DSA 210 Project Final Report: The Impact of Internet Usage in the USA

1. Introduction

1.1. Background & Motivation

Over the past few decades, the global landscape has witnessed a profound and rapid increase in internet access, transforming it into a fundamental component of daily life for billions of individuals worldwide. This pervasive adoption spans diverse sectors, from educational institutions and professional environments to intellectual pursuits and commercial enterprises, offering significant benefits such as enhanced communication capabilities and unparalleled access to vast information repositories. The internet's integration into society has undeniably reshaped how individuals learn, work, and interact, fostering new avenues for progress and development.

However, this digital transformation presents a complex duality. While the positive impacts are extensively documented and widely acknowledged, there are growing concerns regarding the potential negative consequences associated with excessive internet usage. These concerns include a perceived link to declining mental health indicators and an increase in distractions, prompting a critical need to rigorously examine the long-term societal effects of internet penetration. The project is designed to quantify these varied impacts, contributing to a more informed public discourse and policy development. This perspective, which acknowledges both the advantages and disadvantages of technological advancement, is fundamental to understanding the internet's role in contemporary society. It suggests that the project is not merely a technical exercise but a contribution to a relevant public policy discussion, implicitly setting an expectation for findings that may show both positive and negative correlations across different domains.

1.2. Project Objective

The primary objective of this study is to conduct a comprehensive analysis of the impact of increasing internet usage within the United States. This analysis is achieved by enriching a core dataset on internet usage trends with supplementary data across several critical societal dimensions: education performance, productivity metrics, mental health, and overall well-being. The multi-faceted approach taken in this study aims to provide a holistic view of the internet's influence, recognizing that its impact is not isolated but affects various fundamental pillars of society.

The overarching goal is to identify significant patterns and correlations between internet usage and these diverse aspects, utilizing publicly available data. The ultimate aim is to derive meaningful understandings that contribute to a deeper comprehension of the internet's implications on societal well-being and efficiency. By identifying these relationships, the study seeks to foster a balanced and beneficial approach to internet consumption, positioning the report as a valuable resource for guiding policy and behavioral adjustments towards achieving a fruitful internet consumption balance.

1.3. Research Questions

This project is guided by the following key research questions, which aim to systematically explore the multifaceted impact of internet usage:

- How has the rise in internet usage impacted education performance?
- Is there a relationship between internet usage and mental health?
- How does internet usage affect productivity?

2. Data Sources & Methodology

2.1. Data Collection

Primary Internet Usage Dataset

The foundational dataset for this study focuses on internet usage trends within the USA. This data was accessed via the Excel file internet data.xlsx. Key features extracted from this dataset include 'Year', spanning from 1990 to 2022/2023, and 'Internet Users (%)', which represents the percentage of the population with internet access. This metric serves as the independent variable across all subsequent analyses. The data clearly illustrates a significant increase in internet penetration, rising from a nascent 0.78% in 1990 to over 97% by 2022.

Enrichment Datasets

To facilitate a multi-faceted analysis, the primary internet usage data was enriched with supplementary datasets covering productivity, education, and health.

- Productivity:
  Productivity data, encompassing both Total Factor Productivity (TFP) and Labour Productivity, was obtained from the Bureau of Labor Statistics (BLS) through an Excel file titled total-factor-productivity-major-sectors-historical.xlsx. This dataset includes 'Year', 'TFP Index', which ranged from approximately 78 to 105

during the internet era, and 'Labour Productivity Index', which ranged from about 57 to 111 over the same period. These indices provide quantitative measures of efficiency and output per worker within the private business sector.

- Education:
  Education performance data was collected from authoritative sources such as the National Center for Education Statistics (NCES) and the Department of Education, utilizing Excel files including sat scores.xlsx and edu data.xlsx. The key features extracted from these sources include:
  - **SAT Scores:** 'Math' and 'Critical Reading and writing' scores, spanning from 1972 to 2023. The 'Critical Reading and writing' score was derived as an average of individual Critical Reading and Writing scores.
  - **NAEP Math Scores:** Average mathematics test scores for '9-year-old' and '13-year-old' students, with data points available from 1973 to 2022/2023.
- Health:
  Health data, comprising mental health metrics (DALYs and Deaths) and obesity prevalence, was sourced from reputable public health organizations such including the World Health Organization (WHO) and the CDC, through Excel files mental health data.xlsx and Obesity Data.xlsx. The key features from these datasets are:
  - **Obesity:** The 'mean_count' metric, representing the mean obesity count, with data available from 1990 to 2025.
  - **Mental Health DALYs (Disability-Adjusted Life Years):** DALYs for 'Anxiety disorders', 'Depressive disorders', 'Mental disorders' (aggregate), and 'Substance use disorders', covering the period from 1990 to 2021.
  - **Mental Health Deaths:** Death counts specifically for 'Mental disorders' and 'Substance use disorders', also spanning from 1990 to 2021.

