Connectivity, Education and Prosperity - A Data-Driven Exploration of Internet Access, Education Level, and Household Income Dynamics in the United States

Ana Carolina Kosai Fischer   
*PostGrad Diploma in Science in Data Analytics, Natonal College of Ireland*Dublin, Ireland  
x23167866@student.ncirl.ie

Patrick Connolly   
*PostGrad Diploma in Science in Data Analytics, Natonal College of Ireland*Dublin, Ireland  
x18202314@student.ncirl.ie

Laleh Kian   
*PostGrad Diploma in Science in Data Analytics, Natonal College of Ireland*Dublin, Ireland  
x20147350@student.ncirl.ie

*Abstract*—The ability to finish higher education is known to increase a person’s opportunities to find gainful employment which in turn gives them better prospects for their own future and the societies within which they live. The digital divide results in inequalities where households without an internet connection find their employment and further learning opportunities restricted. In the USA, a sizable number of households still exist with no internet connection and their number varies from county to county. This report seeks to investigate the correlation between the percentage of households with no internet access, the proportion of the population over the age of 25 who attain an educational qualification of a High School Diploma or higher, and the income level across the states. To this end, data from the U.S. Census Bureau and the National Environmental Public Health Tracking Network was used. Our approach involves programmatically retrieving, preprocessing, and transforming these datasets using Python. Both Python and R were employed for in-depth analysis and visualization, facilitating the identification of patterns and trends. The results showed the partial mediation effect of income on the relationship between the percentage of households with no internet access and the percentage of the population older than 25 years who had attained an educational qualification of a High School Diploma or higher.

Keywords—education attainment, internet availability, mediation effect, socio-economic influence, household income.

# Introduction

Insufficient financial resources result in a reduced or even non-existent access to the internet. The absence of internet access further lowers the chances of achieving excellence in higher-level education, limiting a person's capacity to secure high-quality employment [1]. This reduction in employment opportunities subsequently contributes to lower household incomes, perpetuating a cyclical pattern where financial constraints delay both educational and economic advancement.

In the United States, a persistent digital divide continues to restrict internet access for a considerable portion of the population [2]. This gap in connectivity restricts individuals from being fully engaged in the digital age, limiting their access to educational resources, employment opportunities, and essential services.

Access to the internet is a crucial factor in determining success in the educational environment [3]. Especially in higher-level courses where digital resources and online collaboration are essential to the learning experience [4]. With that in mind, the digital divide brings challenges that extend beyond technological disparities, affecting the educational trajectory of those who are overly impacted.

Furthermore, the connection between household income and internet access accentuates the density of this issue. Household income levels contribute significantly to the rate of internet adoption at home [5]. As a result, disparities in income contribute to variations in digital access, further aggravating educational inequalities. The understanding of these interconnected dynamics is crucial to formulate effective strategies to address the root causes of the digital divide.

In the contemporary landscape, where digital technologies permeate every aspect of society, the importance of understanding and addressing digital disparities cannot be overstated [3]. The impact of limited internet access extends beyond individual households to create a greater socio-economic pattern. Therefore, addressing the digital divide becomes essential in the promotion of equal opportunities, economic mobility, and social progress.

For this project we analysed, evaluated and also assessed the challenges that might be associated with the processing of large datasets. Our data was web-scraped, wrangled and cleaned before being data mined. The programming language, Python was used for most of the steps such as web-scraping, connecting to a database, cleaning, wrangling, analysing and visualizations. A contribution was made using the programming language R for some cleaning, wrangling and exploratory analysis. We used PostgreSQL as our relational database which was hosted on a virtual machine and we connected to the Oracle VM VirtualBox from our host operating system, using Python.

## Motivation

By investigating the relationship between internet access, educational attainment, and household income, the purpose of this project is to understand the diverse challenges faced by the digital divide in the United States. In line with the above discussions, this study seeks to address the following research questions:

1) Does household internet access contribute to higher educational qualifications?

2) Does the higher income link to higher education, and that the income variable mediates the relationship described in research question 1?

