**The Impact of Social Media on Workplace Productivity**

**Business Task**

A startup app is aiming to help professionals improve their workplace productivity as well as improving their wellbeing and satisfaction. To implement features that effectively drive behaviour change they need to understand the relationship between digital habits and productivity outcomes.

The key insights they would like to be identified are:

* What digital habits are most associated with low productivity?
* What features should be prioritized to best improve user productivity?
* Do user’s who use focus apps have an improvement in productivity? If so what is the average improvement and how can this app do better?
* Is there a difference in patterns across age groups in how digital habits affect productivity? Should the app use different strategies for the age groups?

**Stakeholders**

Primary Client: Startup app owners

Secondary Stakeholders: App developers, investors and end users.

**Dataset**

I have selected the Social Media Vs Productivity dataset from Kaggle for this task. It contains approximately 30,000 rows each showing an individual’s digital activity and productivity.

Link: <https://www.kaggle.com/datasets/mahdimashayekhi/social-media-vs-productivity>

**Data Organisation**

The dataset has different columns for each person for:

* Digital habits (e.g. daily\_social\_media\_time, screen\_time\_before\_sleep, number\_of\_notifications)
* Work patterns (e.g. work\_hours\_per\_day, breaks\_during\_work)
* Productivity (perceived\_productivity\_score, actual\_productivity\_score)
* Wellbeing (e.g. stress\_level, sleep\_hours, days\_feeling\_burnout\_per\_month)
* Tool usage (e.g. uses\_focus\_apps, has\_digital\_wellbeing\_enabled)
* Demographics (age, gender, job\_type)
* Lifestyle indicators (coffee\_consumption\_per\_day, weekly\_offline\_hours, job\_satisfaction\_score)

The data is stored in a CSV file which will make it easy to work with in Excel, Big Query and R.

**Credibility**

* Reliable: The data is consistently structured with well-labelled variables. There is some missing data, but this can be addressed during the cleaning phase.
* Original: The dataset is not first-hand or survey-based but rather a simulateddataset designed to mimic real-world patterns.
* Comprehensive: It contains 18 variables that span digital behaviour, work, health, and lifestyle — offering a broad perspective on productivity.
* Current: It was last updated within the last month, so it can be considered up to date.
* Cited: The dataset was created by Mahdi Mashayekhi and is hosted publicly on Kaggle.

While the data is not suitable for business-critical decision-making due to its simulated nature, it is acceptable for educational or prototype-level analysis like this case study.

The data set is provided under a creative commons license so is fine for this project.

**Data Privacy and Security**

There are no privacy concerns with this dataset because it does not include any personal identifiers. It will be stored locally during analysis.

**Data Cleaning Steps**

|  |  |  |
| --- | --- | --- |
| **Step** | **Action** | **Reason** |
| 1 | Checked for missing values. | In essential columns, I deleted all rows with missing values such as social media usage, stress level and sleep hours. I kept missing values in columns such as job satisfaction score and productivity scores because I did not think they were as essential for the analysis. |
| 2 | Checked for outliers . | I used the filter option to check that all the columns contained realistic values and spelling was correct. |
| 3 | Fixed data types | I changed all columns to be the correct data types. |

A cleaning log has also been added as a separate worksheet to the file. This keeps track of all the cleaning steps for future reference.

After cleaning, the dataset is now well labelled, has no missing values in essential columns and is ready to be analysed.

**Analysing with Excel**

Following cleaning the file, I began my analysis with Excel. My first step to gain a better understanding of my dataset was to make a descriptive statistics worksheet and do some basic calculations on there.

A white grid with black text

AI-generated content may be incorrect.

Following this, I made a pivot table showing the average of daily social media time and average productivity score against age. I then visualised this data to make it easier to interpret.

The graph indicated minimal variation in social media use or productivity across age groups. This suggests that, within this dataset, age alone is not a strong differentiator in how digital habits affect productivity.

I then built a new pivot table comparing users who use focus apps against those who don’t. I calculated average productivity score and average daily social media time for both groups.

A blue and black text with numbers

AI-generated content may be incorrect.

There was a slight improvement in productivity and reduction in social media time among users who use focus apps. Although the differences are modest, this suggests that focus-enhancing tools may have some positive effect. It also indicates that there is room for improvement, and a well-designed productivity app could potentially amplify this effect.

