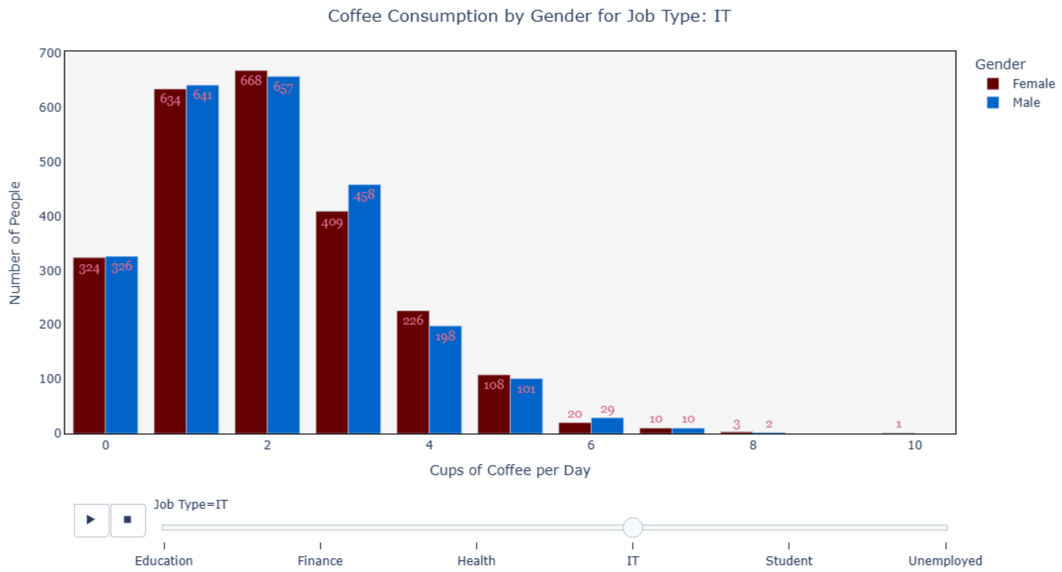
In the finance industry, 2 cups per day still be the peak for both genders where male leading as 668 compared to female just 656. Coffee consumption for females is generally slightly higher at lower levels between 0 to 3 cups but male consume more at higher levels, above 5 cups per day. There are very few individuals who drink 7 and above cups per day, but males appear slightly more likely to do so.

Figure 4.3.3 Coffee Consumption by Gender for Job Type: Health



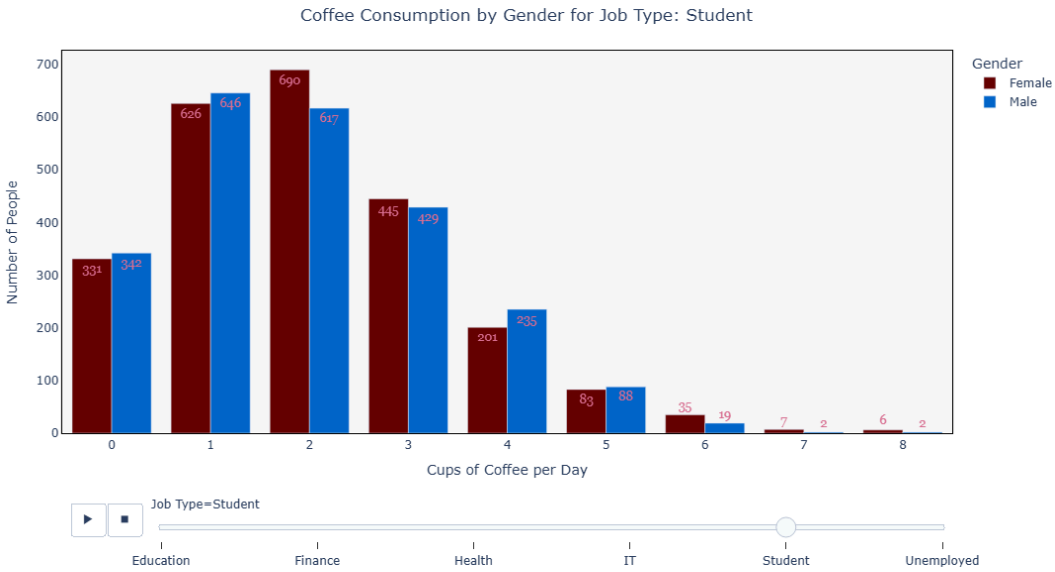
The highest coffee consumption was led by females, likely mirroring other professions, with 2 cups per day. Almost the same, total number of people who took only 1 cup of coffee per day between male and females, with females slightly higher than male. As the total cups increase from 3 cups per day, male take a place as the leader of this category. To conclude, health professions prioritize the moderate intake of coffee due to awareness of caffeine's effects, but data granularity is needed.