The selection of a broad 'Internet Users (%)' metric, which is a high-level aggregate, combined with more granular metrics from the enrichment datasets (e.g., specific SAT sections, different age groups for math, various mental health disorders), allows for a nuanced exploration of the internet's differentiated impacts, moving beyond a monolithic "internet effect." The choice of diverse, time-series enrichment data enables a longitudinal view of the internet's influence, facilitating the identification of trends and potential shifts in impact over time. Furthermore, the use of specific health metrics such as DALYs, death counts, and obesity prevalence, rather than a general "overall well-being" measure, allows for more precise conclusions and potentially more targeted policy recommendations.

2.2. Data Preprocessing

The preprocessing phase was a critical step in preparing the diverse datasets for integrated analysis, ensuring consistency, handling missing values, and standardizing formats.

- **Internet Data:** The initial steps involved loading Excel files, skipping metadata rows, and dropping columns that contained only NaN values. A crucial transformation involved "melting" the wide-format data, where years were represented as columns, into a long format with explicit 'Year' and 'Internet Users (%)' columns. This facilitated time-series analysis and subsequent merging operations. The 'Year' column was consistently converted to an integer data type, and the 'Country Name' column was removed, as the analysis specifically focused on the USA.
- **Productivity Data:** Raw productivity data was meticulously filtered to isolate 'Total factor productivity' and 'Labor productivity' specific to the 'Private business sector'. Irrelevant metadata columns, such as 'NAICS', 'Sector', 'Measure', and 'Units', were systematically removed. The remaining value columns were then appropriately renamed to 'TFP Index' and 'Labour Productivity Index' for clarity and consistency. The 'Year' column in these datasets was also standardized to an integer type to ensure compatibility during merging.
- **Education Data:**
  - **SAT Scores:** For the SAT scores data, values represented by '-' in the 'Critical Reading' and 'Writing' columns were identified and replaced with np.nan to accurately represent missing data. These columns were then converted to a numeric data type, with errors coerced to NaN. A new, combined 'Critical Reading and writing' score was computed as the row-wise average of the individual scores, with NaN values ignored in the calculation. The original 'Critical Reading' and 'Writing' columns were subsequently dropped.
  - **9yo/13yo Math Scores:** Specific rows containing average mathematics test scores for 9-year-olds and 13-year-olds were extracted from the raw data. These datasets were then "melted" to create 'Year' and 'Score' columns, with the year values extracted using regular expressions to ensure a clean four-digit year format, which was then converted to a numeric type. Any NaN values present in the 9-year-old scores dataset were explicitly removed to maintain data integrity.
- **Health Data:**
  - **Obesity Data:** The obesity dataset was filtered to include only 'Country' level data and 'Both' sexes, while excluding specific age groups (e.g., '2 to 14'). The 'mean_count' of obesity cases was aggregated by 'year_id' to obtain annual averages, and the data was further filtered to include years up to and

including 2025.

    - **Mental Health Data (DALYs & Deaths):** This dataset was filtered to include data for 'All ages' and 'Both' sexes. Numerous irrelevant columns, including various identifiers and statistical bounds, were removed. The streamlined data was then separated into two distinct dataframes: one for df_deaths and another for df_dalys, based on the measure_name column. Both dataframes were subsequently pivoted to transform 'cause_name' (e.g., 'Anxiety disorders', 'Substance use disorders') into new columns, with the corresponding 'val' (value) populating these columns.
- **Data Merging:** All preprocessed enrichment datasets (TFP, Labour Productivity, SAT scores, 13-year-old scores, 9-year-old scores, obesity data, DALYs, and deaths data) were systematically merged with the internet usage data. The merging operation was performed primarily on the 'Year' column using an inner join to ensure that only years common to all datasets were included. For the obesity data, a specific merge using left_on='year_id' and right_on='year' was necessary due to differing column names, with year_id subsequently dropped to remove redundancy. To ensure a complete time series for analysis, a continuous range of years (1990-2023) was established as a base, and any remaining NaN values were imputed using a forward fill (ffill()) followed by a backward fill (bfill()). This imputation strategy ensures a complete dataset, which is often necessary for statistical and machine learning models.

2.3. Analytical Approach

The analytical approach employed in this study combines both traditional hypothesis testing and machine learning methodologies to provide a robust and multi-faceted understanding of the relationship between internet usage and various societal outcomes. This multi-pronged approach strengthens the findings by validating relationships through different statistical lenses, moving from exploratory statistical tests to a more formal predictive modeling approach.

Hypothesis Tests Conducted

- **Pearson Correlation Test:** This statistical test was employed to quantify the strength and direction of linear relationships between internet usage and each of the continuous target metrics across productivity, education, and health domains. The purpose was to statistically validate initial observations of association.
    - The test determined if a statistically significant linear relationship exists between 'Internet Users (%)' and:
        - **Productivity:** TFP Index, Labour Productivity Index.
        - **Education:** SAT Math, SAT Critical Reading/Writing, 13-year-old scores,

9-year-old scores.

        - **Health:** DALYs (Anxiety, Depressive, Mental, Substance Use), Deaths (Mental, Substance Use), Obesity Mean Count.

