**Final Paper:**

# Smoking Prevalence and Socio-demographics of U.S. Counties in 2022

Anna Kenney-Hynes
Carson Chapman
Daniel Woodford

SODA 308: Research Design

December 8, 2023

## Introduction

Smoking and vaping are known to have negative effects on the physical and mental health of those who become addicted to tobacco-related products. Smoking creates an increased risk for lung cancer, throat cancer, mouth cancer, decreased lung capacity, and more. Knowing this, the United States government has launched public campaigns and federal regulations intended to reduce smoking in the population. Researchers have also aimed to study aggravating factors that make a person more at risk to take up smoking or vaping. Understanding the demographic nuances in tobacco use is essential for the development of targeted educational initiatives and support programs. By tailoring interventions to specific populations, policymakers can address socio-economic disparities and cultural influences, fostering a more comprehensive and inclusive approach to public health that not only reduces smoking rates but also uplifts the overall well-being of the nation.

## Research Question

Is there a significant relationship between county-level socio-demographic factors such as

1. median annual household income,
2. educational attainment,
3. and the severity of the COVID-19 virus in 2020,

and county-level smoking prevalence among adults in the United States in 2022?

## Purpose

In wake of the COVID-19 pandemic, the purpose of this research is to investigate the relationship between smoking prevalence and various demographic factors at the county level within the United States, as well as the potential impact of coronavirus severity in a community on smoking prevalence. By using county level demographics, this study seeks to provide valuable insights for public health development in a diverse and thorough manner. Using counties allows for greater exploration in local variation, such as demographic characteristics, access to healthcare, cultural norms, socioeconomic conditions, and the presence of smoking-friendly areas. The connection between these variations can help policymakers and public health officials make informed decisions regarding resource allocation and intervention strategies, which is often made at the county and state level.

## Background

Prior research has been conducted using cross-sectional surveys of the U.S. population in order to determine the prevalence of smoking and related factors that contribute to the uptake of smoking and the cessation of tobacco use as well as other distinct patterns of smoking habits. According to Dwyer-Lindgren et al., (2014) total cigarette smoking prevalence varies dramatically between counties even within states. Between 1996 and 2012, the counties with the highest rates of total cigarette smoking were counties in the South and counties with large Native American populations. The counties with the lowest smoking rates were counties in Utah and other Western states. Between 1996 and 2012, total smoking prevalence in the United States decreased, however, statistically significant declines were concentrated within a small number of counties and counties in the top quintile in terms of income experienced faster declines in smoking prevalence compared to counties in the bottom quintile. Jahnle et al. (2017) found that smoking was at least partially associated with socioeconomic status via differential exposure to smoking friendly environments. This corresponds with other research done in this area. According to Garrett et al., (2019) men and women who lived at or above the poverty line had a significantly lower smoking prevalence than those who lived below the federal poverty level. Garrett et al., (2019) also suggests a significant inverse relationship between smoking status and highest education level attained, finding a higher smoking prevalence among populations with lower educational attainment.

The effects of the recent pandemic of COVID-19 and subsequent lockdown on smoking have also been a focal point of research in smoking behavior and consumption. A qualitative study of 25 adult smokers reported increased smoking due to the COVID-19 lockdown (O'Donnell et al., 2021). A study of university students reported a higher risk of substance use and smoking during the pandemic lockdown (Rogés et al., 2021). However, Alomari et al. (2023) found that changes in smoking habits during the COVID-19 pandemic were reported by about 50% of survey participants of adults in Jordan who smoke tobacco, with more participants reporting a decrease than increase in use. A web-based cross-sectional study of Italian adults found that the lockdown increased cigarette consumption by 9.1% and emphasized the association of increased cigarette consumption with increased mental distress among the study's participants (Carreras et al., 2021).

Further research is required with this topic as the relationship between smoking prevalence and other factors such as income and education at the county level has yet to be explored. Previous literature focuses on survey data and reported attributes of respondents who are current smokers. In order to understand the greater relationship between local population demographics and smoking prevalence, county-level demographics should be explored further. The effect of the coronavirus should also be explored as the severity of the pandemic that counties experienced may have impacted their smoking prevalence and tobacco use.

## Theory

According to Dwyer-Lindgren et al, between 1996 and 2012,, statistically significant declines in smoking prevalence were concentrated within a small number of counties and counties in the top quintile in terms of income experienced faster declines in smoking prevalence compared to counties in the bottom quintile. According to Garrett et al., men and women who lived at or above the poverty line had a significantly lower smoking prevalence than those who lived below the federal poverty level. The theory we have developed after reviewing the existing literature is that counties with higher median incomes have better access to smoking cessation programs and support and also greater access to more diverse, recreational stress-relief opportunities. On the other hand, in counties with lower median incomes individuals may perceive smoking as a more affordable stress-relief option and also may be more susceptible to tobacco companies' marketing and pricing strategies. This theory brings us to our first hypothesis:

