Education and Covid 19 Analysis

Lindsey Hill

May 16, 2025

Introduction

For this project, I am analyzing the correlation between Covid-19 infection rates and educational attainment levels across counties in the United States. The goal is to determine whether a possible relationship exists between the level of education in a population and the spread of Covid-19 while considering other influencing factors such as population density. It is my theory that access to higher education provides people with better access to resources, like healthcare and the ability to work from home, that may prevent the spread of disease. This analysis can provide insights into potential social factors that impacted infection rates during the pandemic and suggest possible avenues to better prepare for the next public health crisis.

The key questions I am exploring are as follows:

1. How does the percentage of adults with a bachelor's degree or higher correlate with Covid-19 infection rates per capita at the county level?
2. Do areas with lower high school graduation rates show a stronger relationship with higher infection rates per capita?
3. Are there geographic trends where education level differences impact infection rates more significantly?

Hypothesis

I hypothesize that counties with a higher percentage of adults holding a bachelor's degree or higher may exhibit lower Covid-19 infection rates per capita, while counties with higher percentages of adults without a high school diploma may experience higher infection rates per capita. I propose that higher education levels correlate with greater health literacy, better healthcare access, and more remote work opportunities, which may reduce exposure risks. Conversely, lower education levels may be associated with in-person jobs and limited healthcare access, increasing exposure risks.

Data Sources

The primary data sources for this analysis include:

1. Covid-19 Case Data – The New York Times Covid-19 Data Repository on GitHub provides this data, which includes county-level Covid-19 case counts. The base tables contain the date of the case number report, county, state, FIPS, case count, and death count (The New York Times, 2021). The FIPS is a standardized number that shows geographic information that changes with granularity (US Census Bureau) (Dotsquare LLC). Data for New York City was excluded during data preparation due to its status as an extreme outlier which was skewing results and flattening data.
2. Educational Attainment Data by Age and Sex - ACS provides a Sex by Age by Educational Attainment for the Population 18 years and over from the 2020 5-year Estimate Subject Table. It includes geographic ID and geographic name, as well as estimates of population, sex, and educational attainment by age groups, and margins of error for each estimate (US Census Bureau). This data set organizes this data into 168 columns listed by county.
3. **Educational Attainment Data** – Obtained from the U.S. Census Bureau's American Community Survey (ACS) 2020 5-year Estimates Subject Tables, detailing county-level educational attainment statistics (US Census Bureau). This is the same educational information as above but organized differently and was ultimately not used.

Audience and Purpose

This report presents visualizations and interpretations developed during analysis exploring the correlation between educational attainment and Covid-19 infection rates across U.S. counties during the period of 2020 through 2021. The primary audience includes public health analysts, data scientists, and policymakers interested in the social determinants of public health outcomes. These visualizations make complex patterns in Covid-19 case data more accessible and actionable, especially for decision making around future public health crisis strategies. Specifically, these visualizations suggest expanding educational access could improve a future public health crisis outcome.

Prepare and Process Data

All data preparation, cleaning, analysis, and visualizations were generated or conducted in R using RStudio (Posit Team, 2019). To begin the analysis, I combined two main datasets: Covid-19 case data and county-level education data. I first filtered the Covid-19 dataset from The New York Times to include data from January 2020 through December 2021 and then calculated new monthly case rates per 100,000 people for each county. I sourced the education data from the American Community Survey (ACS) 5-year estimates for 2020; it provided educational attainment breakdowns of adults by county. I removed the education data for people aged 18 to 24 because it is unreasonable to assume an adult should have a bachelor's degree before 25. I recalculated the population after removing those younger than 25, and combined the different sexes, age groups, and appropriate education brackets. I recalculated the new population of adults without those aged 18 to 24 and calculated percentages for adults in each educational bracket, those with a bachelor's degree or higher and those without a high school diploma or equivalent. I standardized both data sets for date formatting, truncated the geographic ID known as FIPS in both to remove unnecessary granularity, and corrected any numbers that were formatted as text. Then, I combined the Covid-19 data sets for 2020 and 2021. I aggregated the cumulative case count to

pull just the new cases using the lag function in R and then adjusted cumulative and new case counts to show monthly instead of daily (Harris, 2024). Grouping by month allows any time related analysis to be less complex because of the granularity and seeks to remove some of the negative new case counts that occurred due to human error or reporting errors in the data. Then, I joined the education data using the matching FIPS information to the Covid-19 data and investigated and removed missing results for discrepancies. I then normalized additional calculations for cumulative Covid-19 cases, new Covid-19 case counts, and counts of adults in both education brackets by calculating the amount of each of these per 100,000 people in the county population. This allows the case information to not be skewed based on county population. I detected a major outlier, New York City, during exploratory analysis and investigated and removed it later.