# Related Work

Previous research has extensively explored the digital divide and its impact on educational attainment. Studies have shown that limited access to the internet can intensify educational inequalities, particularly in higher-level courses [1]. Anderson and Kumar [2] noted that the digital divide persists, even as lower-income Americans make gains in technology adoption. Additionally, Warschauer [3] emphasized the importance of technology in social inclusion and called for a rethinking of the digital divide.

Another research paper published by the Milken Institute in 2017 found that having a reliable internet connection had a positive effect on educational outcomes but that this was conditional on outside variables such as family size, demographics and regional characteristics between urban and rural settings [5].

National statistics have been crucial in understanding the relationship between education and internet access. The National Center for Education Statistics [6] reported on education and internet access in U.S. households in 2019, shedding light on the disparities in access across different demographic groups.

Furthermore, the influence of internet usage on income inequality has been a subject of investigation. Tsaurai and Chimbo [1] assessed the correlation between internet usage and income inequality in transitional countries, providing insights into the complex dynamics.

The Pew Research Centre published an analysis of several surveys in 2010 that found that 93% of households making more than $75,000 per year had an internet connection of some type versus 85% for lower income households and that these higher income households were more likely to use the internet on any given day [7].

The current research aims to study the potential median effect of income on the relationship between the Internet accessibility and education attainment. Previous research and publications referred in above are focused on other research questions different to the income mediation effect.

# Methodology

# *Data Mining Methodology*

Data mining is the process of identifying useful and previously unknown patterns from large datasets with the aim of extracting useful information and knowledge from the data [8]. The purpose of this project is to identify and conduct a series of analysis on a collection of large datasets that are somehow related and that could elucidate the questions proposed. The data mining method known as Knowledge Discovery in Databases (KDD) was employed to understand the datasets and discover any patterns in the data (Fig. 1). The KDD process involves the iterative application of the following steps [9]:

* **Data Selection:** where the relevant data for the analysis is extracted from the data collection.

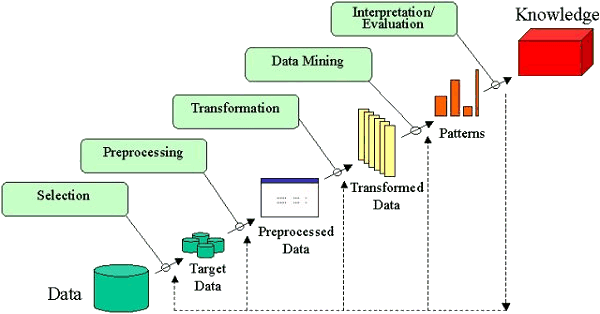


Fig. 1. KDD data mining process [9].

* **Pre-processing:** to remove noisy and irrelevant data, and handle missing data.
* **Transformation:** to transform data into a form suitable for data mining through data smoothing, aggregation and normalization.
* **Data Mining**: focuses on the analytical tools employed to discover trends and useful patterns from a dataset using techniques such as artificial intelligence and specialized algorithmic models.
* **Interpretation and Evaluation:** any patterns and insights that have been identified are categorized as bar graphs, pie charts, histograms etc., to establish the outcomes of the previous steps. This step also measures the effectiveness of the data model used.
* **Knowledge:** the results obtained are presented to the end user and are used to aid in the decision-making process. To achieve the data mining objectives proposed, the research followed the steps listed below:

**Step 1:** data was web-scraped and populated into a \*.csv file which was then downloaded.

**Step 2:** initial data wrangling and cleaning of the individual datasets was performed before the contents of the \*.csv files were uploaded to a PostgreSQL database.

**Step 3:** using SQL commands the data within the PostgreSQL database is queried and loaded into a Python data frame using Pandas.