    - For each pair, a Null Hypothesis (H0) stating no linear relationship was tested against an Alternative Hypothesis (H1) stating a significant linear relationship.

- **ANOVA (Analysis of Variance) Test:** This test was utilized to assess whether there were statistically significant differences in the mean values of productivity and academic performance metrics across different quartiles of internet usage. This provided a group-based perspective on the impact of varying internet penetration levels.
    - The test determined if average productivity (TFP Index, Labour Productivity Index) and academic performance (Math, Critical Reading/Writing, 13-year-old scores, 9-year-old scores) significantly differ across internet usage quartiles (Q1-Q4).

Machine Learning Models Applied

- **Linear Regression:** This supervised learning algorithm was applied to model the linear relationship between 'Internet Users (%)' (as the independent variable) and each of the continuous target variables across the three domains. The primary goal was to quantify the magnitude of the effect (slope), establish a baseline (intercept), and assess the model's explanatory power (R-squared) and prediction accuracy (Mean Squared Error/Root Mean Squared Error).
    - The models aimed to:
        - **Quantify Impact:** Determine the change in a target metric for every 1% increase in internet users (slope).
        - **Assess Explanatory Power:** Evaluate how much of the variance in target metrics is explained by internet usage (R-squared).
        - **Evaluate Prediction Accuracy:** Measure the typical error in predictions (MSE/RMSE).
        - **Generalizability (Education):** For education, cross-validated R-squared was also employed to assess model stability and generalization to unseen data.
    - Linear regression models were applied to all productivity metrics (TFP, Labour Productivity), all academic performance metrics (SAT Math, SAT Reading/Writing, 13-year-old, 9-year-old scores), and all health metrics (DALYs for various disorders, Deaths for mental/substance use, Obesity mean count).

3. Analysis & Findings: Internet vs. Productivity

3.1. Hypothesis Testing

The initial statistical investigations, utilizing Pearson correlation and ANOVA tests, provided compelling evidence of a significant relationship between internet adoption and productivity in the USA.

Pearson Correlation Results

- **Internet vs Total Factor Productivity Index:** A near-perfect positive correlation was observed, with a Pearson correlation coefficient (r) of 0.9794 and a p-value of 0.0000. This indicates that as internet usage increases, Total Factor Productivity (TFP), a key measure of technological efficiency and innovation, tends to increase proportionally.
- **Internet vs Labour Productivity Index:** A very strong positive correlation was found, with an r-value of 0.9640 and a p-value of 0.0000. This implies that higher internet adoption is consistently associated with greater output per worker.

The interpretation of these results is straightforward: both p-values are approximately 0.000, which is substantially below the conventional significance threshold of α=0.05. This provides overwhelming statistical evidence to reject the null hypotheses for both relationships, confirming a statistically significant linear association between internet usage and both TFP and Labour Productivity metrics.

ANOVA Test Results

The ANOVA analysis further supported these findings by examining differences in productivity across distinct internet usage quartiles.

- **TFP Index:** The ANOVA test yielded an F-statistic of 94.89, with an extremely low p-value of 4.27e-15, and an Eta-squared value of 0.9075.
- **Labour Productivity Index:** Similarly, the test for Labour Productivity showed an F-statistic of 95.62, a p-value of 3.86e-15, and an Eta-squared of 0.9082.

The interpretation of these ANOVA results indicates that both productivity metrics showed statistically significant differences across internet usage quartiles (p-value < 0.05). The high F-statistics signify strong group differences, and the exceptionally high Eta-squared values (approximately 0.91) suggest that internet usage explains a very large proportion of the variance in productivity metrics. This reinforces a strong association between internet usage levels and productivity differences, underscoring the crucial role of digital infrastructure in economic output.

Table 1: Summary of Productivity Hypothesis Tests

Metric	Test Type	Pearson (r)	p-value (Pearson)	F-statistic (ANOVA)	p-value (ANOVA)	Eta-squared (ANOVA)
Total Factor Productivity	Pearson / ANOVA	0.9794	0.0000	94.89	4.27e-15	0.9075
Labour Productivity Index	Pearson / ANOVA	0.9640	0.0000	95.62	3.86e-15	0.9082

This table provides a concise overview of the statistical significance and strength of the relationships identified by both correlational and group-difference tests, allowing for quick comparison and validation of findings.

3.2. Machine Learning Models

Linear regression models were applied to quantify the precise impact of internet usage on Labour Productivity and Total Factor Productivity, providing insights into the magnitude of these relationships and the models' predictive capabilities.