Figure 4.3.4 Coffee Consumption by Gender for Job Type: IT



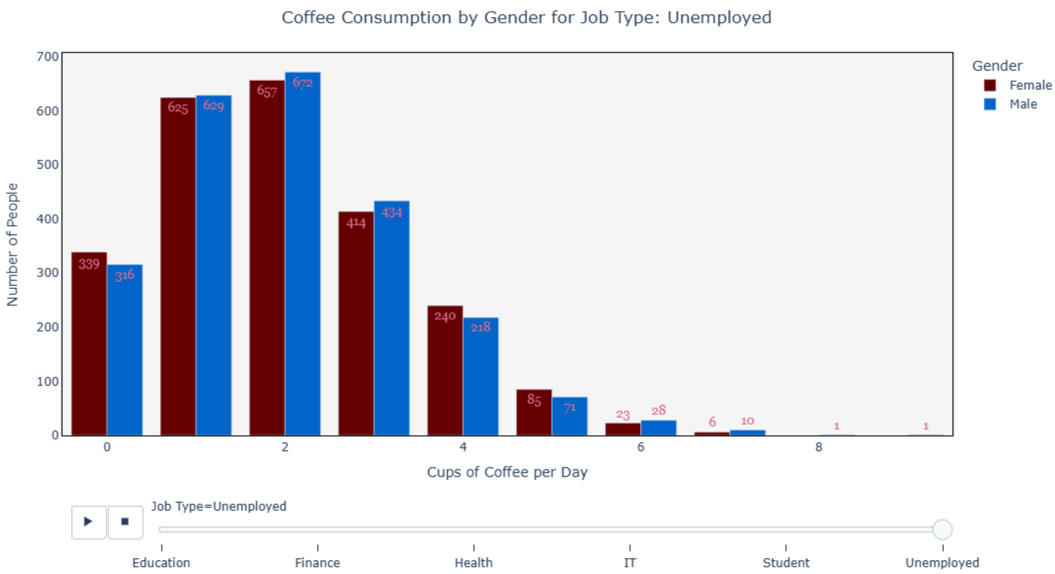
The peak consumption is 2 cups per day, dominated by female, 668 people while male slightly dropped from female, 657 individuals. The trend aligns very well with the general trends. For 1 to 3 cups per day, there are nearly equivalent numbers of people between both genders, male and female. Males slightly lead the distribution in higher intake, 6 cups and above. There might be an outlier where 1 person took 10 cups per day who was a female. To conclude, IT professionals show uniform habits across genders, with a minor male skew in heavy consumption.

Figure 4.3.5 Coffee Consumption by Gender for Job Type: Student



The highest coffee consumption is females at 2 cups level as much as 690 compared to male, 617 at the same levels. Females again dominate in 0 to 3 cups per day while male slightly exceed females in 4 to 5 cups per day. Additionally, a very small proportion for both male and females drink 6 cups and more per day. To conclude, female students tend to fall more for propaganda coffee consumption in higher intake compared to male.

Figure 4.3.6 Coffee Consumption by Gender for Job Type: Unemployed



The highest coffee consumption is male, 672 followed by female, 657 at the same levels of cups consumed, which are 2 cups per day. The number of individuals who are non-coffee drinkers highest in females is 339 people. At the levels of 4 to 5 cups per day, females lead and it drops sharply at 6 cups. Above 6 cups consumption per day is led by male. To conclude, unemployment status does not significantly alter coffee habits where the distribution absolutely mirrors employed groups but with slightly lower totals.

In a nutshell, all visualization follows the same trend where 2 cups per day is the standout preference across all job types and genders. Males, generally gradually outnumber females in higher consumption, especially in professional roles such as IT and Finance. However, females dominate the 0 to 3 cups per day range in most non-professional categories such as Student. Regardless of job or gender, 1 to 3 cups of coffee per day is the norm, peaking at 2 cups. Caffeine consumption is normally being normalised in a workplace due to increased productivity and individual's focus. Due to this, most people who work in professional roles might consume coffee at a higher level compared to non-professional groups.