Clean Data and Figure 1

The data cleaning was almost entirely focused on handling missing information. After joining the data, it was found that the County column of the Covid-19 data included three cities, none of which had a FIPS associated with them. These were found when searching for records with missing FIPS to make sure the join functioned as expected. These cities were Joplin, MO, Kansas City, MO, and New York City, New York. Investigation into the counties that made up Joplin, MO produced two counties, and a search showed these counties were mentioned in the data, so the decision was made to drop the Joplin, MO data in favor of the county specific data to prevent duplication (County of Joplin, Missouri). Investigation into the counties that make up New York City showed that none of the county specific data was included in the dataset (City of New York). Initially, I combined the education information for the 5 counties of New York City and joined that combined data with the Covid-19 data for the city. Later in visualizations, this caused New York City to show as an extreme outlier, flattening charts by a large margin when conducting exploratory analysis, so New York City was removed for the extreme inconsistencies pointing to data that was collected incorrectly or based on a different standard. Kansas City is made up of 14 counties and 11 of the 14 counties were represented in the Covid-19 data already so the decision was made to drop the Kansas City data in favor of the 11 counties that are already represented and more accurately joined with the education data (Legends of America). When searching for missing FIPS data, I also found a lot of data where the county name was listed as “Unknown.” This data was also dropped as there is no way to connect this Covid-19 data with educational data. The rest of the data looks to have cleanly joined with the education data.

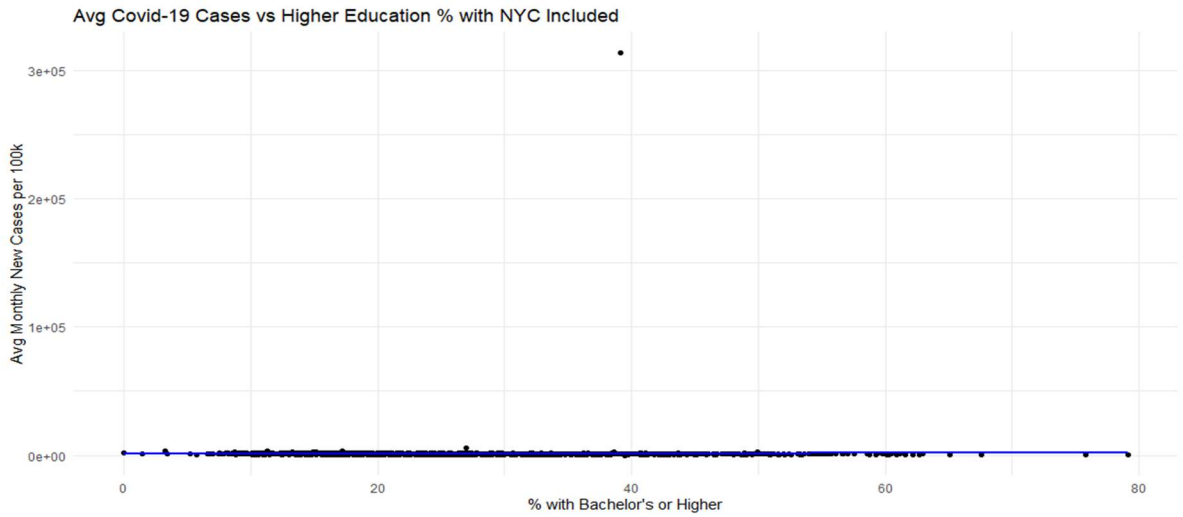

Figure 1- Shows outlier of New York City flattening data. Modeled using RStudio (Posit Team, 2019).