Linear Regression Results for Labour Productivity Index (LPI)

The linear regression model for LPI demonstrated a strong positive relationship with internet usage:

- **Slope (Coefficient):** 0.5059. This implies that for every 1% increase in Internet Users, the Labour Productivity Index is predicted to increase by approximately 0.506 points. A 10% rise in Internet Users is associated with a +5.06 LPI points increase, which is considered a moderate but meaningful effect given the observed LPI range in the dataset (57.0 to 111.6).
- **Intercept:** 55.6741. This represents the theoretical baseline LPI when Internet Users are at 0%.
- **R-squared (R^2):** 0.8834. This indicates that 88.3% of the LPI variance is explained by Internet penetration, suggesting a strong fit of the model. Internet adoption appears to be a key driver of productivity trends.
- **Mean Squared Error (MSE):** 24.9044. The average prediction error is approximately 4.99 LPI points (calculated as the square root of 24.9). This error,

representing about 5% of the LPI range, is considered acceptable for social science data.

Linear Regression Results for Total Factor Productivity (TFP)

The linear regression model for TFP also revealed a highly significant positive relationship:

- **Slope (Coefficient):** 0.2496. Every 1% increase in Internet Users corresponds to a 0.25-point rise in TFP. A 10% increase in Internet Users leads to a +2.5 TFP points increase, which is a smaller but systematic effect given the TFP range in the dataset (approximately 78 to 105).
- **Intercept:** 78.1768. This is the theoretical baseline TFP when Internet Users are at 0%.
- **R-squared (R^2):** 0.9464. This indicates an exceptional fit, with 94.6% of TFP variance explained by Internet penetration, suggesting it is a dominant predictor.
- **Mean Squared Error (MSE):** 2.4839. The average prediction error is approximately 1.57 TFP points (calculated as the square root of 2.48), representing about 1.5% of the TFP range, which indicates high prediction precision.

Table 2: Summary of Productivity Linear Regression Models

Metric	Slope (Coefficient)	Intercept	R-squared (R^2)	MSE
Labour Productivity Index	0.5059	55.6741	0.8834	24.9044
Total Factor Productivity	0.2496	78.1768	0.9464	2.4839

This table quantifies the predictive relationships, showing the magnitude of impact (slope) and the model's explanatory power and accuracy (R^2, MSE). It directly supports the interpretation of how internet usage influences productivity.

3.3. Discussion

The consistent findings from both Pearson correlation and linear regression models strongly suggest a robust positive relationship between internet adoption and productivity metrics (Total Factor Productivity and Labour Productivity) in the USA.

The exceptionally high correlation coefficients (r > 0.96) and R-squared values (LPI: 0.8834, TFP: 0.9464) indicate that internet penetration is a dominant factor associated with productivity gains.

The high R-squared values and strong positive correlations across both TFP and Labour Productivity are particularly striking. TFP often reflects technological advancements and efficiency improvements at a systemic level, while Labour Productivity reflects the output per worker. The strong connection observed suggests that the internet is not merely a tool but a foundational infrastructure that enables the adoption and effective utilization of other productivity-enhancing technologies and processes. This implies that investments in internet infrastructure and digital literacy could yield significant economic returns by fostering a more productive workforce and efficient business operations. The evidence supports the idea that internet access accelerates the adoption of productivity-enhancing tools and acts as a powerful policy lever for long-term productivity gains.

This observation also provides an interesting perspective on the historical "productivity paradox." In the early stages of the information technology revolution, there was a period where substantial investments in IT did not immediately translate into economy-wide productivity gains, leading to questions about the true economic impact of these technologies. The strong correlations observed in this study, particularly with TFP, suggest that the internet's impact has matured and become clearly measurable in recent decades, potentially offering a resolution to this paradox. While the consistency and strength of the relationship warrant further investigation into the specific mechanisms through which these gains are realized (e.g., increased access to information, automation, improved communication, or new business models enabled by the internet), the current findings lay a strong foundation for such inquiries.

While the statistical evidence is strong, the observed trends could be influenced by confounding variables, such as broader technological advancements, research and development investments, or shifts in economic structures that co-vary with internet adoption.

4. Analysis & Findings: Internet vs. Education

4.1. Hypothesis Testing

Both Pearson correlation and ANOVA tests provided statistically significant evidence of a relationship between internet usage and academic performance.

Pearson Correlation Results

- **Internet vs SAT Math:** A strong positive correlation was observed (r = 0.7746, p-value = 0.0000).
- **Internet vs SAT Reading/Writing:** A moderate positive correlation was found (r = 0.3894, p-value = 0.0228).
- **Internet vs 13-year-old Score:** A strong positive correlation was noted (r = 0.7368, p-value = 0.0000).
- **Internet vs 9-year-old Score:** A very strong positive correlation was identified (r = 0.7871, p-value = 0.0000).

The interpretation of these results indicates that all p-values were statistically significant ($p < 0.05$), leading to the rejection of the null hypothesis. This consistently points to a positive correlation between internet usage and academic performance metrics, with particularly strong relationships observed in mathematics and for younger student scores.

ANOVA Test Results

The ANOVA analysis further confirmed these relationships by examining differences in academic performance across internet usage quartiles.

- **Math:** F-statistic = 13.76, p-value = 8.02e-06, indicating a significant difference.
- **Critical Reading and writing:** F-statistic = 14.45, p-value = 5.28e-06, indicating a significant difference.
- **Score (13-year-old):** F-statistic = 18.53, p-value = 5.45e-07, indicating a significant difference.
- **Score (9-year-old):** F-statistic = 28.78, p-value = 5.79e-09, indicating a significant difference.