4.4. Histogram (Days Feeling Burnout in a Month)

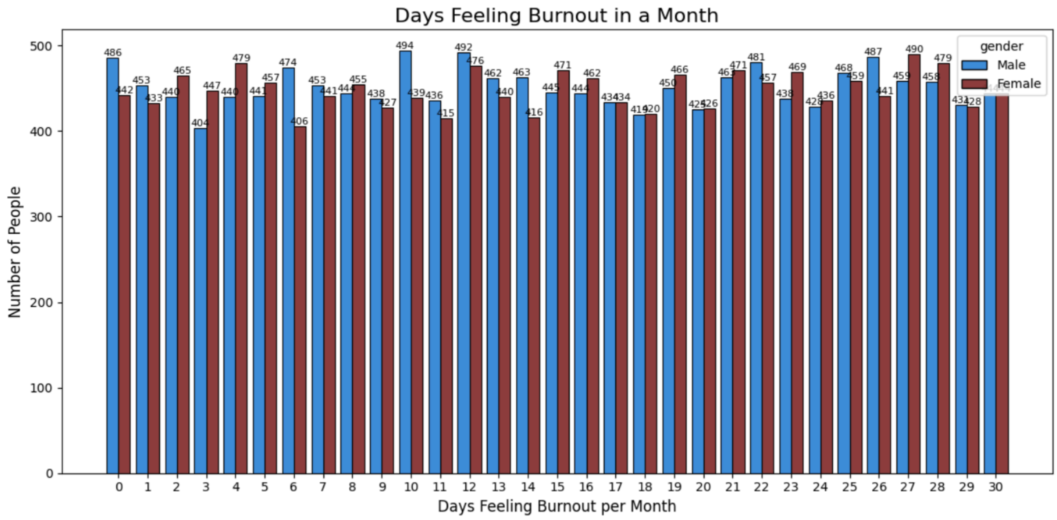
Code :

```
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np

filtered_df = df[df['gender'].isin(['Male', 'Female'])]

bins = np.arange(-0.5, 31.5, 1)

custom_palette = {'Female': '#660000', 'Male': '#0066cc'}

plt.figure(figsize=(12, 6))
ax = sns.histplot(
    data=filtered_df,
    x='days_feeling_burnout_per_month',
    hue='gender',
    bins=bins,
    multiple='dodge',
    shrink=0.8,
    element='bars',
    palette=custom_palette,
    edgecolor='black'
)

for p in ax.patches:
    height = p.get_height()
    if height > 0:
        ax.text(
            p.get_x() + p.get_width() / 2,
            height,
            int(height),
            ha='center',
            va='bottom',
            fontsize=8
        )

plt.title('Days Feeling Burnout in a Month', fontsize=16)
plt.xlabel('Days Feeling Burnout per Month', fontsize=12)
plt.ylabel('Number of People', fontsize=12)
plt.xticks(np.arange(0, 31, 1))
plt.tight_layout()
```

```
plt.show()
```

Results :

Figure 4.4.1 Bar Chart of Days Feeling Burnout in a Month



Interpretation :

This histogram shows the distribution of burnout frequency among people with different genders within a 30 days month. The histogram breaks down the number of males and females for a specific number of days they feel burnout ranging from 0 to 30 days. The blue bars represent the male gender while the brown bars represent female. This chart allows us to explore how burnout is experienced between genders and identify if there are any trends or differences.

Based on the visualization, at 0 days of feeling burnout, there are 486 males reported not feeling burnout at all while females are only 442 people. This shows that females are the ones who frequently experience burnout rather than male. A similar pattern shows at 3 days of feeling burnout where 447 females reported to be burnout while only 404 males feel burnout. The highest number of days females feel burnout is 27 days with the amount of 490. It shows that females are likely to report higher burnout frequencies. This could point toward the fact where females are either being more sensitive while handling a task. This emotional engagement can make them easily feel stressed and burnout.

Next, the highest days of male feeling burnout is only 10 days with the amount of male reported to feel burnout are 494 people while females are 439 people.