**Step 4:** the queried datasets were analysed and visualizations were produced from these datasets using Python and R.

# *Data Sources and Descriptions*

As previously mentioned, four datasets were selected for this project. The datasets are based on the socioeconomic data from the United States of America and include significant community characteristics down to county level within individual states. A description of each dataset and their respective sources are outlined below:

* **Number of Households with No Internet Access** dataset [[link](https://ephtracking.cdc.gov/DataExplorer/?query=d3d59990-0c41-4e82-ba52-d49d02b91b0c)] dataset was retrieved from the U.S. Census Bureau and the American Community Survey (ACS). Web-scraping was employed to extract the raw data from the National Environmental Public Health Tracking Network and then the outcome was filtered using two key features, namely Counties and Years. The original dataset had 3143 entries, one for each county, and ten different community characteristics for the period from 2017 to 2021.
* **Socioeconomic Status** dataset [[link](https://ephtracking.cdc.gov/DataExplorer/?query=244fb3b2-fc69-483b-a06d-9de385fb8043)] was also extracted from the U.S. Census Bureau and the American Community Survey (ACS). The selected measure was the percentage of the population with age equal or greater than 25 years that holds a High School Diploma (or equivalent) or higher. Once again, a raw data entry was obtained for each county within each state over a period of time from 2009 to 2021.
* **Inflation-Adjusted Median Household Income in the past 12 months** dataset [[link](https://api.census.gov/data/2021/acs/acsse/variables.html)] was sourced using a U.S. Census Bureau API to fulfill the project requirements for a semi-structured dataset. Household related data is estimated and is collected from a sample of the population in the United States with varying margins of error.
* **Number Of People By Demographic Group** [[link](https://ephtracking.cdc.gov/DataExplorer/?query=244fb3b2-fc69-483b-a06d-9de385fb8043)] was a fourth dataset extracted from the U.S. Census Bureau and the American Community Survey (ACS) to weight the statistics analysis based on the population of each state. The weighting was intended to ensure that the data accurately represents the larger population and to help correct any imbalance of the underrepresented demographic groups, providing more accurate and liable results.

# *Data Pre-processing*

In each dataset, a thorough examination of data quality was performed, assessing attributes for their correct data types, identifying any missing values, ensuring the predictors were numeric values, and detecting any potential outliers within the data. Each individual dataset was cleaned and its contents used to populate a table within the PostgreSQL database which in turn allowed the tables to be queried and interrogated for data analysis while allowing for visualizations of the data to be presented.

The Number of Households with No Internet Access dataset was imported as a \*.csv file using a Pandas data frame. As shown in Fig. 2, the working data subset was created by filtering the period of 2017-2021. The last column was filled with “NaN” values and this was subsequently dropped. Also, all leading zeroes were removed.

The Socioeconomic Status dataset was also imported as a \*.csv file using a Pandas data frame, and it was grouped by the attribute of interest, namely ‘Percent of Population >=25 Years of Age with High School Diploma (or Equivalent) or Higher’. Initial cleansing of the dataset ensured that there were no missing values and that any attributes that offered little information value to the study, such as, “StateFIPS”, “CountyFIPS” and “Data Comment” were removed.

The third dataset, Median Household Income, was retrieved as a JSON file through an API URL available in the U.S. Census Bureau website. The import was done using the Python libraries such as Pandas, Requests, OS and Json. To extract the information necessary for comparison with the two other datasets, the variable “K201902\_001E” was sourced along with the Counties and States names. The dataset was subsequently pre-processed to split the County and State information into two different attributes, and to search for any missing values. The cleaned dataset was exported as \*.csv file using Pandas data frame for further analysis required to address the research questions.

Furthermore, the fourth dataset was also imported as a \*.csv file using a Pandas data frame. The attribute “Value” was transformed from an object data type into an integer before checking for missing values and outliers.

A screenshot of a table

Description automatically generated

Fig. 2. Attributes within Household with No Internet Access dataset.

# Results and Evaluation

# *Descriptive Results*

Figures 3, 4, and 5 below highlight key insights about the states in the USA, exploring the correlation between the percentage of households without internet access, the educational attainment of the population aged 25 and above, and the income levels across states..

Fig. 3 identifies the top 10 states in the USA that have the highest percentage of households without an internet connection. Mississippi stands out as the state with the highest percentage (30.4%) of households lacking internet access, with several other Southern states also featuring the rank.