The interpretation of these ANOVA results consistently showed statistically significant differences in academic performance across different internet usage quartiles for all metrics. Higher F-statistics indicate stronger group differences, further supporting the influence of internet usage levels on academic outcomes.

Table 3: Summary of Education Hypothesis Tests

Metric	Test Type	Pearson (r)	p-value (Pearson)	F-statistic (ANOVA)	p-value (ANOVA)
SAT Math	Pearson /	0.7746	0.0000	13.76	8.02e-06

	ANOVA				
SAT Critical Reading/Writing	Pearson / ANOVA	0.3894	0.0228	14.45	5.28e-06
13-Year-Old Scores	Pearson / ANOVA	0.7368	0.0000	18.53	5.45e-07
9-Year-Old Scores	Pearson / ANOVA	0.7871	0.0000	28.78	5.79e-09

This table provides a concise summary of the statistical evidence for the relationship between internet usage and academic performance, highlighting the varying strengths of correlation across different subjects and age groups.

4.2. Machine Learning Models

Linear regression models were employed to quantify the specific impact of internet usage on various academic performance metrics, assessing the magnitude of change, explanatory power, and prediction accuracy.

Linear Regression Results for Academic Performance

- **Math Performance:**
  - **Slope:** 0.1884. This indicates that for every 1% increase in Internet users, the Math score is predicted to increase by 0.188 points. A 10% increase in Internet users is associated with a 1.9-point gain in Math scores (e.g., from 503 to 505).
  - **R^2:** 0.6000. This represents the strongest relationship observed among the academic metrics, meaning that Internet usage explains 60% of the variance in Math scores.
  - **RMSE:** ±4.89 points.
  - **Cross-Validated R^2:** 0.4413 ± 0.4358, suggesting reasonable generalization of the model.
- **Critical Reading/Writing:**
  - **Slope:** 0.1653. For every 1% increase in Internet users, the Critical Reading/Writing score is predicted to increase by 0.165 points.
  - **R^2:** 0.1516. This represents the weakest link among the academic metrics, indicating that Internet usage explains only 15% of the variance in Critical Reading/Writing scores.

    - **RMSE:** ±12.44 points, which is a relatively large prediction error.
    - **Cross-Validated R²:** 0.0906 ± 0.2321, indicating lower stability and generalizability of this model.
- **13-Year-Old Scores:**
    - **Slope:** 0.1105. For every 1% increase in Internet users, the 13-year-old scores are predicted to increase by 0.111 points.
    - **R²:** 0.5429. Internet usage explains 54.3% of the variance in 13-year-old scores.
    - **RMSE:** ±3.23 points.
    - **Cross-Validated R²:** 0.1583 ± 0.3089.
- **9-Year-Old Scores:**
    - **Slope:** 0.1452. This slope is stronger than that observed for 13-year-olds.
    - **R²:** 0.6196. This represents the strongest R² among all academic metrics, indicating that Internet usage explains 62% of the variance in 9-year-old scores.
    - **RMSE:** ±3.62 points.
    - **Cross-Validated R²:** 0.2068 ± 0.1693.

Table 4: Summary of Education Linear Regression Models

Test Category	Slope	Intercept	R²	MAE	RMSE	CV R² (Stability)
Math	0.1884	503.2055	0.6000	3.7467	4.8948	0.4413 ± 0.4358
Critical Reading/ Writing	0.1653	496.5211	0.1516	9.8101	12.4387	0.0906 ± 0.2321
13-Year-Old Scores	0.1105	272.4708	0.5429	2.3428	3.2260	0.1583 ± 0.3089
9-Year-Old Scores	0.1452	229.1724	0.6196	2.8724	3.6210	0.2068 ± 0.1693

This table provides detailed performance metrics for each academic model, allowing for a quantitative comparison of internet's predictive power across different educational outcomes and age groups. The inclusion of MAE and RMSE provides insight into prediction accuracy.

4.3. Discussion

The analysis of internet usage and academic performance reveals a nuanced relationship. Overall, the findings suggest a positive correlation between increased internet adoption and improved academic outcomes, particularly in mathematics and for younger students. The linear regression models indicate that internet access moderately supports math skills and elementary education, implying that digital tools may be instrumental in closing learning gaps in early STEM (Science, Technology, Engineering, and Mathematics) fields.

A notable observation is the differential impact across age groups: 9-year-olds appear to benefit more from internet access than 13-year-olds, as evidenced by a stronger slope (0.145 vs. 0.111) and a higher R^2 (0.62 vs. 0.54) for the younger cohort. This suggests that internet access may have a greater impact during foundational learning stages, potentially by providing early exposure to educational content and interactive learning tools.