It has the difference of 55 people. This shows that male know how to cope with their emotions better than female rather than being exhausted and feeling burnout.

Lastly, the visualization shows on the highest days of feeling burnout which is 30 days, the distribution is equal, meaning that male and female reported feeling burnout for 30 days has the same amount. It is clear that feeling burnout is a normal experience across genders, but the distribution on how often they feel is different. Males tend to feel burnout for the average of 10 days. Females, on the other hand, appear to feel burnout for the average of 27 days. This could reflect the differences on how they cope with their emotions.

In conclusion, the histogram shows that feeling burnout is a common problem for both males and females but the amount of days they feel is different based on how they handle their emotion and how much their workload is. Males mostly feel burnout for only short days for example only 10 days while females tend to feel burnout more, like 27 days in a month. These differences show that it is important to have mental health support that considers the unique needs of both genders. By understanding how males and females function differently while experiencing burnout at the workplace, it is important to create better ways to help each group to prevent burnouts.

4.5. Pie Chart (Social Media Preferences by Gender)

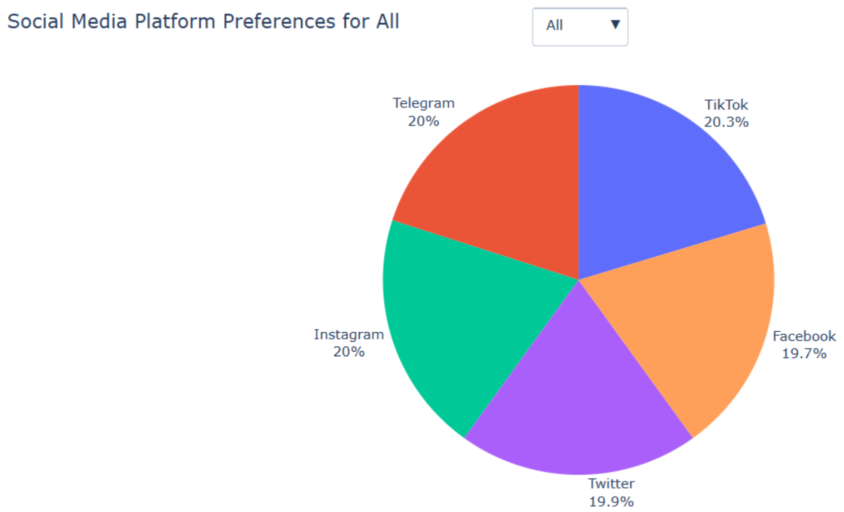
Codes :

```
import pandas as pd
import plotly.graph_objects as go

genders = ['All'] + sorted(df['gender'].unique().tolist())

pie_traces = []
for gender in genders:
    if gender == 'All':
        temp_df = df
    else:
        temp_df = df[df['gender'] == gender]

    counts = temp_df['social_platform_preference'].value_counts()

    pie_traces.append(go.Pie(
        labels=counts.index,
        values=counts.values,
        name=gender,
        textinfo='label+percent',
        textposition='outside',
        insidetextorientation='radial',
        hole=0,
        visible=(gender == 'All')
    ))

dropdown_buttons = []
for i, gender in enumerate(genders):
    visibility = [False] * len(genders)
    visibility[i] = True
    dropdown_buttons.append(dict(
        label=gender,
        method='update',
        args=[{'visible': visibility},
              {'title': f'Social Media Platform Preferences for
{gender}'}]
    ))

fig = go.Figure(data=pie_traces)
fig.update_layout(
```

```
    title=f'Social Media Platform Preferences for All',
    updatemenus=[dict(
        active=0,
        buttons=dropdown_buttons,
        x=0.5,
        y=1.2,
        xanchor='center',
        yanchor='top'
    )],
    showlegend=False
)


fig.show()
```

Result :