**$H_1$:** Counties with *lower median annual household income* in 2022 will have a *higher smoking prevalence.*

According to Jahnel et al., participants with a lower educational attainment level were more likely to encounter places where smoking was allowed compared to participants with higher educational attainment levels and participants who encountered more smoking-permissive environments smoked more cigarettes per day. Garrett et al., (2019) suggests a significant inverse relationship between smoking status and highest education level attained, finding a higher smoking prevalence among populations with lower educational attainment. The theory we have developed after reviewing the existing literature on the relationship between educational attainment level and smoking prevalence is that counties with higher educational attainment levels may be more informed about the health risks associated with smoking and are more likely to have better work and economic conditions. Individuals in counties with lower educational attainment levels may be more likely to encounter places where smoking is allowed, tolerated or even encouraged. Garrett et al., encountering more smoking-permissive environments is related to smoking more cigarettes per day. This theory brings us to our second hypothesis:

**$H_2$:** Counties with *higher levels of educational attainment* in 2022 will have *a lower smoking prevalence.*

The literature regarding COVID-19 presents opposing results and theories about the pandemic's influence on smoking prevalence. According to Alomari et al. (2023), changes in smoking habits during the COVID-19 pandemic were reported by about 50% of survey participants of adults in Jordan who smoke tobacco, with more participants reporting a decrease than increase in use. Carreras et al. (2021) found that the lockdown increased cigarette consumption by 9.1% and emphasized the association of increased cigarette consumption with increased mental distress among the study's participants. The theory

we have developed after reviewing the existing literature on the relationship between educational attainment level and smoking prevalence is that lower COVID-19 severity in a county may be indicative of successfully implementing proactive public health measures, emphasizing health consciousness of the community which may extend to a lower acceptance of smoking. Individuals in counties with higher COVID-19 severity, may undergo higher levels of stress due to health concerns, economic uncertainties, and disruptions to daily life, and may use smoking as a coping mechanism. This theory brings us to our third and final hypothesis:

> **$H_3$**: Counties that experienced *greater levels of severity of the COVID-19 virus* will have a *higher smoking prevalence* in 2022 than counties that experienced lower levels of COVID-19 severity.

## Data

We are measuring the county smoking prevalence as the percentage of the adult population who smoke. The data source for the adult smoking measure we are using is County Health Rankings published by the University of Wisconsin-Madison Population Health Institute. County Health Rankings uses the CDC's single-year modeled county-level estimates for adult smoking prevalence. Estimates for adult smoking prevalence are based on the responses to the Behavioral Risk Factor Surveillance System (BRFSS) telephone survey that reported that they currently smoke every day or some days and have smoked at least 100 cigarettes in their lifetime. The prevalence of smoking is reported as an age adjusted rate in order to fairly compare countries with differing age structures. The percentage of adult smokers of the county is the dependent variable for testing all three hypotheses.

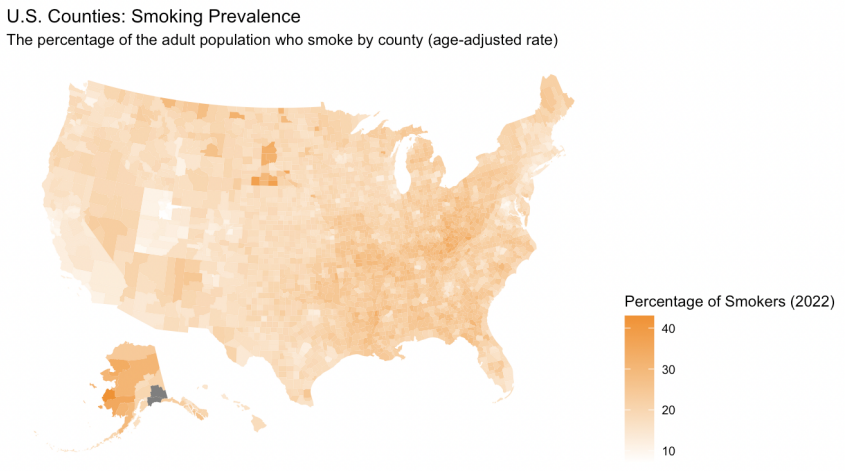



Figure 1: U.S. Counties: Smoking Prevalence Map

Figure 1 is a choropleth map visualizing the percentage of smokers across counties in the United States. A choropleth map is a type of thematic map that represents data using different shades or patterns to indicate the variation in a variable across geographic regions. (See Appendix B for implementation of

this choropleth map in R) The regions are typically shaded or colored based on the intensity of the variable being represented. Darker orange colors represent higher values of percentage of smokers, while lighter colors represent lower values. Choropleth maps are useful for visualizing spatial patterns and trends, making it easier to understand geographic variations in data. This map also allows us to visualize what median household income data is missing and for which counties with regional context. The map is missing the data for smoking prevalence from three counties. Two of the three counties are in Alaska and are missing due to the county splitting in 2019 from Valdez-Cordova county to two counties, Chugach and Copper River, with separate FIPS codes, making the data for the different variables incompatible as smoking prevalence has not updated its data for this change. The other missing county, which cannot be seen on this map, is Kalawao County, Hawaii. This county is the smallest county in the U.S. by land area and has a population of approximately 82 persons.