Conversely, a significant area of concern is the negligible impact observed on language arts, specifically critical reading and writing skills. The linear regression model for Critical Reading/Writing showed the weakest relationship, with internet usage explaining only 15% of the variance in scores, and a large prediction error. This suggests that the development of these skills might depend more heavily on traditional, offline practices and interactions, or that the nature of internet engagement for these skills is less conducive to improvement. Furthermore, the benefits of internet access for adolescents (13-year-olds) appear to plateau faster compared to younger children, indicating that the positive effects may diminish as students progress to higher educational stages.

5. Analysis & Findings: Internet vs. Health

5.1. Hypothesis Testing

Pearson correlation tests were conducted to assess the linear relationship between internet adoption and various health metrics.

Pearson Correlation Results for DALYS (Disability-Adjusted Life Years)

- **Anxiety disorders:** A very strong positive correlation was found with a Pearson correlation coefficient (r) of 0.8049 and a p-value of 2.80×10−8.
- **Depressive disorders:** Showed a very strong positive correlation with an r-value

of 0.9135 and a p-value of 3.04×10−13.

- **Mental disorders (aggregate):** Exhibited the strongest positive correlation with an r-value of 0.9394 and a p-value of 2.00×10−15.
- **Substance use disorders:** Had a very strong positive correlation with an r-value of 0.8415 and a p-value of 1.61×10−9.

The statistical interpretation of these DALYs results indicates that all p-values are significantly less than 0.001, providing overwhelming evidence to reject the null hypothesis (which states no linear correlation). This signifies an astronomically low probability of observing these correlations by chance. All correlations (r > 0.8) demonstrate very strong positive relationships, meaning that as internet adoption increases, DALY rates for these disorders tend to increase proportionally. The strongest effect was observed for aggregate mental disorders.

Pearson Correlation Results for Deaths

- **Substance use disorders:** A very strong positive correlation was observed with an r-value of 0.8744 and a p-value of 6.22×10−11.
- **Mental disorders (aggregate):** Showed a very strong positive correlation with an r-value of 0.9359 and a p-value of 4.00×10−15.

Statistically, the extremely low p-values for deaths (e.g., ≤6.22×10−11) indicate that these correlations are not due to random chance, leading to the rejection of the null hypothesis. For mental disorder deaths, the probability of observing this relationship if the null hypothesis were true is less than 1 in 10 billion. The r-value of 0.87 for substance use deaths suggests an 87% positive covariance between internet adoption and fatal outcomes. For mental disorder deaths, the r-value of 0.94 indicates a nearly perfect linear relationship, with 94% of the variance in deaths explained by internet adoption in this dataset.

Pearson Correlation Results for Obesity

- **Obesity Mean Count:** An extremely strong positive correlation was found with a Pearson correlation coefficient (r) of 0.9688 and a p-value of 2.44×10−20.

The p-value of 2.44×10−20 provides overwhelming evidence to reject the null hypothesis, indicating that the probability of this correlation occurring by chance is less than 1 in 100 quintillion. The correlation coefficient of 0.97 signifies an almost perfect positive linear relationship, suggesting that 97% of obesity rate variance is explained by internet adoption trends.

Table 5: Summary of Health Pearson Correlation Tests

Disorder Category	Pearson (r)	p-value	Strength
DALYS			
Anxiety disorders	0.8049	2.80×10−8	Very Strong
Depressive disorders	0.9135	3.04×10−13	Very Strong
Mental disorders (agg.)	0.9394	2.00×10−15	Very Strong
Substance use disorders	0.8415	1.61×10−9	Very Strong
Deaths			
Substance use disorders	0.8744	6.22×10−11	Very Strong
Mental disorders (agg.)	0.9359	4.00×10−15	Very Strong
Obesity			
Obesity Mean Count	0.9688	2.44×10−20	Extremely Strong

This table provides a concise summary of the statistical evidence for the relationship between internet usage and health outcomes, highlighting the varying strengths of correlation across different health categories.

5.2. Machine Learning Models

Linear regression models were applied to quantify the precise impact of internet usage on DALYs, deaths, and obesity prevalence, providing insights into the magnitude of these relationships and the models' predictive capabilities.

Linear Regression Results for DALYS

- **Anxiety Disorders:** Slope = 7,193.46; R^2 = 0.648; MSE = 2.72×1010. This indicates a moderate positive link, where DALYs increase with internet use.

- **Depressive Disorders:** Slope = 10,614.41; R^2 = 0.834; MSE = 2.16×1010. This shows a strong positive link.
- **All Mental Disorders:** Slope = 24,125.28; R^2 = 0.883; MSE = 7.50×1010. This represents a very strong link.
- **Other Mental Disorders:** Slope = 1,381.80; R^2 = 0.905; MSE = 1.95×108. This category shows the strongest explanatory power from internet usage.
- **Substance Use Disorders:** Slope = 50,399.94; R^2 = 0.708; MSE = 1.01×1012. This indicates a large effect, but with high variance in predictions.

