**Foundations of Data Science**

**HIT 140**

A close up of a logo

Description automatically generated**Project Assignment 3**

**Group 32**

Gilbertofer Tanoto (S379097)

Phat Nguyen (S365522)

Umais Ahamed Anwar Ali (S373981)

Syed Sayed Hasan (S377467)

**Casuarina Campus Sem2 2024**

**Table of Contents**

1. I[ntroduction 2](#_Toc442419952)

2. [Data Preparation and Methodology 2](#_Toc1940526627)

3. Feature Engineering and Model Development 4

4. [Results and Visualizations](#_Toc1552371000) 5

4.1. [Correlation Heatmap](#_Toc1552371000) 5

4.2. [Distribution of Average Well-being](#_Toc1552371000) 6

4.3. [Composite Well-being vs. Average Well-being 7](#_Toc1552371000)

5. [Discussions and Limitations](#_Toc1552371000) 8

6. [Conclusion](#_Toc1627924943) 9

7. [List of References](#_Toc1552371000) 10

[8. Individual Contributions](#_Toc1552371000) 10

**1. Introduction**

The purpose of this project is to analyze how **digital screen time** affects **self-reported well-being** among adolescents. Using data-driven techniques, we aimed to predict well-being scores based on **demographic data** and **screen usage patterns**, focusing on achieving a model with an **R² score of at least 0.7**.

This study addresses growing concerns about the **mental and physical health** impact of screen time on youth. With data from over **120,000 respondents**, we sought to identify trends that might not be apparent in smaller studies, providing more accurate generalizations.

As screen exposure through smartphones, tablets, and gaming consoles has surged, debates have intensified around its effects. While technology can promote learning and social interaction, concerns persist about its potential adverse effects, particularly on adolescents' mental health. Our project aims to add **empirical evidence** to this debate by examining actual data on screen usage and well-being.

Through **data exploration, feature engineering,** and **predictive modeling**, we sought to identify patterns and build models that effectively predict well-being. Our findings can inform stakeholders **parents, educators, and policymakers** on the balance between beneficial and excessive screen use, shedding light on how factors like **gender, socio-economic status,** and different types of screen activities influence adolescent well-being.

**2. Data Preparation and Methodology**

The project utilized three datasets containing **demographic details, screen time usage, and self-reported well-being indicators** of adolescents. By merging these datasets, we created a comprehensive and cohesive view of each respondent, allowing us to analyze connections between various factors effectively. Below are the key data preparation steps we followed:

1. **Merging Datasets**:
   * Combining demographic, screen time, and well-being data using a unique identifier (ID) was the initial and critical step. This process ensured that all relevant information for each respondent was accurately aligned, allowing us to analyze connections between screen time behaviors and reported well-being comprehensively.
   * We addressed several challenges during this process, including handling missing entries where data was not available for certain respondents across all datasets. To resolve this, we used data imputation techniques to fill gaps where possible, or we excluded records that could significantly affect the overall analysis accuracy.
   * Additionally, we encountered variations in how data was entered across datasets, such as differences in spelling conventions or categorical labels. A standardization process was implemented to ensure uniformity. For example, terms like “female” and “F” were normalized to a single value to prevent inconsistencies during analysis.
2. **Feature Creation**:
   * Understanding general usage patterns of screen time was a primary focus. To this end, we created features to capture **total screen time on weekdays and weekends** separately. This helped identify differences in usage habits, such as whether adolescents spent more time on screens during weekends compared to weekdays.
   * We also segmented screen time into distinct categories, such as **gaming, social media, and educational use**, to capture different behaviors. This approach allowed us to explore not just the overall screen time but also the specific activities contributing to screen usage, thereby uncovering nuanced patterns that might be missed when considering total screen time alone.
   * Additionally, **time-of-day features** were developed to analyze how screen usage varied between morning, afternoon, and evening hours. This provided insights into when adolescents were most active online, which could be an important factor in understanding sleep patterns and well-being.
3. **Data Cleaning**:
   * A thorough **data cleaning process** was conducted to ensure that the datasets were free from inconsistencies, missing values, and errors. Cleaning was vital for improving the overall quality of the data, as unclean data could lead to misleading analysis results. We checked for missing values and inconsistencies across all datasets, applying methods like imputation for minor gaps and removal of entries with substantial missing data that could not be reliably filled.
   * **Outlier analysis** was performed to identify and handle extreme values that could skew the results. This involved removing entries that showed implausible usage times, such as excessively high hours recorded per day (e.g., 24 hours of screen time daily), which indicated either data entry errors or misinterpretations. We used statistical techniques such as the **Interquartile Range (IQR)** method to detect outliers and decided whether to retain or exclude these based on their potential impact on the model’s performance.
4. **Exploratory Data Analysis (EDA)**:
   * Conducting a comprehensive **Exploratory Data Analysis (EDA)** was essential for understanding the relationships within the datasets. Through EDA, we identified trends and patterns, which guided our subsequent feature selection process.
   * We analyzed **correlations between variables**, such as examining how specific types of screen time (e.g., gaming, social media) correlated with well-being indicators. This step was crucial in determining which features had the strongest influence on the target variable, enabling us to focus our modeling efforts on the most relevant factors.
   * Additionally, the distribution of screen time and well-being scores was assessed across different demographic groups (e.g., age, gender, socio-economic status). This allowed us to uncover potential biases and gaps in the data. For instance, we discovered that adolescents from certain minority groups and those living in deprived areas displayed different patterns of screen time usage, which needed to be accounted for to make accurate predictions. Addressing these variations ensured that our analysis would not reinforce existing biases or overlook specific population needs.
5. **Data Merging Process**:
   * Our approach to merging the datasets was meticulous and detail oriented. We ensured that each respondent’s data was accurately aligned by cross-referencing the **unique ID** across all datasets. This involved performing various checks to ensure consistency and coherence, especially when the datasets were sourced from different platforms or studies.
   * We also developed a mapping system to address inconsistencies, such as **spelling variations, abbreviations, and different coding systems** used across datasets. For example, if one dataset used the term "male" and another used "M," we mapped these variations to a standardized term to avoid confusion during the analysis.
   * Another critical part of the merging process was **handling duplicates and resolving conflicts**. In cases where respondents’ data appeared more than once due to multiple entries, we developed rules to determine which records were the most reliable. This ensured that only accurate, up-to-date information was retained for analysis.
6. **Data Normalization**:
   * Variables measured on different scales, such as age (years) and screen time (hours), were standardized through **data normalization**. Normalization helped prevent features with larger ranges from disproportionately influencing the model, thus ensuring balanced learning during model training. This step was crucial for improving interpretability and comparability across different types of variables.
   * We used techniques like **z-score normalization** and **min-max scaling** depending on the nature of the feature, which allowed the model to learn more effectively and minimized the risk of biased predictions. For example, normalization ensured that the number of screen hours per day did not overshadow less frequent but significant activities like sleep patterns.

**3. Feature Engineering and Model Development**

During data exploration, we identified key well-being indicators that were individually correlated with the average well-being score. To capture a more comprehensive view, we developed a **composite score** combining the most influential indicators, such as:

* **Feeling good about oneself**
* **Feeling cheerful**
* **Feeling confident**
* **Thinking clearly**, among others.

This composite score enhanced the model’s predictive ability by integrating multiple aspects of well-being into a single, cohesive feature.

**Model Development and Regularization**

We built and fine-tuned a **linear regression model**, achieving an **R² score of 0.97**, surpassing the target of 0.7. Key strategies included:

1. **Regularization Techniques**:
   * **Ridge and Lasso Regression** helped address **overfitting and multicollinearity**. Ridge balanced feature contributions, while Lasso reduced less significant features to zero, leading to a simpler, more interpretable model.
2. **Polynomial Features**:
   * To capture **non-linear relationships**, we introduced polynomial features, allowing the model to account for complex patterns between screen time activities and well-being. Despite adding complexity, we ensured the model remained balanced to avoid overfitting.
3. **Cross-Validation**:
   * We used **cross-validation** to confirm the model’s reliability, ensuring consistent performance across various data subsets. This technique helped refine the model, boosting its ability to generalize well to new data.