Fig. 4 showcases the top 10 states with the lowest percentage of the population (25 years or older) holding a High School Diploma or higher. Massachusetts leads with only 8.2% of its population meeting the educational threshold for the research. Curiously, the other states with lower educational attainment are distributed across the North, West, and South regions of the USA.

Similarly, the Fig. 5 shows the top 10 states in the USA with the highest median income per household. The District of Columbia takes the lead with an impressive median income of $90,088 per year. Notably, many of the states with the highest median income are concentrated in the Northeast region of the USA.

A graph of a number of people

Description automatically generated

Fig. 3. Top 10 states with the highest percentage of households with no internet access.

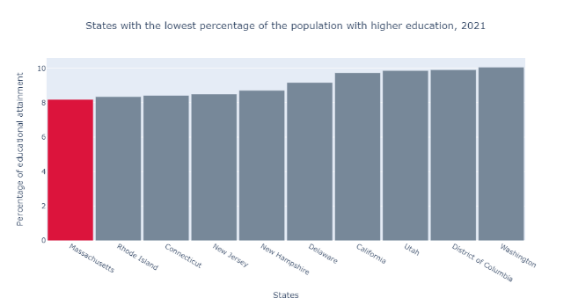


Fig. 4. Top 10 states with the lowest percentage of the population with a higher education diploma or equivalent.

A graph with numbers and text

Description automatically generated

Fig. 5. Top 10 states with the highest median income per household.

To visualize the variables of interest in this study, the boxplots of Education Attainment (%), No Internet Access (%), and Median of Household Annual Income ($), and state populations are presented in the Figures 6 to 8. As pictured below, there are no outliers observed in the boxplots. The outliers shown in the population boxplot (Fig. 9) indicate that the consequent analysis should consider this variable as a weight factor.

A graph with a box and a diagram

Description automatically generated with medium confidence

Fig. 6. Boxplot of the predictor “Educational attainment (%)”.

A graph with a blue rectangle and yellow dots

Description automatically generated

Fig. 7. Boxplot of the predictor “No internet access (%)”.

A graph with a blue square and yellow dots

Description automatically generated

Fig. 8. Boxplot of the predictor “Median annual income($)”.

A graph with numbers and a diagram

Description automatically generated with medium confidence

Fig. 9. Boxplot of weight factor “Population”.

# *Hypothesis*

The analysis explored the "mediation" effect of the variable "Median annual income" on the relationship between "No internet access" and "Educational attainment". This exploration involved the development of various Linear Regression models, where each one aimed at deciphering the dynamics of this mediation effect.

It is important to note that the interpretations provided below are subject to the assumptions of linear regression being met. Linear regression assumes a linear relationship between the independent and dependent variables, independence of observations, homoscedasticity, and normality of residuals [10].

In the initial regression model, we hypothesized the significance of the relationship between "No internet access" and "Educational attainment". The results unveiled a notable relationship (b = -0.476, t = -4.69, p < 0.001). Fig. 10 visually represents this connection through a scatter plot and the associated linear regression, where the dot sizes reflect the population of the corresponding states.

The negative coefficient of the first regression model indicates an inverse relationship between the percentage of households with no internet access and the educational attainment. In other words, as the percentage of households without internet access increases, the level of educational attainment tends to decrease. The t-value being significantly different from zero and the low p-value (< 0.001) suggest that this relationship is statistically significant.

Subsequently, the second regression model examined the relationship between "No internet access" and "Median annual income". The outcome highlighted yet another significant relationship (b = -2262.4, t = -9.912, p < 0.001). Fig. 11 complements this finding with a scatter plot and linear regression, employing dot sizes to signify the population of the respective states.

In this second model, the negative coefficient implies a negative relationship between the percentage of households with no internet access and the median annual income. As the percentage of households without internet access increases, the median annual income tends to decrease. The statistically significant t-value and low p-value (< 0.001) indicate that this relationship is also significant.

The next regression model analysed the significance of relationship between the variables “Educational attainment” and “Median annual income”. The outcome underscored a meaningful association (b = 9.65e-05, t = 2.30, p = 0.0255). Fig. 12 provides a visual representation of this relationship through a scatter plot and the accompanying linear regression, with dot sizes indicating the population of the respective states.