All mental health disorders show positive slopes, indicating rising DALYs with increased Internet use. Substance Use Disorders exhibit the strongest absolute effect, with an increase of approximately 50,400 DALYs for every 1% increase in Internet penetration. "All Mental Disorders" also show a significant effect, with an increase of about 24,125 DALYs per 1% Internet increase. "Other Mental Disorders" has the best model fit with an R^2 of 0.905, meaning Internet penetration explains over 90% of the variance in DALYs for this category. Substance Use Disorders have the highest uncertainty in predictions, with an MSE of 1.01×1012.

Linear Regression Results for Deaths

- **Mental Disorders:** Slope = 0.189; R^2 = 0.876; MSE = 4.92. This indicates a strong positive relationship, where each 1% increase in Internet Users is associated with a 0.189-unit increase in mental disorder deaths.
- **Substance Use Disorders:** Slope = 657.29; R^2 = 0.764; MSE = 1.29×108. This shows a very large effect size, with each 1% Internet increase linked to approximately 657 more substance-use deaths.

The high R^2 of 0.876 for mental disorder deaths suggests that 87.6% of the variance is explained by Internet penetration. For substance use disorders, the dramatically larger effect size indicates a strong association, though the R^2 of 0.764 suggests other factors also contribute.

Linear Regression Results for Obesity

- **Slope:** 356,997.45. This indicates that each 1% Internet increase is associated with approximately 357,000 more obese individuals.
- **Intercept:** 18,422,630. This represents the theoretical baseline obesity count at 0% Internet.
- **R^2:** 0.939. This is a near-perfect fit, meaning 93.9% of obesity variance is explained by Internet penetration.

- **MSE:** 8.28×1012. The typical prediction error is approximately ±2.88 million people.

The staggering slope for obesity suggests a strong correlation: a 10% increase in Internet users correlates with 3.57 million more obese individuals, implying a significant impact of digital lifestyles on weight management. The near-perfect R^2 suggests internet adoption explains almost all variability in obesity trends within the dataset.

Table 6: Summary of Health Linear Regression Models

Disorder Category	**Slope (Coefficient)**	**Intercept**	**R-squared (R^2)**	**MSE**
DALYS				
Anxiety Disorders	7,193.46	1,753,718	0.648	2.72×1010
Depressive Disorders	10,614.41	1,957,840	0.834	2.16×1010
All Mental Disorders	24,125.28	6,129,229	0.883	7.50×1010
Other Mental Disorders	1,381.80	372,679	0.905	1.95×108
Substance Use Disorders	50,399.94	1,172,487	0.708	1.01×1012
Deaths				
Mental Disorders	0.189	24.21	0.876	4.92
Substance Use Disorders	657.29	4,311.71	0.764	1.29×108
Obesity				
Mean Count	356,997.45	18,422,630	0.939	8.28×1012

This table provides detailed performance metrics for each health model, allowing for a quantitative comparison of internet's predictive power across different health outcomes.

5.3. Discussion

The analysis reveals strong and consistent positive correlations between internet usage and various adverse health outcomes in the USA, including mental health DALYs, deaths related to mental health and substance use, and obesity prevalence.

For **mental health DALYs**, all analyzed disorders (Anxiety, Depressive, All Mental, Other Mental, and Substance Use) show increasing DALYs with rising internet usage. Substance Use Disorders exhibit the strongest absolute effect, with an increase of approximately 50,400 DALYs for every 1% increase in Internet use, making it the most sensitive category. The high R-squared values, particularly for "Other Mental Disorders" (0.905), suggest that internet penetration explains a substantial portion of the variance in DALYs for these conditions. This suggests that higher internet penetration is associated with greater mental health burdens.

Regarding **deaths related to mental health**, a 1% increase in internet users is associated with a 0.189-unit increase in mental disorder deaths, with internet penetration explaining 87.6% of the variance. For **substance use disorder deaths**, the effect size is dramatically larger, with each 1% internet increase linked to approximately 657 more deaths. This substantial slope for substance use disorders demands urgent investigation into digital risk factors.

In the context of **obesity prevalence**, the analysis indicates a staggering correlation: each 1% increase in internet users correlates with approximately 357,000 more obese individuals. This suggests that digital lifestyles may significantly impact weight management, with internet adoption explaining almost all variability in obesity trends within the dataset (R^2 = 0.939).

The observed relationships are complex and could be influenced by a multitude of interconnected factors. For instance, the increase in mental health burdens might reflect increased digital stressors such as social media pressures and cyberbullying, or it could be due to improved disorder diagnosis and reporting in more digitally advanced societies. For substance use, online drug markets and digital communities that normalize risky behavior could be contributing factors. In the case of obesity, increased screen time displacing physical activity, the rise of food delivery apps, targeted advertising of unhealthy foods, and sleep disruption due to 24/7 connectivity are all plausible co-occurring factors. Other broader societal changes, such as economic shifts or urbanization, may also confound these relationships, driving both internet adoption and health outcomes concurrently.

6. Conclusion & Insights

6.1. Summary of Key Findings

In the domain of **Productivity**, the study found exceptionally strong positive correlations between internet usage and both Total Factor Productivity (TFP) and Labour Productivity. Both Pearson correlation coefficients were above 0.96, and linear regression models showed R-squared values of 0.9464 for TFP and 0.8834 for Labour Productivity. This suggests that internet penetration is a dominant factor associated with significant productivity gains, potentially acting as a foundational infrastructure that enables broader technological advancements and efficiencies.