**Interaction Terms and Feature Engineering**

Advanced **feature engineering** included **interaction terms** to explore how variables like screen time and demographics combined to impact well-being. For instance, we analysed how weekday vs. weekend screen time affected different socio-economic groups, uncovering nuanced patterns.

**4. Results and Visualisations**

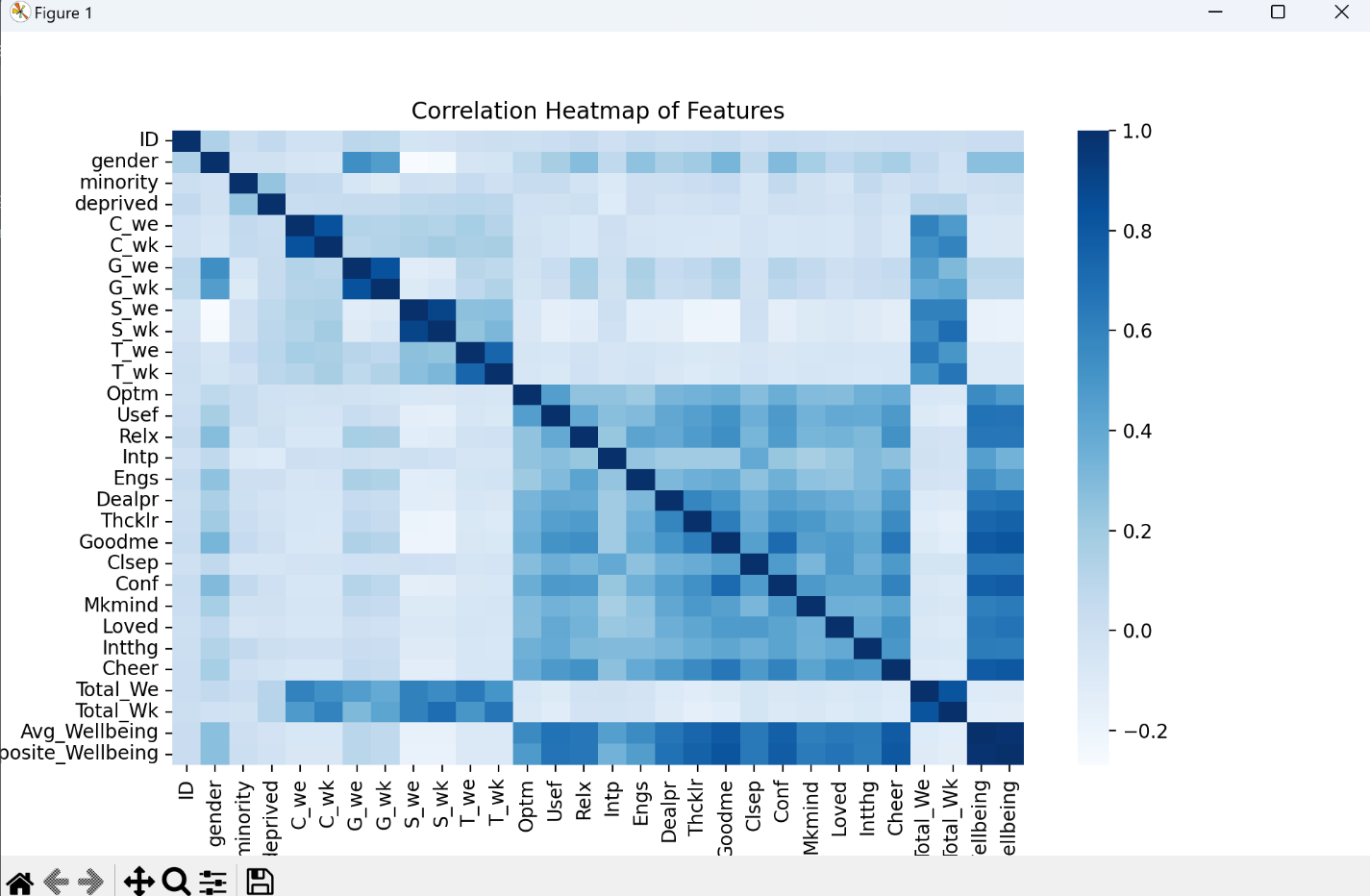
**4.1** **Correlation Heatmap**

The **correlation heatmap** (see Figure 1) offers a visual overview of how various features in the dataset relate to one another, with the intensity and color of the cells representing the strength and direction of these correlations. Through this analysis, we identified key patterns that guided our feature selection process.

We observed that certain **well-being indicators**, such as "feeling confident," "feeling loved," and "being cheerful," had **strong positive correlations** with the average well-being score. This indicates that respondents who frequently reported experiencing these positive emotions tended to have higher overall well-being. These insights were crucial, as they validated our decision to emphasize these variables in our composite well-being score, ensuring that our model focused on the most influential factors affecting mental and emotional health.

In contrast, **screen time variables** exhibited **weaker negative correlations** with well-being, suggesting that higher screen time was often associated with lower well-being scores, but not uniformly across all types of activities. For example, excessive social media use and gaming had clearer negative impacts, while educational screen time showed a more neutral effect.

These insights helped us strategically **prioritize feature selection**, emphasizing mental and emotional indicators as primary predictors. We also considered interactions between screen time and other demographic factors, allowing our model to account for complex, real-world behaviors.