The independent variable of the first hypothesis is the median annual household income of the county. Median annual household income data is drawn from the Small Area Income and Poverty Estimates (SAIPE) program which produces single-year estimates of income and poverty for all U.S. states and counties with data from the American Community Survey. Median household income is based on one year of survey data and is created using complex statistical modeling. Modeling generates more stable estimates for places with small numbers of residents or survey responses.

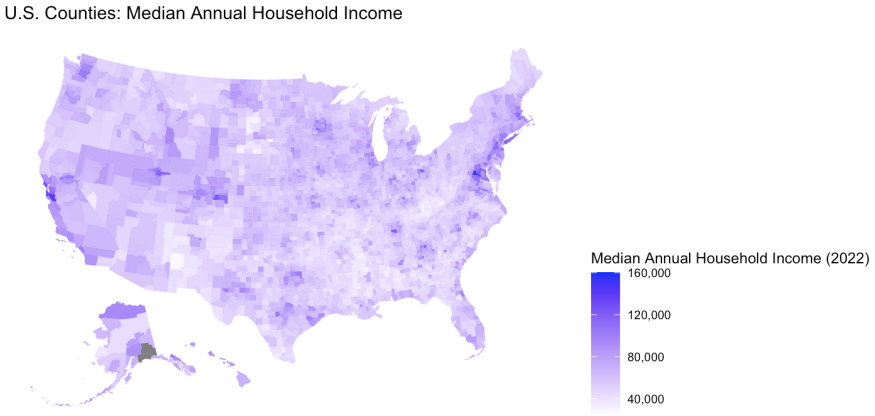



Figure 2: U.S. Counties: Median Household Income Map

Figure 2 is a choropleth map visualizing median household income across counties in the United States. Median household income data is missing for the same three counties as smoking prevalence, Chugach, Copper River, and Kalawao.

The concept in the second hypothesis is educational attainment level is operationalized as the percentage of adults aged 25 and older with a bachelor's degree or higher. The data source for the independent variable for the second hypothesis is the Census Bureau's American Community Survey estimates that are derived from survey data collected over a 5-year period from 2017 to 2021.

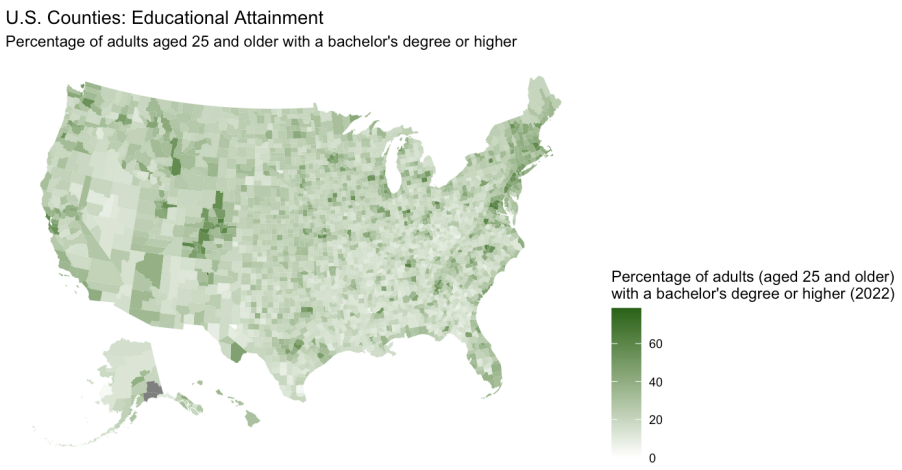



Figure 3: U.S. Counties: Median Household Income Map

Figure 3 is a choropleth map visualizing educational attainment levels across counties in the United States measured as percentage of adults in a county, aged 25 and older, with a bachelor's degree or higher. Data for educational attainment is only missing for two counties: Chugach and Copper River.

The concept of the third hypothesis, the severity of the COVID-19 virus in a county, can be measured by the number of deaths due to COVID-19 in 2020, per 100,000 population. The data source for the independent variable in the third hypothesis is the National Center for Health Statistics - Mortality Files.