For **Education**, the analysis indicates a positive correlation between internet usage and academic performance, particularly in mathematics and for younger students (9-year-olds). Pearson correlation coefficients for Math and 9-year-old scores were strong (r=0.77 and r=0.79, respectively), and linear regression models showed R-squared values of 0.600 for Math and 0.620 for 9-year-old scores. This suggests that internet access moderately supports math skills and elementary education, potentially aiding early STEM learning. However, the impact on critical reading and writing was notably weaker (R^2=0.152), and benefits for adolescents (13-year-olds) appeared to plateau faster than for younger children.

In the **Health** domain, the findings present a concerning picture, with strong positive correlations between internet usage and various adverse health outcomes. Pearson correlation coefficients for mental health DALYs ranged from 0.80 to 0.94, and for deaths, they were above 0.87. Obesity prevalence showed an extremely strong correlation (r=0.97). Linear regression models further quantified these relationships, indicating that as internet usage increases, DALYs, deaths from mental health and substance use disorders, and obesity rates tend to rise proportionally. For instance, a 1% increase in internet users correlated with approximately 357,000 more obese individuals.

6.2. Answering Research Questions

Based on the findings, the research questions posed in the introduction can be addressed as follows:

- How has the rise in internet usage impacted education performance?

The rise in internet usage has shown a positive association with education performance, particularly in mathematics and for younger students (9-year-olds), where strong correlations and explanatory power were observed. This suggests that digital tools may support foundational learning. However, the impact on critical reading and writing skills appears to be less pronounced, and the benefits for adolescents may not be as sustained as for younger children.

- Is there a relationship between internet usage and mental health?

  There is a strong positive relationship between internet usage and mental health challenges. Increased internet adoption correlates significantly with higher rates of mental health DALYs (for anxiety, depression, general mental disorders, and substance use disorders) and an increase in deaths related to mental health and substance use. This indicates that as internet penetration grows, so do various indicators of mental health burden.
- How does internet usage affect productivity?

  Internet usage strongly affects productivity positively. The analysis revealed near-perfect positive correlations and high explanatory power (R-squared values above 0.88) between internet adoption and both Total Factor Productivity and Labour Productivity. This suggests that the internet serves as a significant catalyst for economic efficiency and output per worker.

6.3. Broader Implications

The overall picture emerging from this study is one of a profound and complex digital transformation shaping life in the USA. The internet appears to be a powerful engine for economic growth and educational advancement in certain areas, particularly in fostering productivity and supporting early-stage mathematical learning. This suggests that continued investment in digital infrastructure and literacy could yield substantial economic returns and educational benefits. The high explanatory power of internet usage on productivity metrics indicates that the internet has become a foundational element of modern economic activity, enabling widespread efficiency gains and potentially resolving historical questions about technology's impact on productivity.

However, this progress is accompanied by significant public health concerns. The strong positive correlations between internet usage and rising rates of mental health disorders, substance use-related issues, and obesity highlight a critical societal challenge. This suggests that while the internet offers immense benefits, its pervasive nature may also contribute to increased digital stressors, sedentary lifestyles, and potentially unhealthy digital environments. The observed relationships compel a deeper examination of the mechanisms through which internet usage influences these health outcomes, considering factors beyond simple access, such as content consumption patterns, online social dynamics, and digital well-being practices.

The findings underscore the dual nature of digital transformation. While the internet undeniably drives progress in areas like productivity and certain aspects of education, its widespread adoption also appears to coincide with, and potentially contribute to, growing public health burdens. This necessitates a balanced perspective, acknowledging the internet's role as both a driver of progress and a potential source of societal challenges. Understanding these complex interconnections is crucial for developing holistic strategies that maximize the internet's benefits while mitigating its adverse effects, ultimately striving for a more balanced and sustainable digital future for the USA.

7. References

Note: All the data mentioned below is available in my GitHub's ‘Data’ folder.

- World Bank. Internet Users (%). (Data accessed via Excel files: internet data.xlsx)
- Bureau of Labor Statistics. Total Factor Productivity and Labour Productivity. (Data accessed via Excel file: total-factor-productivity-major-sectors-historical.xlsx)
- National Center for Education Statistics (NCES) / Department of Education. SAT Scores. (Data accessed via Excel file: sat scores.xlsx)
- National Center for Education Statistics (NCES) / Department of Education. Average Mathematics Test Scores of 9-year-old and 13-year-old students. (Data accessed via Excel file: edu data.xlsx)
- World Health Organization (WHO) / Centers for Disease Control (CDC). Mental Health Data (DALYs & Deaths). (Data accessed via Excel file: mental health data.xlsx)
- World Health Organization (WHO) / Centers for Disease Control (CDC). Obesity Data. (Data accessed via Excel file: Obesity Data.xlsx)