Differently from the first two models, the positive coefficient suggests a positive relationship between the educational attainment and median annual income. As the educational attainment increases, the median annual income tends to increase. The statistically significant t-value and low p-value (0.0255) indicate that this relationship is statistically significant as well.

In summary, the models described above investigated the relationship between “No internet access”, “Educational attainment” and “Median annual income” using three distinct simple linear regression models.

Finally, a thorough Multiple Linear Regression analysis was executed, introducing the variable "Median annual income" as a potential mediator to the existing linear regression model between "No internet access" and "Educational attainment". The outcome revealed an overall significant model (F(2, 48) = 14.13, p < 0.001), as represented in Fig. 13. Additionally, significance was observed for both variables, namely "No internet access" (b = -0.774, t = -4.56, p < 0.001) and "Median annual income" (b = -1.318e-04, t = -2.15, p = 0.0365).

A graph with black dots and red line

Description automatically generated

Fig. 10. Scatterplot of the variables “Education attainment (%)” vs. “No internet access (%)”.

A graph of black dots and red line

Description automatically generated

Fig. 11. Scatterplot of the variables “Median annual income ($)” vs. “No internet access (%)”.

A graph with black dots and red line

Description automatically generated

Fig. 12. Scatterplot of the variables “Educational attainment (%)” vs. “Median annual income ($)”.

The coefficient for "No internet access" indicates that, holding "Median annual income" constant, a one-unit increase in the percentage of households without internet access is associated with a decrease of 0.774 units in the dependent variable. The coefficient for "Median annual income" suggests that, holding "No internet access" constant, a one-unit increase in median annual income is associated with a decrease of 0.0001318 units in the dependent variable. The overall model is statistically significant (p < 0.001), indicating that the combination of these predictors significantly predicts the dependent variable.

Comparing the results of the Multiple Linear Regression model to the initial model indicated a change in the coefficient of the variable "No internet access" between the two models. This change implies that "Median annual income" acts as a "partial mediator" in the relationship between "No internet access" and "Educational attainment". It's crucial to note that, in this analytical context, if the primary predictor (No internet access) loses significance in the multiple linear regression model, the mediation effect is named as a "full effect". On the contrary, if the variable being tested for mediation lacks significance in the multiple linear regression model, it suggests no mediation effect.

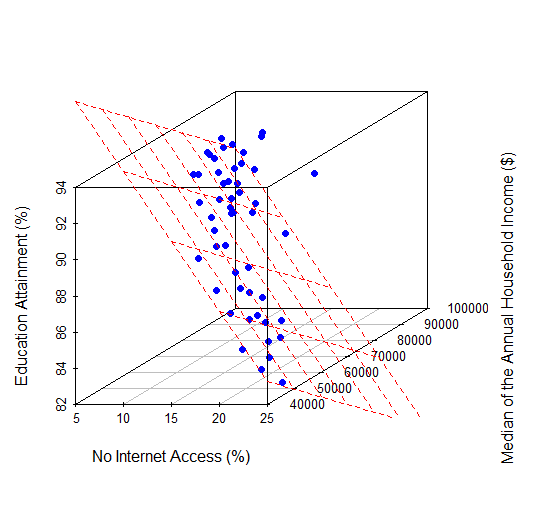


Fig. 13. 3-D Scatterplot of the variables within the multiple linear regression model.

# Conclusions and Future Work

The primary objective of this research was to investigate the relationship between internet access, educational attainment, and household income, with a specific focus on understanding whether internet access contributes to higher educational qualifications and whether higher income is linked to higher education, mediating the relationship between internet access and educational attainment.

The analysis revealed that internet access itself does not significantly contribute to the percentage of the population with educational attainment. In fact, the household income was identified as a mediator, suggesting that it plays a role in influencing the relationship between internet access and education attainment.

In comparison to previous studies, our findings provide a unique perspective on the dynamics between internet access, education, and income. While some studies have suggested a direct link between internet access and educational outcomes, our research suggests that household income may be a crucial factor in understanding this relationship. This deviation from existing literature underscores the complexity of socioeconomic interactions in shaping educational attainment.