Exploratory Analysis

Exploratory Data Analysis (EDA) focused on visualizing and comparing Covid-19 infection rates to educational attainment. Several visualizations were created including line graphs comparing monthly new case rates in the top 10 and bottom 10 counties by education level and scatter plots comparing the average monthly new Covid-19 cases. The monthly new cases were compared with the percentage of adults with a bachelor's degree or higher, the monthly cases with the percentage of adults without a high school diploma, and county population size.

Initial findings suggested an inverse relationship between higher education and Covid-19 rates, and a possible correlation between lower educational attainment and higher case counts. However, results varied depending on how many counties were included (e.g., top 10 vs. top 100) and the small sample gave inconsistent results when compared to data about adults missing high school education , indicating the need to account for confounding variables like geography and population density. When testing average Covid-19 monthly new cases as compared to the population with a bachelor's degree or higher for all counties represented in the cleaned data, it was found that counties with higher education rates trended to have lower average monthly new Covid-19 cases per 100,000 people. The opposite was also found to be true, counties with higher numbers of adults without a high school education or equivalent trended to having higher average monthly new Covid-19 cases per 100,000 people. The average monthly new Covid-19 cases per 100,000 was then tested against the population of each county and I found it very interesting that the average new cases for every 100,000 decreased when the county contained more population, suggesting education levels of the population affect the data more than population size. However, it should be noted that population density was not measured and that may prove to be more significant.

Incorporate Modeling & Algorithms

To quantify observed relationships, linear regression models were applied using R's lm() function (Porras, 2024). This produced a statistically significant negative coefficient suggesting that for each percentage point increase in bachelor's degree attainment, the average number of monthly new Covid-19 cases per 100,000 decreased. Specifically, on the Avg Covid 19 Cases Vs Higher Education %, the coefficient for the slope was -8.27 and the p value showing it is significant is 2e-16. A similar model was run for the percentage of the population without a high school diploma. For the chart showing the average new cases against the percentage of adults without a high school education, the coefficient for the slope was 3.6994 and the p value showing it is significant is 2.3e-5. Regression models were used to estimate trends, and residuals were analyzed to assess fit.

Visualization Rationale

Given the size and complexity of the datasets, which included county-level statistics for over 3,000 U.S. counties, visualizations enhance interpretability and uncover trends not immediately apparent in raw data alone. Scatter plots were chosen specifically to illustrate relationships between variables, and to show those relationships against the other factor of population size. A choropleth, like a heat map but with regions restricted by governing or other boundaries, was also used to summarize the large-scale geographical patterns and see what states showed a higher level of correlation between case numbers and education (Standard Co, 2018).

Figure 2: County Average Covid-19 Cases vs. Higher Education %

This scatter plot shows a clear negative trend: as the percentage of adults with a bachelor's degree or higher increases, the average monthly Covid-19 cases per 100,000 population decreases. This supports the hypothesis that higher education levels may lead to reduced exposure or better preventive behavior during the pandemic. Education likely provides indirect protection through better job flexibility like work-from-home options, greater health literacy, and easier access to medical resources.