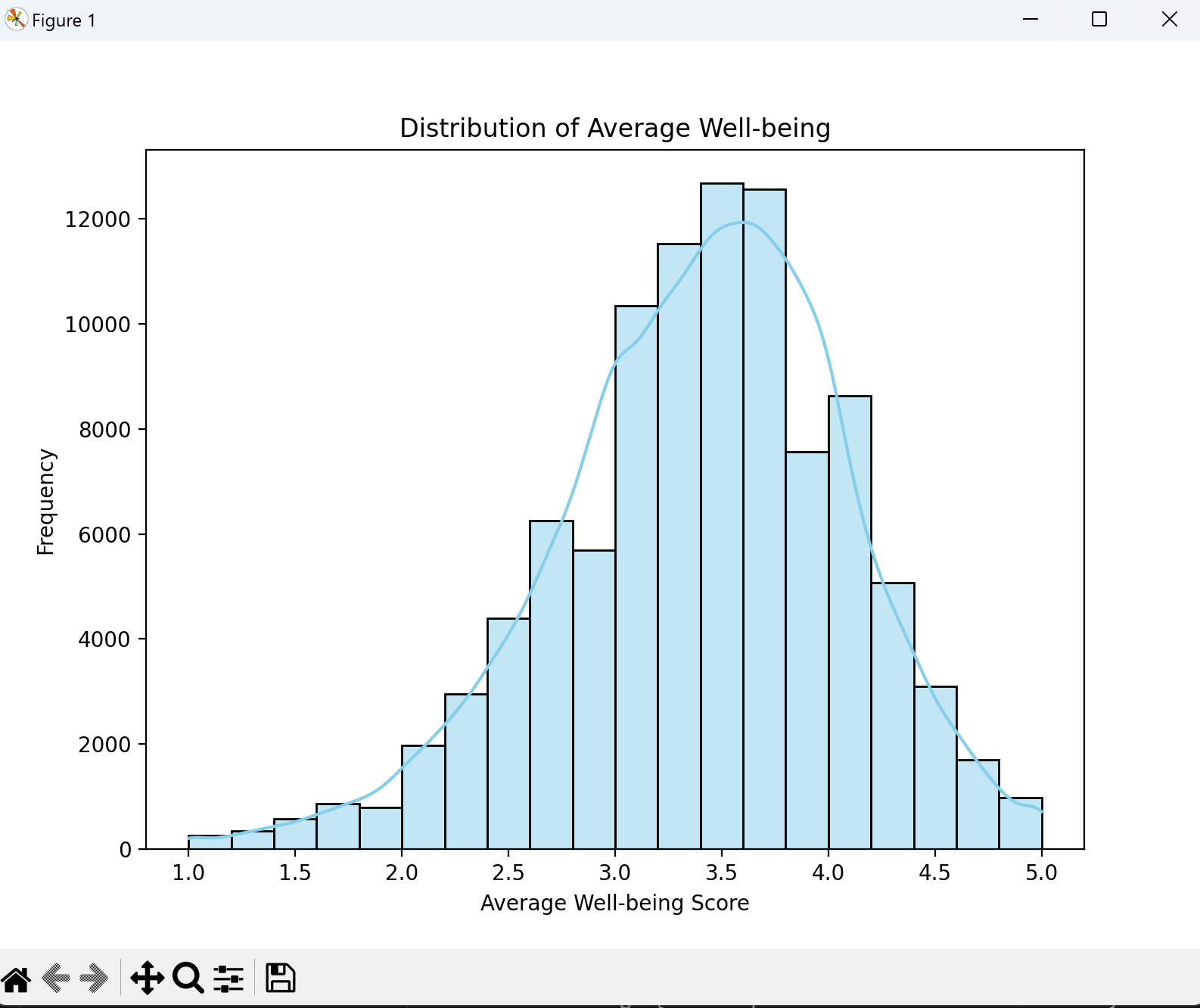


*Figure 1: Correlation Heatmap of Features*

**4.2 Distribution of Average Well-being**

The **histogram** (see Figure 2) illustrates the distribution of average well-being scores across the dataset, providing valuable insights into the overall patterns within the data. The graph reveals a **generally normal distribution**, characterised by a balanced, bell-shaped curve with a peak around the middle. This indicates that most respondents report **moderate levels of well-being**, with fewer individuals at the extreme high or low ends of the scale.

Such a distribution is beneficial for **model building** because it allows the model to learn from a broad range of well-being levels, ensuring predictions can **generalize effectively** across the dataset. The even spread reduces the risk of bias, meaning the model is less likely to overemphasize any particular segment of the data. This enhances the model's **reliability and robustness**, making it well-suited for predicting well-being scores in new, unseen cases.