Addressing the research questions, our study provides answers to both. First, household internet access alone does not seem to contribute significantly to higher educational qualifications. Second, the higher income indeed appears to be linked to higher education, and the income variable mediates the relationship described in the first research question. These insights offer a valuable contribution to the ongoing understanding of the dynamics of factors influencing educational outcomes.

As with any research, it is essential to acknowledge potential limitations. Deviations from linear regression assumptions and caution in inferring causation are critical considerations. Additionally, the sensitivity of coefficients to sample size highlights the importance of future studies with larger datasets to confirm and generalize these findings.

Future researches could extend the current investigation by conducting a more detailed analysis based on counties rather than states. Utilizing county-level data would provide a larger dataset, enabling a more thorough exploration of the relationships between household internet access, educational attainment, and income.

Furthermore, expanding the regression model to incorporate additional variables, such as gender, could offer a richer understanding of the socioeconomics influences on educational outcomes. Including a broader set of variables in the model would contribute to a more comprehensive understanding of the dynamics between socioeconomic factors and educational achievements.

In conclusion, the proposed future work aims to build upon the current study's foundation, refining and expanding the analysis to uncover more detailed insights into the intricate relationships between household internet access, educational attainment, and income.

##### References

1. K. Tsaurai and B. Chimbo, "Investigating the Influence of Internet Usage on Income Inequality in Transitional Countries", *EuroEconomica*, vol. 40, no. 2, pp. 113-128, 2021.
2. M. Anderson and M. Kumar, "Digital Divide Persists Even as Lower-Income Americans Make Gains in Tech Adoption", *Pew Research Center*, 2019.
3. M. Warschauer, "Technology and Social Inclusion: Rethinking the Digital Divide", *MIT Press*, 2003.
4. P. DiMaggio and E. Hargittai, "From the ‘Digital Divide’ to ‘Digital Inequality’: Studying Internet Use as Penetration Increases", *Princeton University Center for Arts and Cultural Policy Studies Working Paper Series*, 2001.
5. J. Lee, "Internet Usage Effect on Educational Attainment: Evidence of Benefits," Milken Institute, 2017. [Online]. Available: [https://milkeninstitute.org](https://milkeninstitute.org/sites/default/files/reports-pdf/Internet-Usage-and-Educational-Attainment-FINAL.pdf) [Accessed on: 07 December, 2023].
6. National Center for Education Statistics, "Education and Internet Access in U.S. Households: 2019", *U.S. Department of Education*, 2021.
7. J. Jansen, “Use of the internet in higher-income households,” Pew Research Center, 2010. [Online]. Available: [https://www.pewresearch.org](https://www.pewresearch.org/internet/2010/11/24/use-of-the-internet-in-higher-income-households/) [Accessed on: 07 December, 2023].
8. R. Kimmons and G. Veletsianos, "Public Internet Data Mining Methods in Instructional Design, Educational Technology, and Online Learning Research," in TechTrends, vol. 62, pp. 492-500, 2018. [Online]. Available: [https://link.springer.com/article/10.1007/s11528-018-0307-4](https://link.springer.com/article/10.1007/s11528-018-0307-4%20) [Accessed on: 07 December, 2023].
9. U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From Data Mining to Knowledge Discovery in Databases,” AI Magazine , vol. 17, no. 3, 1996. Available: [From Data Mining to Knowledge Discovery in Databases](https://www.google.com/search?q=From+Data+Mining+to+Knowledge+Discovery%3A+An+Overview&rlz=1C1CHBF_enIE923IE923&oq=From+Data+Mining+to+Knowledge+Discovery%3A+An+Overview&gs_lcrp=EgZjaHJvbWUyBggAEEUYOdIBBzk4NWowajeoAgCwAgA&sourceid=chrome&ie=UTF-8) [Accessed on: 07 December, 2023].
10. D. C. Montgomery, E. A. Peck, and G. G. Vining, "Introduction to Linear Regression Analysis," in Introduction to Linear Regression Analysis. John Wiley & Sons, 2012.
11. A. Gelman and J. Hill, Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, 2007.