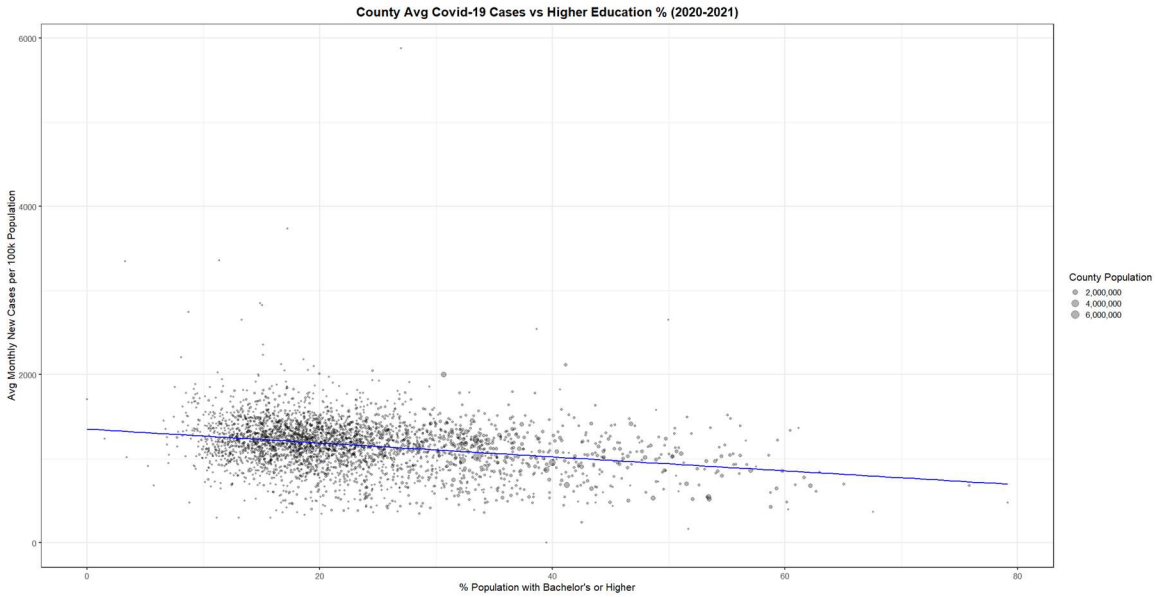

Figure 2- Modeled using RStudio (Posit Team, 2019).

Figure 3: County Average Covid-19 Cases vs. % Without High School Diploma

This chart shows a positive correlation: counties with a higher proportion of adults without a high school diploma or equivalent education tend to have higher average monthly case rates. Lower education levels may be associated with occupations requiring physical presence, reduced health knowledge, and less healthcare access; these are factors likely contributing to increased exposure and vulnerability.

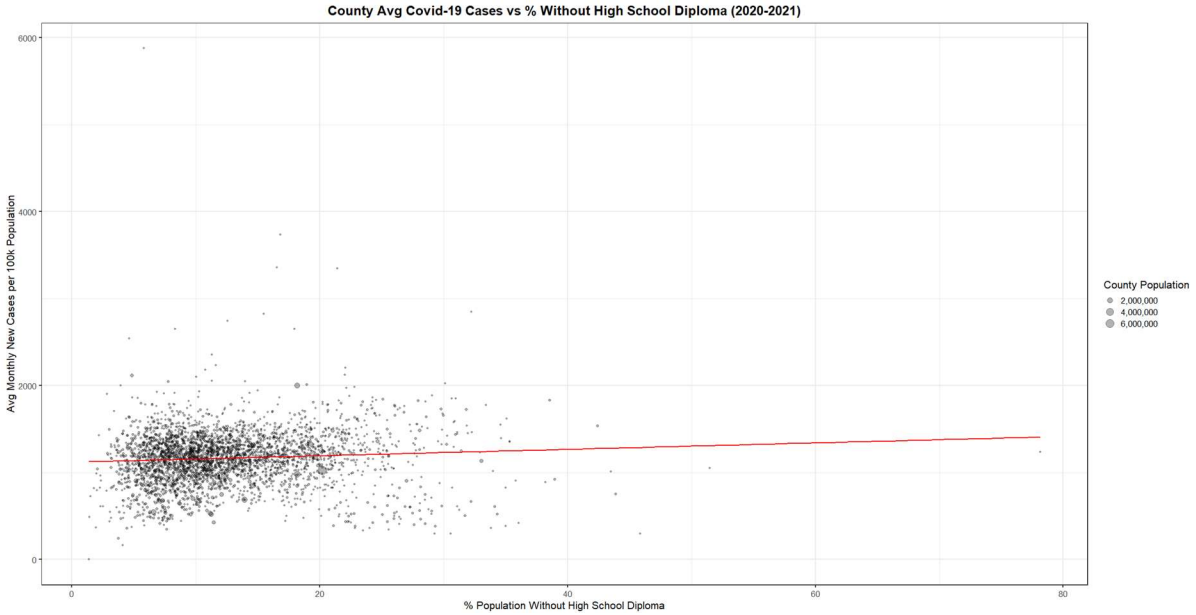

Figure 3- Modeled using RStudio (Posit Team, 2019).

Figure 4: Average Monthly Covid-19 Cases per 100k vs. County Population

To test whether county population size skewed infection rates, this visualization plots population against average monthly cases per 100,000 people. Surprisingly, the chart shows a decreasing trend, suggesting larger populations were associated with lower average case rates per capita. Population size is not a confounding variable that invalidates the education hypothesis. Instead, education level appears to be a more consistent predictor of infection rates than raw population count. It should be noted, however, that population density may prove to be a more important factor and is different from population size. However, population density information was not readily available in this project's scope and timeline but should be considered for future expansion.