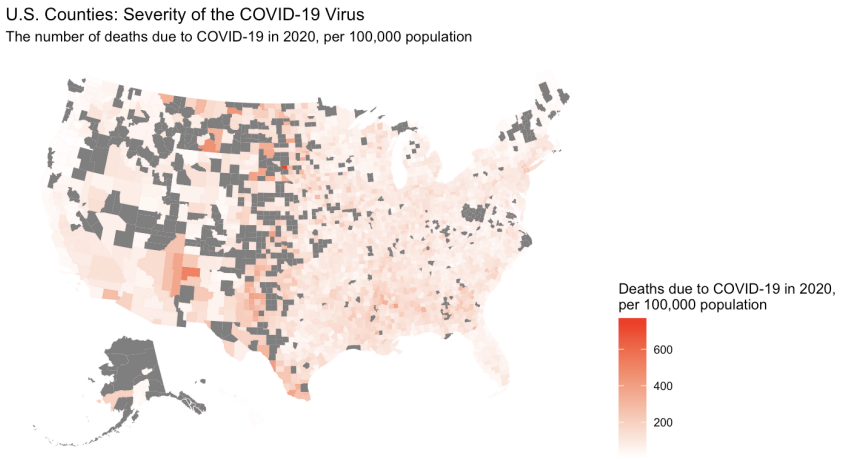



Figure 4: U.S. Counties: Severity of the COVID-19 Virus Map

Figure 4 is a choropleth map visualizing severity of the COVID-19 virus across counties in the United States. This visualization is particularly important as it highlights the areas where we do not have the data for, for deaths due to COVID-19 in 2020 per 100,000 population. To be exact we do not have the data for 578 of the 3142 counties we have which is about 18% of the counties in the United States.

## Methods

The three independent variables and the dependent variable are all continuous measures. We decided to use R and R studio to conduct our hypothesis testing. (See Appendix A) We merged the data from various sources, utilizing the FIPS code as the primary key attribute in R. We planned to first explore the correlation coefficients between our dependent variable and each of the independent variables. (See Appendix C for implementation of correlation coefficient analysis in R) We also decided to use linear regression to explore the relationship between the independent variables and dependent variable. In addition to running a bivariate linear regression model, we will also run a multiple regression model incorporating all three independent variables and other control variables that may influence the percentage of smokers. (See Appendix D and E for implementation of simple and multiple linear regression in R) In order to account for heteroskedasticity, we also decided to calculate robust standard errors to correct the bias of heteroskedasticity and interpret resulting p-values to determine significance of the regression coefficients. (See Appendix F for calculation of robust standard errors in R)

## Results

The three independent variables and the dependent variable are all continuous measures. In order to analyze the relationship between smoking prevalence and the three possible determinants, we first found the correlation coefficients.

**Table 1:** Correlation Coefficients (Pearson's) and Significance Values

| | Percentage of Smokers |
|---|---|
| Median Household Income | -0.6912086*** |
| Percent of Adults with a Bachelor's Degree or Higher | -0.6987278*** |
| COVID-19 Death Rate | 0.2453013*** |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

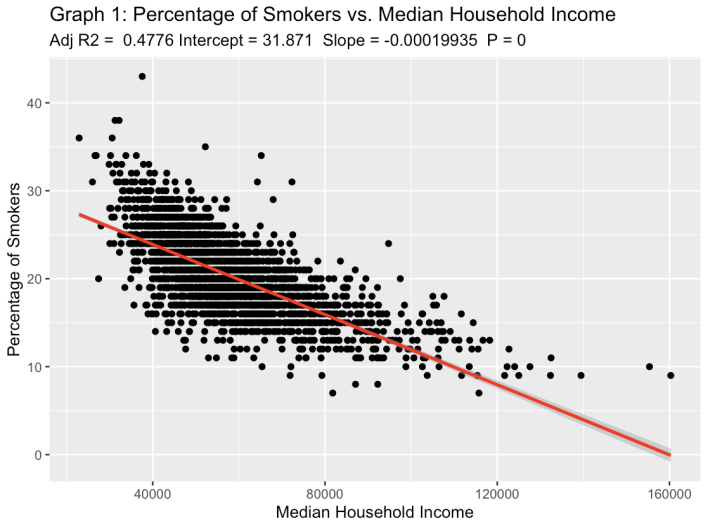
Since all three independent variables have a statistically significant relationship, we can reject the null hypothesis and conclude that there is a statistically significant relationship between the variables. Median Household Income and Percent of Adults with a Bachelor's degree or higher have an inverse association with Smoking prevalence based on the correlation coefficients. The COVID-19 Death Rate, our measure for COVID-19 severity of a county, has a positive association with smoking prevalence.

We applied linear regression to explore the relationship between the independent variables and dependent variables. Regression expresses the relationship between the independent variable and the dependent variable in the form of an equation. We investigated three simple linear regression models.

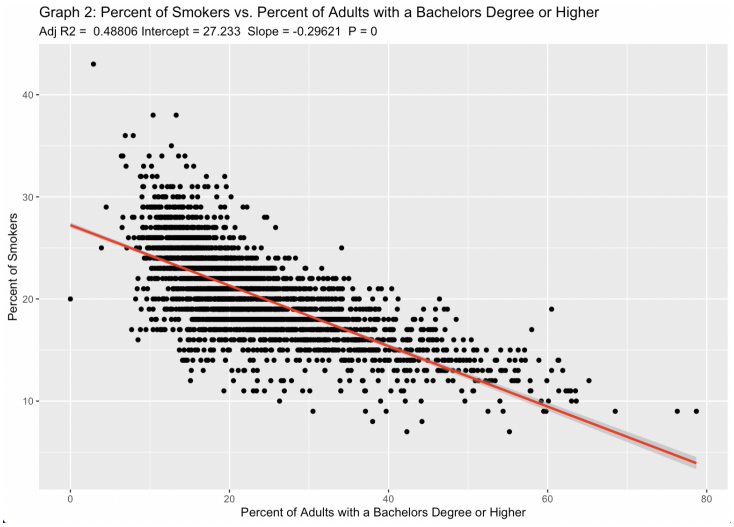
$$y_{Percentage\ of\ Smokers} = \beta_0 + \beta_1 x_{Median\ Household\ Income} + \varepsilon$$

$$y_{Percentage\ of\ Smokers} = \beta_0 + \beta_1 x_{Percent\ of\ population\ with\ Bachelor's\ degree\ or\ higher} + \varepsilon$$