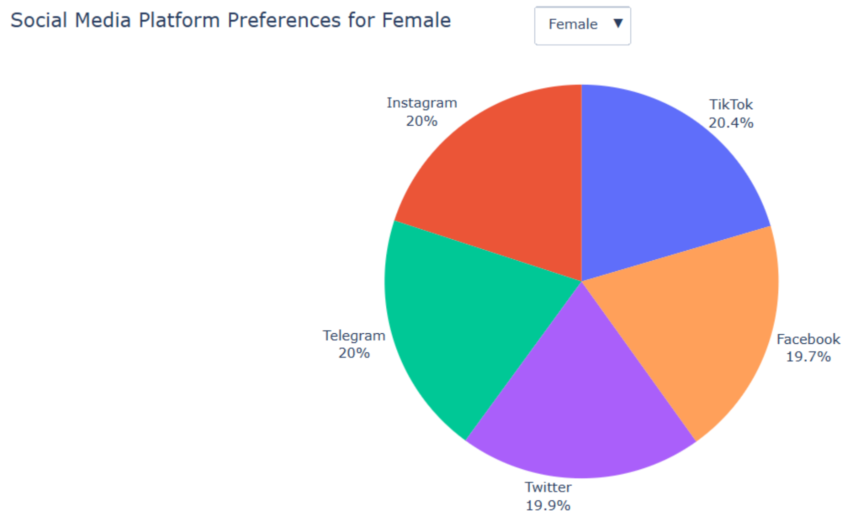
Figure 4.5.1 : Social Media Platform Preferences for All



Interpretation :

This interactive pie chart illustrates the distribution of social media platform preferences for both genders which are female and male. It is shown that TikTok records the highest percentage with 20.3% compared to other social media platforms. This suggests that a remarkable portion of individuals gravitate toward TikTok than other platforms, likely due to its highly engaging short videos, viral trends, and personalized algorithms. Whereas, the social media platform with the lowest percentage is Facebook with 19.7% indicating a decline in its relevance since Facebook is one of the oldest mainstream social media platforms. Thus, the chart highlights a clear shift in social media preferences, with users favoring newer, more interactive platforms like TikTok over older ones like facebook. This reflects changing user behaviour and the growing demand for engaging, fast-paced, and personalized digital content.

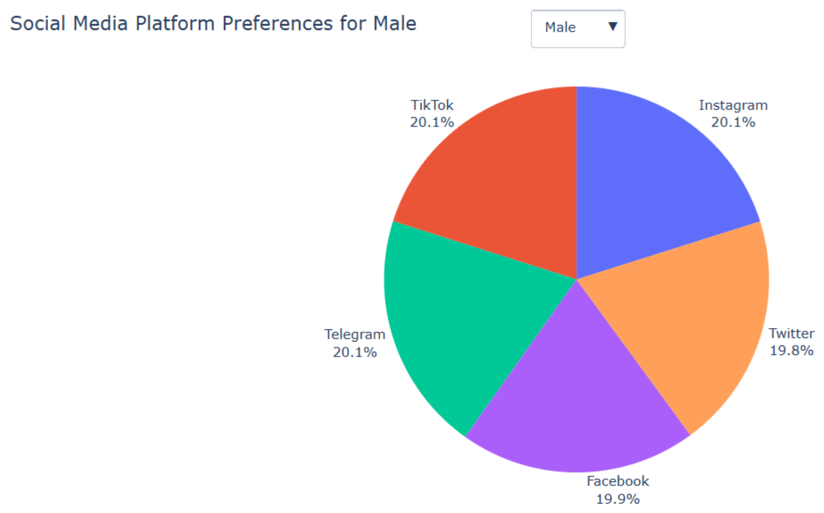
Figure 4.5.2 : Social Media Platform Preferences for Female



Interpretation :

Based on the interactive pie chart, Tiktok is the most preferred social media platform, capturing the largest share of user preferences. This highlights TikTok's widespread popularity, likely driven by its addictive short-form content, viral trends, and algorithmic content recommendations that keep users engaged. To add, it is particularly popular among females as it offers an engaging platform for self-expression, creativity, and community, with content that resonates deeply through relatable themes such as beauty, fashion, mental health, and lifestyle, while also empowering users to voice out their opinions, build personal brands, and connect with others in a meaningful ways. There is also an interesting pattern where Instagram and Telegram have the same percentage with 20%. The reason why the percentage of female users might be the same could be because both platforms appeal similar to interests among women. Many female users enjoy using apps that allow them to connect with friends, share their routines, and communicate easily. Instagram offers visually engaging features to share memories such as photo sharing stories, and reels, while Telegram provides strong messaging capabilities , large group chats, and privacy options. For example, there are many personal shoppers that use Telegram as their business platform to engage with the customers which mostly are females. These different but complementary features attract female users equally. To conclude, the data reflects how different platforms cater to users' varying needs, with TikTok leading due to its engaging and relatable content, especially among females. Meanwhile, the equal preference for Instagram and Telegram highlights how both platforms successfully meet similar social, communication and aesthetic needs, making them equally appealing to female users.