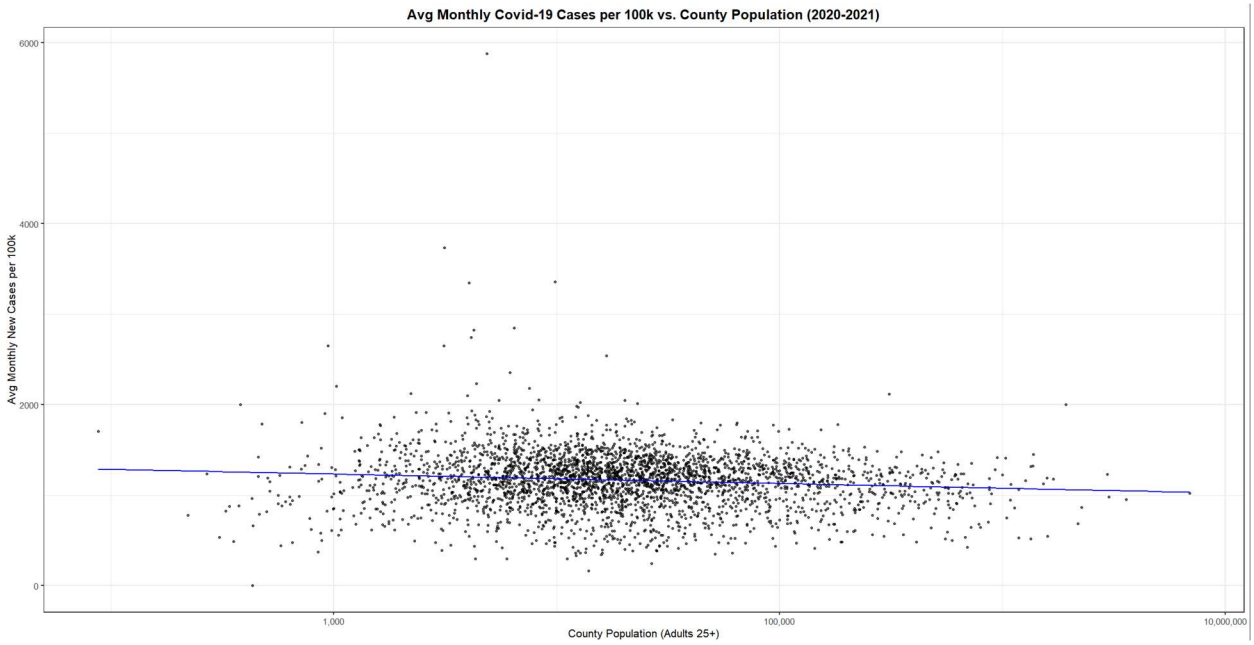

Figure 4- Modeled using RStudio (Posit Team, 2019).

Figure 5: Education Level vs Average Monthly Covid-19 Cases per 100k State-Level Choropleth

This choropleth map shows the strength of the correlation between education levels, both high and low attainment, and Covid-19 case rates across U.S. states. The abundance of medium to deep red shades, representing a strong positive correlation between low education and high case rates, provides strong visual support for the hypothesis. The near absence of blue, indicating a strong negative correlation with higher education, reinforces the conclusion. The consistency of this trend across states underscores the systemic nature of the education-infection relationship and highlights the potential for educational equity as a tool in public health planning.