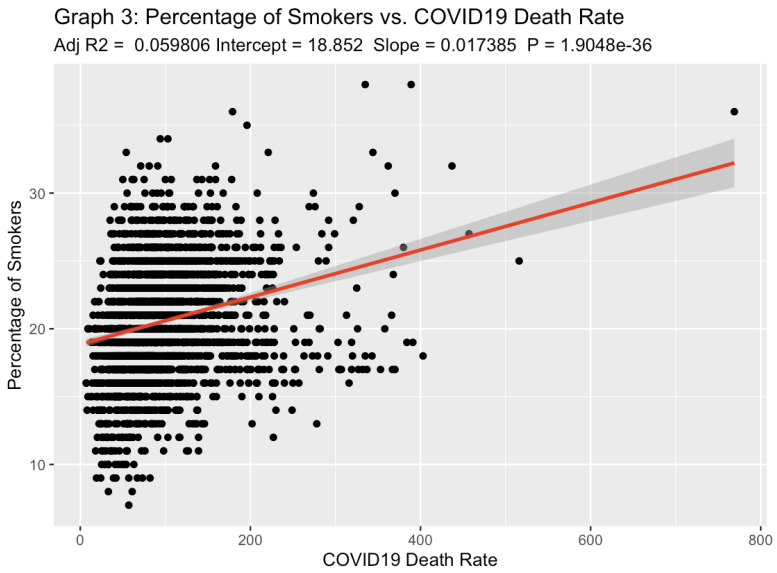
$$y_{Percentage\ of\ Smokers} = \beta_0 + \beta_1 x_{Number\ of\ deaths\ due\ to\ COVID-19\ in\ 2020,\ per\ 100{,}000\ population} + \varepsilon$$




Graph 1 presents the linear regression model between median household income and percentage of smokers. For every unit increase in median household income, the percentage of smokers decreases by 0.0001994 or for every $10,000 increase in median household income, the percentage of smokers decreases down by about 2 percentage points.




Graph 2 presents the linear regression model for percent of adults with a bachelor's degree or higher and percentage of smokers. For every percent increase in percentage of adults with a bachelor's degree or higher, the percentage of smokers decreases by 0.296 percentage points.




Graph 3 presents the linear regression model for COVID-19 death rate and percentage of smokers. For every unit increase in COVID-19 death rate, the percentage of smokers increases by 0.0174. We can also see an issue here with heteroskedasticity which we need to account for with robust standard errors in order to confirm statistical significance.

**Table 2: Simple Linear Regression**

| | *Dependent variable:* | | |
|---|---|---|---|
| | Percentage of Smokers | | |
| | (1) | (2) | (3) |
| Median Household Income | -0.0002*** | | |
| | (0.00000) | | |
| Percent of Adults with a Bachelors Degree or Higher | | -0.296*** | |
| | | (0.005) | |
| COVID19 Death Rate | | | 0.017*** |
| | | | (0.001) |
| Constant | 31.871*** | 27.233*** | 18.852*** |
| | (0.221) | (0.136) | (0.157) |
| Observations | 3,140 | 3,141 | 2,564 |
| $R^2$ | 0.478 | 0.488 | 0.060 |
| Adjusted $R^2$ | 0.478 | 0.488 | 0.060 |
| Residual Std. Error | 3.040 (df = 3138) | 3.010 (df = 3139) | 4.035 (df = 2562) |
| F Statistic | 2,870.839*** (df = 1; 3138) | 2,994.501*** (df = 1; 3139) | 164.033*** (df = 1; 2562) |
| *Note:* | | | *p<0.1; **p<0.05; ***p<0.01 |

Table 2 above presents the beta coefficients and statistical significance of each of the three simple linear regression models. We can see a statistically significant relationship between the dependent variable and all three independent variables with all three p-values less than the baseline of 0.01.

In addition to the linear regression models, we also ran a multiple regression model. The multiple regression model incorporates all three independent variables.

$$y_{Percentage\ of\ Smokers} = \beta_0 + \beta_1 x_{Median\ Household\ Income} + \beta_2 x_{Percent\ of\ population\ with\ Bachelor's\ degree\ or\ higher} + \beta_3 x_{number\ of\ deaths\ due\ to\ COVID-19\ in\ 2020,\ per\ 100{,}000\ population} + \varepsilon$$

Multiple regression allows us to investigate the relationship between a single dependent variable and several independent variables. Even if the individual simple linear models showed significance in a variable, this significance may be eliminated if it is related to one of the other independent variables. If this is the case, when there are fixed values of the other independent variables, changes in this independent variable will not significantly affect the percentage of smokers. Multiple regression aids in determining the importance of each of the predictors to the relationship, with the effect of the other predictors statistically eliminated.