**Analysis in R**

**The Impact of Social Media on Workplace Productivity**

Oliver Leng

2025-06-16

**Introduction**

This report explores the impact of digital habits on workplace productivity and aims to provide insight into how an upcoming focus app could best tackle digital habits and convert that into better productivity. The analysis focuses on: - How social media use correlates with work hours and productivity - How these effects vary across age groups - Which digital habits (notifications, screen time before bed) are most associated with lower productivity

**Social Media Time and Work Hours**

Social\_Media\_Vs\_Productivity <- Social\_Media\_Vs\_Productivity **%>%**  
 **mutate**(  
 daily\_social\_media\_time = **as.numeric**(daily\_social\_media\_time),  
 social\_media\_range = **case\_when**(  
 daily\_social\_media\_time **<** 1 **~** "<1 hr",  
 daily\_social\_media\_time **<** 2 **~** "1-2 hrs",  
 daily\_social\_media\_time **<** 3 **~** "2-3 hrs",  
 daily\_social\_media\_time **<** 4 **~** "3-4 hrs",  
 TRUE **~** "4+ hrs"  
 ),  
 social\_media\_range = **factor**(social\_media\_range, levels = **c**("<1 hr", "1-2 hrs", "2-3 hrs", "3-4 hrs", "4+ hrs"))  
 )  
  
social\_media\_vs\_work <- Social\_Media\_Vs\_Productivity **%>%**  
 **group\_by**(social\_media\_range) **%>%**  
 **summarise**(  
 avg\_work\_hours = **mean**(work\_hours\_per\_day, na.rm = TRUE),  
 count = **n**()  
 ) **%>%**  
 **ungroup**()  
  
**ggplot**(social\_media\_vs\_work, **aes**(x = social\_media\_range, y = avg\_work\_hours, fill = social\_media\_range)) **+**  
 **geom\_col**() **+**  
 **scale\_fill\_manual**(values = **c**(  
 "<1 hr" = "#1b9e77",  
 "1-2 hrs" = "#d95f02",  
 "2-3 hrs" = "#7570b3",  
 "3-4 hrs" = "#e7298a",  
 "4+ hrs" = "#66a61e"  
 )) **+**  
 **labs**(title = "Average Work Hours by Social Media Time",  
 x = "Social Media Use",  
 y = "Average Work Hours") **+**  
 **theme\_minimal**() **+**  
 **theme**(legend.position = "none")

A graph of different colored squares

AI-generated content may be incorrect.

This visualization shows that there is actually no real change at all in work hours vs social media hours. This suggests that daily work hours remain relatively consistent regardless of social media usage, implying that social media use may not directly reduce time spent working but could still affect productivity.

**Social Media Time Vs Productivity**

**ggplot**(Social\_Media\_Vs\_Productivity, **aes**(x = daily\_social\_media\_time, y = actual\_productivity\_score)) **+**  
 **geom\_point**(alpha = 0.6) **+**  
 **geom\_smooth**(method = "lm", color = "blue") **+**  
 **labs**(title = "Social Media Time vs. Productivity",  
 x = "Daily Social Media Time (hrs)",  
 y = "Actual Productivity Score")

## `geom\_smooth()` using formula = 'y ~ x'

A graph showing a line between a black and a blue line

AI-generated content may be incorrect.

This chart does show a weak negative correlation, meaning that the more time is being spent on social media, the less productive the individual is like to be although there is a wide spread on the data points, showing a big variation in productivity across all social media times.

**Productivity by Age Group and Social Media Time**

*# Create age groups*  
Social\_Media\_Vs\_Productivity <- Social\_Media\_Vs\_Productivity **%>%**  
 **mutate**(age\_group = **case\_when**(  
 age **<** 30 **~** "Under 30",  
 age **<** 50 **~** "30-49",  
 TRUE **~** "50+"  
 ))  
  