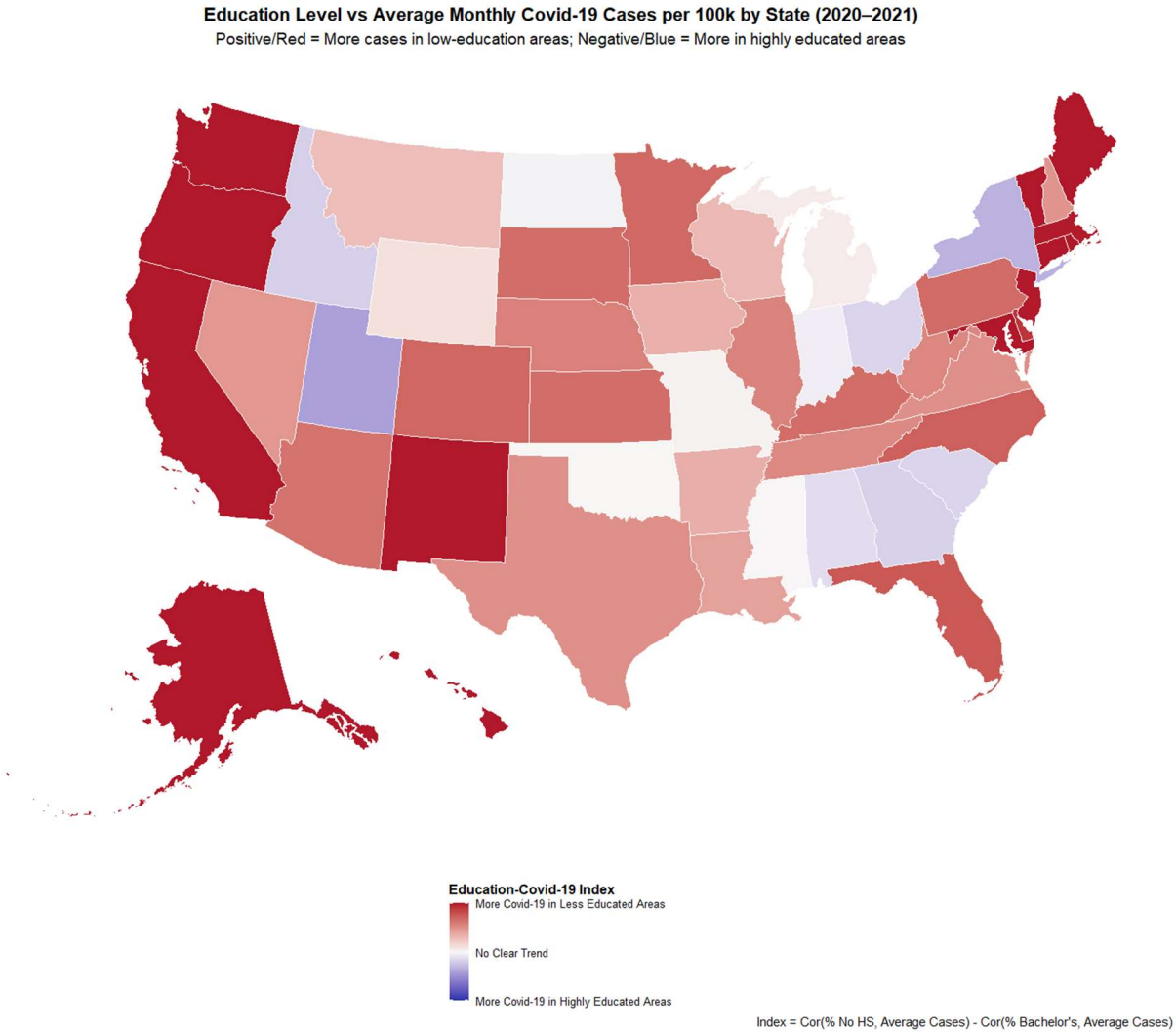

Figure 5- Modeled using RStudio (Posit Team, 2019).

Conclusion

The objective of this project was to determine whether there is a relationship between educational attainment and the spread of Covid-19 at the county level. The initial hypothesis stated that counties with higher percentages of adults holding at least a four-year college degree would experience lower Covid-19 case rates per 100,000 people. Through data preparation, normalization, time series analysis, and linear regression modeling, I found a consistent negative correlation between higher education levels and Covid-19 case rates across both 2020 and 2021. This supports the original hypothesis. Counties with higher rates of adults lacking a high school diploma consistently experienced higher case rates over time, while counties with greater proportions of adults with bachelor's degrees tended to have lower case rates. Notably, these trends were independent of total population size.

To quantify these relationships, linear regression models were applied. The model for bachelor's degree attainment produced a statistically significant negative coefficient. Conversely, the model for adults without a high school diploma showed a significant positive relationship with case rates. These results demonstrate a meaningful and statistically significant relationship between education levels and Covid-19 spread.