**Table 3: Multiple Linear Regression**

| | *Dependent variable:* |
|---|---|
| | Percentage of Smokers |
| Median Household Income | -0.0001*** |
| | (0.00001) |
| Percent of Adults with a Bachelors Degree or Higher | -0.173*** |
| | (0.008) |
| COVID19 Death Rate | -0.003*** |
| | (0.001) |
| Constant | 31.780*** |
| | (0.260) |
| Observations | 2,564 |
| $R^2$ | 0.609 |
| Adjusted $R^2$ | 0.608 |
| Residual Std. Error | 2.604 (df = 2560) |
| F Statistic | 1,328.520*** (df = 3; 2560) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

After examining the estimates of the multiple regression beta coefficients, we can see that each variable still has a statistically significant relationship with the percentage of smokers. Median household income and percent of adults with a bachelor's degree or higher have an inverse relationship with percentage of smokers. However, the coefficient of COVID-19 death rate changed drastically. With median household income and percent of adults with a bachelor's degree or higher held constant, COVID-19 death rate presents a statistically significant inverse association with percentage of smokers.

In the multiple regression model with all three variables, COVID-19 severity switched direction. This change in the direction of the coefficient may indicate multicollinearity. Multicollinearity occurs when independent variables in a regression model are highly correlated with each other. Multicollinearity can lead to issues in estimating the individual effects of each variable. We decided to run a correlation matrix to report the correlation coefficients between the independent variables. (See Appendix G for implementation of correlation matrix in R)

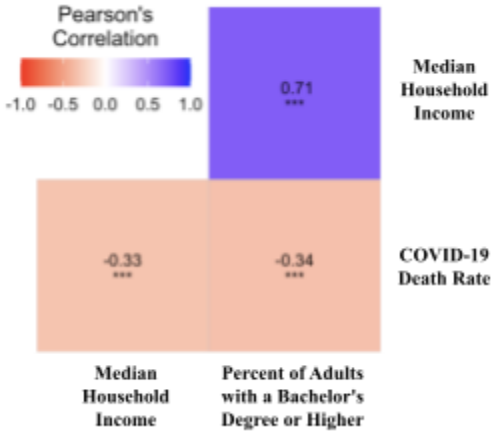



Figure 5: Correlation Matrix with Independent Variables

There is a high correlation coefficient (0.71) between median household income and educational attainment level which suggests multicollinearity.

**Table 4: Multiple Linear Regression without Median Household Income**

| | *Dependent variable:* |
|---|---|
| | Percentage of Smokers |
| Percent of Adults with a Bachelors Degree or Higher | -0.301*** |
| | (0.006) |
| COVID19 Death Rate | -0.0002 |
| | (0.001) |
| Constant | 27.548*** |
| | (0.206) |
| Observations | 2,564 |
| $R^2$ | 0.525 |
| Adjusted $R^2$ | 0.525 |
| Residual Std. Error | 2.868 (df = 2561) |
| F Statistic | 1,417.329*** (df = 2; 2561) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

**Table 5: Multiple Linear Regression without Educational Attainment Level**

| | *Dependent variable:* |
|---|---|
| | Percentage of Smokers |
| Median Household Income | -0.0002*** |
| | (0.00000) |
| COVID19 Death Rate | 0.001 |
| | (0.001) |
| Constant | 32.158*** |
| | (0.284) |
| Observations | 2,564 |
| $R^2$ | 0.532 |
| Adjusted $R^2$ | 0.532 |
| Residual Std. Error | 2.848 (df = 2561) |
| F Statistic | 1,456.434*** (df = 2; 2561) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

We decided to run two multiple linear regression models, removing at least one of the two variables that suggest multicollinearity. The results of these models can be seen in table 4 and 5. The

significance of the relationship between smoking prevalence and COVID-19 death rate is no longer statistically significant in both models.

**Table 6: Multiple Linear Regression without COVID-19 Severity**

| | *Dependent variable:* |
|---|---|
| | Percentage of Smokers |
| Median Household Income | -0.0001*** |
| | (0.00000) |
| Percent of Adults with a Bachelors Degree or Higher | -0.178*** |
| | (0.007) |
| Constant | 31.031*** |
| | (0.204) |
| Observations | 3,140 |
| $R^2$ | 0.565 |
| Adjusted $R^2$ | 0.565 |
| Residual Std. Error | 2.775 (df = 3137) |
| F Statistic | 2,036.715*** (df = 2; 3137) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

As can be seen in table 6, the multiple linear regression model without COVID-19 severity still shows significance in the relationships between smoking prevalence and the other two independent variables. It is important to note that as the COVID-19 severity measure has missing data, models with the COVID-19 death rate variable have less observations. The model without COVID-19 has 3,140 observations. Since COVID-19 death rate has missing data, the total observations for models with COVID-19 severity is 2,564. Missing data can also affect the stability of estimates.