Figure 4.5.3 : Social Media Platform Preferences for Male



Interpretation :
The visualization above shows that TikTok, Instagram and Telegram share the same percentage which is 20.1%. The equal preferences among males for these platforms likely reflects the diverse ways they engage with digital platforms with TikTok offering entertainment and trends, Instagram providing usual updates and social connections, and Telegram serving as tool for private communication accessing niche content suggesting that no single platform fully meets all their interests, leading to balanced usage across the three. Furthermore, it also shows that Twitter has a slightly lower percentage than Facebook even though Facebook can be considered as one of the oldest social media platforms. One possible reason for this is that Facebook still offers a familiar space for connecting with friends and family through posts, photos, and comments. It is often used for maintaining personal relationships and staying updated with social circles. In contrast, Twitter is more centered around public conversations, news, and opinions. This style of content may not appeal as much to users who prefer more personal and private experience. Not only that, Twitter requires a more active engagement to feel connected, while Facebook allows users to interact more passively. These differences in platform function and user experience may explain why Facebook maintains a higher preference than Twitter among males in the dataset. Hence, the equal

preference for TikTok, Instagram and Telegram among males indicates that each platform fulfills different aspects of their digital needs, indicating in balanced usage. The slightly higher preference for Facebook over Twitter can be attributed to Facebook's more personal and familiar environment, which contrasts with Twitter’s fast-paced and public nature. These patterns highlight how platform features and user engagement styles influence social media preferences among male users.

5.0 Summary

5.1 Summary for each of the integer attributes

Codes :

```
import pandas as pd

int_cols = data_clean2.select_dtypes(include='int').columns
data_int = data_clean2[int_cols]

summary = data_int.describe().round(2)

print("Summary statistics for integer columns :\n", summary)

variability = summary.loc['std'].sort_values(ascending=False).round(2)
print("\nVariables sorted by variability (standard deviation):\n",
variability)
```

Output :

```
Summary statistics for integer columns :
             age  number_of_notifications  breaks_during_work  \
count  28822.00                 28822.00            28822.00
mean      41.48                    59.95                4.99
std       13.83                     7.73                3.17
min       18.00                    30.00                0.00
25%       30.00                    55.00                2.00
50%       41.00                    60.00                5.00
75%       53.00                    65.00                8.00
max       65.00                    90.00               10.00

       coffee_consumption_per_day  days_feeling_burnout_per_month
count                    28822.00                        28822.00
mean                         2.00                           15.56
std                          1.41                            9.25
min                          0.00                            0.00
25%                          1.00                            8.00
50%                          2.00                           16.00
75%                          3.00                           24.00
max                         10.00                           31.00

Variables sorted by variability (standard deviation):
 age                               13.83
days_feeling_burnout_per_month     9.25
number_of_notifications            7.73
breaks_during_work                 3.17
coffee_consumption_per_day         1.41
Name: std, dtype: float64
```

Interpretation :

This analysis only focuses on the integer-type attributes present in the dataset. It is shown that the age attribute has the highest variability with a standard deviation of 13.83, indicating that the ages of respondents are widely spread across a large range. This makes this data more unpredictable in analyses of this attribute because the age data is less consistent. In contrast, the coffee_consumption_per_day attribute has the lowest variability with a standard deviation of 1.41. From this observation, we can say that most of the respondents consume a similar amount of coffee daily, making this attribute more consistent compared to others and easier to make accurate predictions because this attribute is predictable.

5.2 Overall Summary for all attributes except for categorical attributes

Coding :

```
import pandas as pd

exclude_cols = ['gender', 'job_type', 'social_platform_preference',
                'uses_focus_apps', 'has_digital_wellbeing_enabled']
data_selected = data_clean2.drop(columns=exclude_cols)

summary_all = data_selected.describe(include='all').round(2)
print("\nOverall summary statistics for selected DataFrame:\n")
print(summary_all)

variability_all_sorted =
summary_all.loc['std'].sort_values(ascending=False).round(2)
print("\nAll numeric variables sorted by variability (standard
deviation):\n")
print(variability_all_sorted)
```