My overall approach involved acquiring county-level education and Covid-19 case data, aligning those datasets, performing exploratory statistical analysis, and visualizing trends. These results emphasize the importance of considering social and educational factors in pandemic preparedness. Educational attainment likely acts as a powerful indirect shield against disease spread, likely due to the cumulative effects of increased health literacy, job flexibility, and access to healthcare. Future studies can build on this analysis by incorporating additional variables such as population density, employment type, income levels, or healthcare facility access.

Although the project met its goal of supporting the initial hypothesis, it also revealed limitations in the dataset and analytical scope. Matching geographic identifiers, addressing missing values, and aligning demographic with temporal data required substantial preprocessing. Nonetheless, the project successfully showed that educational attainment is a statistically significant factor in explaining county-level variations in Covid-19 case rates.

Lessons Learned

This project offered several valuable insights into data analysis and the real-world challenges of working with public health and demographic datasets. First, I learned the importance of ensuring geographic consistency when combining data from different sources such as deciding whether to include independent cities which don't align neatly with county-based data. I also gained practical experience in handling missing values and deciding when to exclude or impute data based on relevance and integrity.

Working in R allowed me to become more comfortable with data manipulation and visualization tools, though I faced a learning curve with the regression formulas and some chart techniques. If I were to repeat this analysis, I would incorporate population density to improve explanatory power and possibly investigate vaccination rates and how that played into later case counts alongside education. I would also experiment with possibly building a predictive model.

Finally, I recognized the importance of clear visualizations and communication. While I generated several useful charts, more work could have gone into making them intuitive and visually appealing. In future projects, I would devote more time to designing visuals that convey insights more clearly and effectively to a broader audience.

Bibliography:

CDC. (n.d.). *CDC Covid Data tracker*. Centers for Disease Control and Prevention. https://Covid.cdc.gov/Covid-data-tracker/#datatracker-home

City of New York. (n.d.). *New York City counties*. New York City Counties · NYC311. https://portal.311.nyc.gov/article/?kanumber=KA-02877

Cmdlinetips. (2021, June 18). *9 tips to make better scatter plots with GGPLOT2 in R*. Python and R Tips. https://cmdlinetips.com/2019/11/9-tips-to-make-better-scatter-plots-with-ggplot2-in-r/

County of Joplin, Missouri. (n.d.). *History of Joplin: Joplin, mo - official website*. History of Joplin | Joplin, MO - Official Website. https://www.joplinmo.org/173/History-of-Joplin

Dotsquare LLC. (n.d.). *Look up the census geoid for addresses or properties*. Geocodio. https://www.geocod.io/lookup-census-geoid-for-an-address-or-property/

Harris, M. (2024, January 17). *The "lag" function in R*. Stats with R. https://www.statswithr.com/r-functions/the-lag-function-in-r

Posit Team. (2019). RStudio: Integrated Development Environment for R. computer software, Boston; Posit Software, PBC. http://www.posit.co/

Legends of America. (n.d.). *Kansas City Metropolitan Area*. https://www.legendsofamerica.com/kansas-city-metropolitan-area/

The New York Times. (2021). *Nytimes/COVID-19-data: A repository of data on coronavirus cases and deaths in the U.S.* GitHub. https://github.com/nytimes/Covid-19-data

Porras, E. M. (2024, July 29). *How to do linear regression in R*. DataCamp. https://www.datacamp.com/tutorial/linear-regression-R

Standard Co. (2018, November 27). *Heatmaps vs Choropleths*. Heatmaps vs choropleths. https://www.standardco.de/notes/heatmaps-vs-choropleths

US Census Bureau. (2023, May 1). *American National Standards Institute (ANSI), Federal Information Processing Series (FIPS), and other standardized geographic codes*. Census.gov. https://www.census.gov/library/reference/code-lists/ansi.html

US Census Bureau. (n.d.-a). *B15001: SEX BY AGE BY EDUCATIONAL ... - Census Bureau Table*. Census.gov. https://data.census.gov/table/ACSDT5Y2020.B15001?t=Educational+Attainment&g=010XX00US$0500000&y=2020

US Census Bureau. (n.d.-b). *S1501: Educational Attainment - Census Bureau table*. S1501 Educational Attainment. https://data.census.gov/table/ACSST1Y2022.S1501