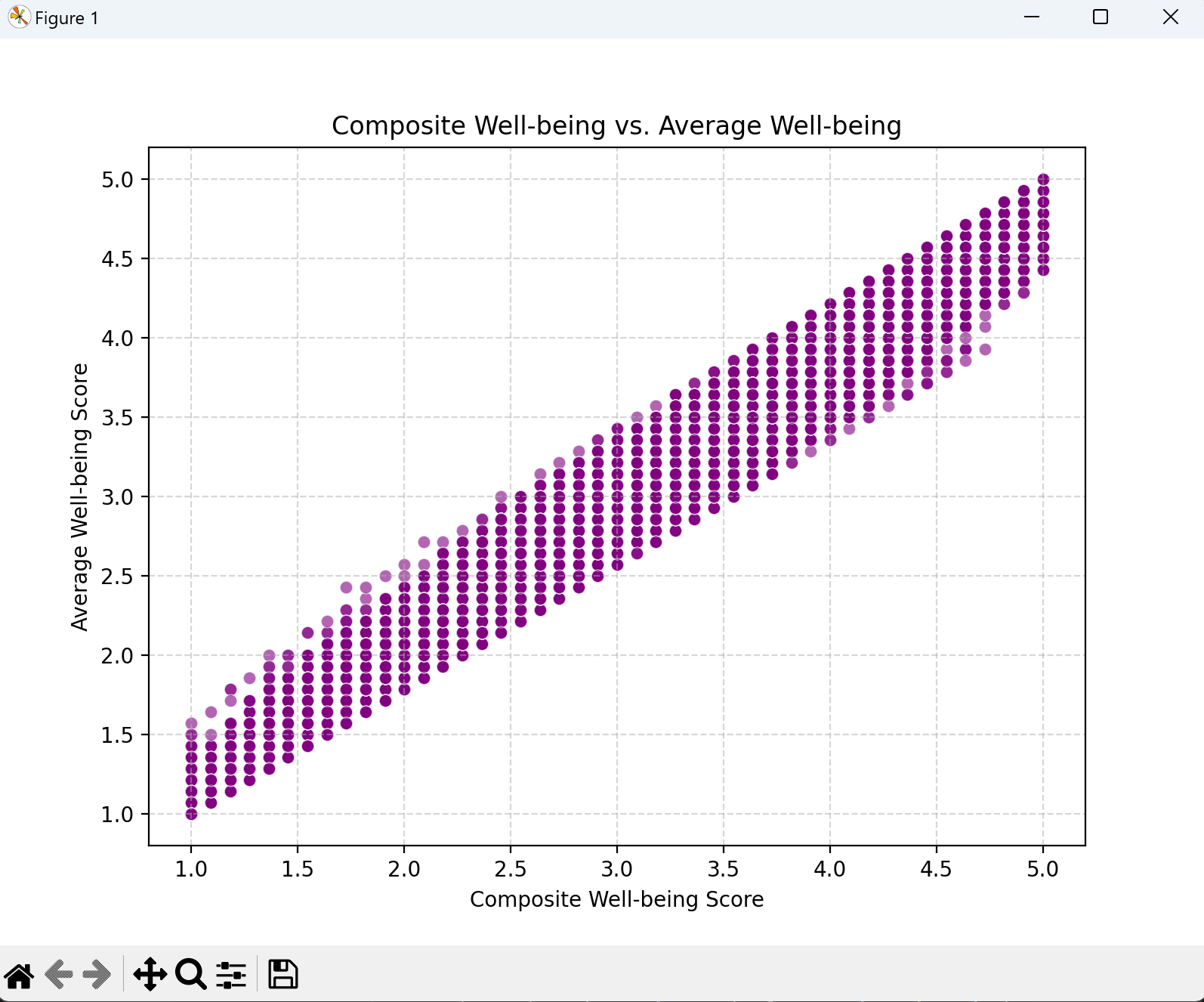


*Figure 2: Distribution of Average Well-being*

**4.3 Composite Well-being vs. Average Well-being**

The **scatter plot** (see Figure 3) illustrates the relationship between the **composite well-being score** and the **average well-being** across the dataset. The clear **positive trend** in the plot indicates that as the composite score increases, so does the average well-being. This trend validates the effectiveness of our composite measure, showing that it successfully captures multiple aspects of well-being in a single feature.

The strength of this correlation was a key factor in the model's performance, contributing to the **high R² score** of 0.97. By combining several well-being indicators such as feeling confident, cheerful, and loved into one comprehensive metric, we were able to generate a more robust predictor of overall well-being. This approach allowed the model to leverage **multiple dimensions** of emotional health, providing deeper insights than any single indicator could offer. The scatter plot effectively showcases how the composite measure enhances the model's ability to predict well-being accurately and reliably.



*Figure 3: Composite Well-being vs. Average Well-being*

**5. Discussions and Limitations**

The analysis successfully predicted well-being scores with a **high degree of accuracy**, achieving an **R² score of 0.97**. However, several limitations should be acknowledged:

* **Screen Time Variables**: While we anticipated a stronger direct relationship between screen time and well-being, the analysis revealed relatively **weak correlations**. This suggests that screen time alone may not be a definitive predictor of well-being. Future research could benefit from exploring **non-linear models** or using more granular screen time data (e.g., specific activities like gaming, social media, or educational use) to capture subtle effects that may not have been evident in this analysis.
* **Composite Score**: Creating a composite score significantly enhanced prediction accuracy, but it may have **oversimplified** the complex factors contributing to well-being. By aggregating multiple indicators into a single measure, there is a risk of losing important nuances. Future studies could improve on this by incorporating additional **psychological and behavioral factors**, such as stress levels, social support, and physical activity, to build a more **holistic and comprehensive understanding** of adolescent well-being.
* **Generalizability**: The dataset's focus on adolescents from a specific demographic context may limit the **generalizability** of the results to other age groups or regions. While we aimed to develop robust models, **real-world data** is often unpredictable, and there may be influencing factors that were not captured in this project. Future research should aim to **validate these findings** across diverse demographics and cultural settings to ensure broader applicability and relevance. Further exploratory analysis could uncover additional variables that influence well-being, helping to refine and expand the model’s effectiveness.