Output :

```
Overall summary statistics for selected DataFrame:

            age  daily_social_media_time  number_of_notifications  \
count  28822.00                 28822.00                 28822.00
mean      41.48                     3.11                    59.95
std       13.83                     1.97                     7.73
min       18.00                     0.00                    30.00
25%       30.00                     1.80                    55.00
50%       41.00                     3.03                    60.00
75%       53.00                     4.22                    65.00
max       65.00                    17.97                    90.00

       work_hours_per_day  perceived_productivity_score  \
count            28822.00                      28822.00
mean                 6.99                          5.52
std                  2.00                          1.97
min                  0.00                          2.00
25%                  5.64                          3.86
50%                  6.99                          5.51
75%                  8.36                          7.18
max                 12.00                          9.00

       actual_productivity_score  stress_level  sleep_hours  \
count                   28822.00      28822.00      28822.0
mean                        4.96          5.36          6.5
std                         1.81          2.84          1.4
min                         0.30          1.00          3.0
25%                         3.51          3.00          5.6
50%                         4.95          5.00          6.5
75%                         6.40          8.00          7.4
max                         9.85         10.00         10.0
```

```
       actual_productivity_score  stress_level  sleep_hours  \
count                   28822.00      28822.00      28822.0
mean                        4.96          5.36          6.5
std                         1.81          2.84          1.4
min                         0.30          1.00          3.0
25%                         3.51          3.00          5.6
50%                         4.95          5.00          6.5
75%                         6.40          8.00          7.4
max                         9.85         10.00         10.0

       screen_time_before_sleep  breaks_during_work  \
count                  28822.00            28822.00
mean                       1.02                4.99
std                        0.63                3.17
min                        0.00                0.00
25%                        0.57                2.00
50%                        1.01                5.00
75%                        1.44                8.00
max                        3.00               10.00

       coffee_consumption_per_day  days_feeling_burnout_per_month  \
count                    28822.00                        28822.00
mean                         2.00                           15.56
std                          1.41                            9.25
min                          0.00                            0.00
25%                          1.00                            8.00
50%                          2.00                           16.00
75%                          3.00                           24.00
max                         10.00                           31.00

       weekly_offline_hours  job_satisfaction_score
count              28822.00                28822.00
mean                  10.35                    4.97
std                    7.28                    2.02
min                    0.00                    0.00
25%                    4.52                    3.53
50%                   10.00                    4.96
75%                   15.28                    6.41
max                   40.96                   10.00


All numeric variables sorted by variability (standard deviation):

age                               13.83
days_feeling_burnout_per_month     9.25
number_of_notifications            7.73
weekly_offline_hours               7.28
breaks_during_work                 3.17
stress_level                       2.84
job_satisfaction_score             2.02
work_hours_per_day                 2.00
daily_social_media_time            1.97
perceived_productivity_score       1.97
actual_productivity_score          1.81
coffee_consumption_per_day         1.41
sleep_hours                        1.40
screen_time_before_sleep           0.63
Name: std, dtype: float64
```

Interpretation :

In this analysis, we observe the summary of overall attributes except for categorical attributes. This observation demonstrates that the age attribute has the highest variability, with a standard deviation of 13.3. This suggests that age is less reliable and more unpredictable. Meanwhile, the attribute with the least fluctuation is screen_time_before_sleep, with a standard deviation of 0.63. This shows that most of the respondents had a similar amount of screen time before bed, making this attribute more reliable and consistent.

6.0 References

Celebi, S. I., & Terkan, R. (2020). Social Media and Employee Productivity at Workplace. *International Review of Management and Marketing*, *10*(6), 37-41. https://doi.org/10.32479/irmm.10806

How Does Social Media Affect Productivity? (2023, March 23). Flash Hub. Retrieved June 8, 2025, from https://flashhub.io/how-does-social-media-affect-productivity/

Maria, S., & Corina, D. (2023, 7). Study on the Behavior and Preferences of Social Media Users, after the Pandemic Era. *Proceedings of the International Conference on Business Excellence*, *17*(1), 1873-1887. https://doi.org/10.2478/picbe-2023-0165

GeeksforGeeks. (2025, April 14). *Box Plot in Python using Matplotlib*. GeeksforGeeks. https://www.geeksforgeeks.org/box-plot-in-python-using-matplotlib/