*# Summarize average productivity by age group and social media time range*  
Social\_Media\_Vs\_Productivity <- Social\_Media\_Vs\_Productivity **%>%**  
 **mutate**(social\_media\_range = **case\_when**(  
 daily\_social\_media\_time **<** 1 **~** "<1 hr",  
 daily\_social\_media\_time **<** 2 **~** "1-2 hrs",  
 daily\_social\_media\_time **<** 3 **~** "2-3 hrs",  
 daily\_social\_media\_time **<** 4 **~** "3-4 hrs",  
 TRUE **~** "4+ hrs"  
 ))  
  
age\_productivity\_summary <- Social\_Media\_Vs\_Productivity **%>%**  
 **group\_by**(age\_group, social\_media\_range) **%>%**  
 **summarise**(  
 avg\_productivity = **mean**(actual\_productivity\_score, na.rm = TRUE),  
 count = **n**()  
 )

## `summarise()` has grouped output by 'age\_group'. You can override using the  
## `.groups` argument.

**print**(age\_productivity\_summary)

## # A tibble: 15 × 4  
## # Groups: age\_group [3]  
## age\_group social\_media\_range avg\_productivity count  
## <chr> <chr> <dbl> <int>  
## 1 30-49 1-2 hrs 4.93 1270  
## 2 30-49 2-3 hrs 4.96 1602  
## 3 30-49 3-4 hrs 5.00 1652  
## 4 30-49 4+ hrs 5.02 2710  
## 5 30-49 <1 hr 4.97 1379  
## 6 50+ 1-2 hrs 4.91 999  
## 7 50+ 2-3 hrs 4.97 1239  
## 8 50+ 3-4 hrs 4.92 1280  
## 9 50+ 4+ hrs 4.90 2086  
## 10 50+ <1 hr 5.11 1087  
## 11 Under 30 1-2 hrs 4.96 732  
## 12 Under 30 2-3 hrs 4.88 912  
## 13 Under 30 3-4 hrs 4.96 976  
## 14 Under 30 4+ hrs 4.80 1565  
## 15 Under 30 <1 hr 4.84 777

*# Optional: Plot average productivity by age group and social media use*  
**library**(ggplot2)  
**ggplot**(age\_productivity\_summary, **aes**(x = social\_media\_range, y = avg\_productivity, fill = age\_group)) **+**  
 **geom\_col**(position = "dodge") **+**  
 **labs**(title = "Productivity by Age Group and Social Media Use",  
 x = "Daily Social Media Time",  
 y = "Average Productivity Score") **+**  
 **theme\_minimal**()

A graph of a number of people

AI-generated content may be incorrect.

Productivity scores appear relatively stable across age groups for each social media use range, suggesting that age is not a strong moderator in the relationship between social media use and productivity.

**Relationship Between Digital Habits and Productivity**

*# Reshape data to long format*  
long\_data <- Social\_Media\_Vs\_Productivity **%>%**  
 **select**(actual\_productivity\_score, daily\_social\_media\_time, number\_of\_notifications, screen\_time\_before\_sleep) **%>%**  
 **pivot\_longer**(  
 cols = **-**actual\_productivity\_score,  
 names\_to = "digital\_habit",  
 values\_to = "habit\_value"  
 )  
  
*# Plot*  
**ggplot**(long\_data, **aes**(x = habit\_value, y = actual\_productivity\_score)) **+**  
 **geom\_point**(alpha = 0.3, color = "steelblue") **+**  
 **geom\_smooth**(method = "lm", se = FALSE, color = "red") **+**  
 **facet\_wrap**(**~** digital\_habit, scales = "free\_x") **+**  
 **labs**(  
 title = "Relationship Between Digital Habits and Productivity",  
 x = "Digital Habit Value",  
 y = "Productivity Score"  
 ) **+**  
 **theme\_minimal**()

## `geom\_smooth()` using formula = 'y ~ x'

## Warning: Removed 1503 rows containing non-finite values (`stat\_smooth()`).

## Warning: Removed 1503 rows containing missing values (`geom\_point()`).

A graph of blue dots

AI-generated content may be incorrect.

This visualization shows that the most important variable in productivity for the start up app to target is the daily social media score because it shows the strongest correlation with decreased productivity of all the factors.

**Data Source**

This dataset was made by the Kaggle user Mahdi Mashayekh and can be found here: <https://www.kaggle.com/datasets/mahdimashayekhi/social-media-vs-productivity>

**Act and Share**

I have made a full PowerPoint presentation sharing the analysis with stakeholders and recommending actions based off this for the final part of the project. The PowerPoint presentation can also be found in my GitHub repository.