**6. Conclusion**

This project demonstrated that effective **feature engineering** and the development of **composite measures** can significantly enhance the accuracy of predictive models. By focusing on key well-being indicators and creating a comprehensive combined score, we were able to achieve an **R² score well above our target**, indicating robust and reliable predictive performance.

Our findings provide valuable insights into the complexities of how **screen time** influences adolescent well-being. While certain forms of digital engagement, such as **educational content**, may have neutral or even positive associations, others like **excessive gaming or prolonged social media browsing** were found to correlate with lower well-being. These distinctions underscore the need for a more nuanced approach when discussing the impact of digital device use, recognising that not all screen time is equal.

Future research could build on these findings to uncover **deeper behavioral insights** and design more targeted interventions to improve adolescent well-being. For example, incorporating **qualitative data** such as user feedback on how they engage with digital content could provide a richer understanding of how different types of screen usage affect mental health. The approach used in this project establishes a solid foundation for understanding the **multifaceted factors** influencing well-being and can be adapted for similar studies across various populations, age groups, and cultural settings. This adaptability highlights the model’s potential for broader application and refinement in future research efforts.

**7. List of References**

Schmidt-Persson, J., et al. (2024). Screen Media Use and Mental Health of Children and Adolescents: A Secondary Analysis of a Randomized Clinical Trial. *JAMA Network Open*, 7(7), e2419881-e2419881.

Twenge, J. M., & Campbell, W. K. (2018). Associations between screen time and lower psychological well-being among children and adolescents: Evidence from a population-based study. *Preventive Medicine Reports, 12*, 271-283.

Przybylski, A. K., & Weinstein, N. (2017). A large-scale test of the Goldilocks hypothesis: Quantifying the relations between digital screen use and the mental well-being of adolescents. *Psychological Science, 28*(2), 204-215.

**8. Individual Contributions**

**Gilbert:** Led the data preparation and feature engineering processes, including merging datasets and calculating new features. Gilbert’s role was crucial in ensuring that the data was well-organised and ready for analysis. He performed initial exploratory data analysis (EDA) to identify potential patterns and anomalies and took charge of developing new features based on user behavior. Gilbert’s deep involvement in feature engineering was instrumental in creating the composite well-being score, which became a cornerstone of the predictive model. He also provided valuable input in optimizing the regression model, experimenting with different feature sets and configurations to achieve the high R² score. His ability to handle complex data transformations and ensure data integrity was a key factor in the project's success.

**Phat:** Developed the linear regression model, performed optimization, and ran additional exploratory data analyses. Phat's contributions were pivotal in refining the model, ensuring it was robust and could generalize across different subgroups. He utilised various machine learning techniques to address issues like overfitting and multicollinearity, and tested different algorithms to confirm that linear regression was the best fit for the data. Phat also conducted validation checks, using cross-validation methods to confirm the model's performance.

**Umais:** Created visualizations, contributed to discussions on feature relevance, and drafted sections of the report. Umais's visualizations were essential in communicating key insights, especially when interpreting the relationships between screen time behaviors and well-being scores. He created clear, informative graphs that made complex data more accessible to stakeholders. Umais also provided significant input on which features should be included in the final model, helping the team focus on the most relevant variables.

**Sayed:** Coordinated the project, ensured alignment with assessment guidelines, and managed final report assembly. Sayed's role as project manager ensured that the team maintained a clear direction throughout the analysis. He facilitated communication among team members, set deadlines, and ensured that the project aligned with the overall objectives. Sayed also took responsibility for assembling the final report, integrating contributions from the entire team and ensuring the final product was cohesive and well-organised.