A change in the direction of the relationship between a predictor variable (such as COVID-19 severity) and the response variable (smoking prevalence) when moving from a simple linear regression to a multiple regression model can also be indicative of heteroscedasticity. Heteroscedasticity refers to when the variability of the errors is not constant across all levels of the independent variables. In the context of regression analysis, this means that the spread of residuals changes as the values of the predictor variables change. In order to account for heteroskedasticity to confirm the relationships are statistically significant, we analyzed the p-values of the multiple regression model with robust standard errors.

Table 7: Multiple Linear Regression with Robust Standard Errors with COVID-19 Severity

| | **Coefficients** | **Robust_SE** | **t_value** | **p_value** |
|---|---|---|---|---|
| (Intercept) | 31.7798934880868 | 0.350280114655214 | 90.7270842918622 | 0.0000 |
| Median.Household.Income | -0.000120089118623953 | 5.91028840022666e-06 | -20.3186562976093 | 0.0000 |
| Percent.of.adults.with.a.bachelors.degree.or.higher | -0.172848584918991 | 0.00770948448293894 | -22.4202520027771 | 0.0000 |
| COVID19.death.rate | -0.00272108063075576 | 0.00163430497924222 | -1.66497726270004 | 0.0960 |

From the results in table 7, we can see that although the relationship between median household income and percent of adults with a bachelor's degree or higher with percentage of smokers remained statistically significant, COVID-19 death rate no longer presents a statistically significant relationship with percentage of smokers. This may mean that the change in the direction of the relationship between COVID-19 death rate and percentage of smokers in the multiple regression model is due to biased standard error. Biased standard errors can also lead to incorrect inference about the statistical significance of coefficients.

## Conclusion

Our research aimed to explore the relationship between county-level socio-demographic factors and smoking prevalence in the United States in 2022, with a specific focus on the impact of median annual household income, educational attainment, and the severity of the COVID-19 virus in 2020. Through our analysis, we found compelling evidence supporting significant associations between these factors and smoking prevalence.

Our first hypothesis suggested that counties with lower median annual household income in 2022 would have a higher smoking prevalence. This hypothesis was supported by our findings, indicating an inverse relationship between median household income and smoking prevalence. The data suggested that as median household income increased, the percentage of smokers decreased. This aligns with existing literature, emphasizing the potential influence of economic conditions on smoking behavior.

The second hypothesis proposed that counties with higher levels of educational attainment in 2022 would have a lower smoking prevalence. Our analysis supported this hypothesis, revealing a negative association between the percentage of adults with a bachelor's degree or higher and smoking prevalence. This implies that higher educational attainment may contribute to increased awareness of the health risks associated with smoking, as well as improved economic conditions, ultimately leading to lower smoking rates.

The third and final hypothesis suggested that counties experiencing greater levels of severity of the COVID-19 virus in 2020 would have a higher smoking prevalence in 2022. Initially, our analysis showed a positive association between the COVID-19 death rate and smoking prevalence, indicating that counties with higher COVID-19 severity had higher smoking rates. However, after addressing multicollinearity and heteroskedasticity issues, the relationship between COVID-19 severity and smoking prevalence became statistically insignificant.

Multicollinearity, particularly between median household income and educational attainment, posed challenges in our analysis. Upon running multiple regression models and addressing multicollinearity, we found that the relationship between COVID-19 severity and smoking prevalence lost

its statistical significance. This suggests the importance of considering the interplay between socio-demographic variables when studying their impact on smoking prevalence.

Our study contributes valuable insights into the complex interconnections between socio-demographic factors and smoking prevalence at the county level. The findings underscore the significance of economic conditions and educational attainment in a county in shaping smoking behavior. While the association with COVID-19 severity was not found statistically significant in our analysis, further research may explore the nuanced relationship between health crises and smoking patterns. Policymakers and public health officials can utilize these findings to tailor interventions and allocate resources effectively, fostering a targeted and inclusive approach to reducing smoking rates and improving overall community well-being.

## References


Alomari, M. A., Khabour, O. F., Alzoubi, K. H., & Maikano, A. B. (n.d.). *The impact of covid-19 pandemic on tobacco use: A population-based study*. PLOS ONE. https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0287375

Carreras, G., Lugo, A., Stival, C., Amerio, A., Odone, A., Pacifici, R., Gallus, S., & Gorini, G. (2021). Impact of covid-19 lockdown on smoking consumption in a large representative sample of Italian adults. *Tobacco Control*, *31*(5), 615–622. https://doi.org/10.1136/tobaccocontrol-2020-056440

*Correlation (Pearson, Kendall, Spearman)*. Statistics Solutions. (2021a, August 10). https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/correlation-pearson-kendall-spearman/

Dwyer-Lindgren, L., Mokdad, A. H., Srebotnjak, T., Flaxman, A. D., Hansen, G. M., & Murray, C. J. (2014). Cigarette smoking prevalence in US counties: 1996-2012. *Population Health Metrics*, *12*(1). https://doi.org/10.1186/1478-7954-12-5

Garrett, B. E., Martell, B. N., Caraballo, R. S., & King, B. A. (2019). Socioeconomic differences in cigarette smoking among sociodemographic groups. *Preventing Chronic Disease*, *16*. https://doi.org/10.5888/pcd16.180553

*Homoscedasticity*. Statistics Solutions. (2021b, August 3). https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/homoscedasticity/

*How to interpret regression analysis results: P-values and coefficients*. Minitab Blog. (n.d.). https://blog.minitab.com/en/adventures-in-statistics-2/how-to-interpret-regression-analysis-results-p-values-and-coefficients

*Introduction to R and rstudio*. Introduction to R - ARCHIVED. (2017, September 8). https://hbctraining.github.io/Intro-to-R/lessons/01_introR-R-and-RStudio.html

Jahnel, T., Ferguson, S. G., Shiffman, S., Thrul, J., & Schüz, B. (2018). Momentary smoking context as a mediator of the relationship between SES and smoking. *Addictive Behaviors*, *83*, 136–141. https://doi.org/10.1016/j.addbeh.2017.12.014

Johnston, S. (2015, April 23). *A quick and easy function to plot LM() results with GGPLOT2 in R*. Johnston Lab. https://sejohnston.com/2012/08/09/a-quick-and-easy-function-to-plot-lm-results-in-r/

Kassambara, Leo, Sara, Mbugua, F., Kassambara, & Visitor. (2018, March 10). *Multiple linear regression in R*. STHDA. http://www.sthda.com/english/articles/40-regression-analysis/168-multiple-linear-regression-in-r/#:~:text=Multiple%20regression%20analysis%20allows%20researchers,of%20other%20predictors%20statistically%20eliminated.

Miller, S. (2021, June 25). *What is R used for? exploring the R programming language*. Codecademy Blog. https://www.codecademy.com/resources/blog/what-is-r-used-for/

*Multiple regression analysis*. Multiple Regression Analysis - an overview | ScienceDirect Topics. (n.d.). https://www.sciencedirect.com/topics/economics-econometrics-and-finance/multiple-regression-analysis#

O'Donnell, R., Eadie, D., Stead, M., Dobson, R., & Semple, S. (2021). 'I was smoking a lot more during lockdown because I can': A qualitative study of how UK smokers responded to the COVID-19 lockdown. *International Journal of Environmental Research and Public Health*, *18*(11), 5816. https://doi.org/10.3390/ijerph18115816

Porras, E. M. (2022, December 5). *R linear regression tutorial: LM function in R with code examples*. DataCamp. https://www.datacamp.com/tutorial/linear-regression-R

Rogés, J., Bosque-Prous, M., Colom, J., Folch, C., Barón-Garcia, T., González-Casals, H., Fernández, E., & Espelt, A. (2021). Consumption of alcohol, cannabis, and tobacco in a cohort of adolescents before and during COVID-19 confinement. *International Journal of Environmental Research and Public Health*, *18*(15), 7849. https://doi.org/10.3390/ijerph18157849

*RStudio desktop*. Posit. (2023, September 22). https://posit.co/download/rstudio-desktop/

*What is R?*. R. (n.d.). https://www.r-project.org/about.html

## Appendix A: R and RStudio

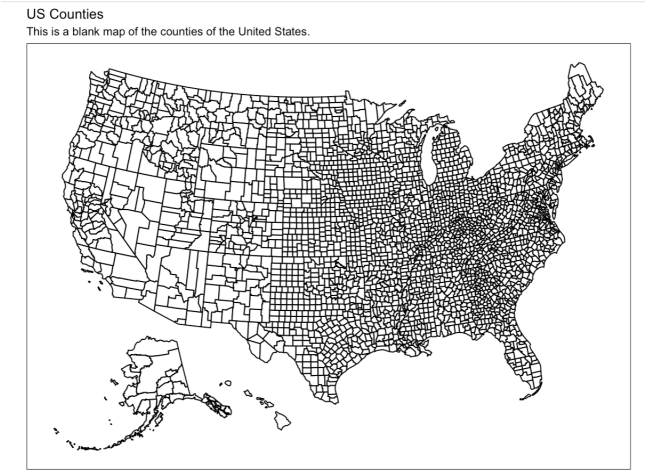
R is an open-source programming language and environment designed specifically for statistical computing and data analysis. It was developed by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is now maintained by the R Development Core Team. R supports common programming constructs such as loops, conditional statements and offers an extensive range of statistical functions and libraries for tasks like regression analysis, hypothesis testing, and data modeling. R also has exceptional data visualization capabilities and includes packages for the creation of highly customizable and publication-quality plots. R has the ability to import data from various file formats and provides functionality of exporting results to different formats. R is available for Windows, macOS, and various Unix-like operating systems.

RStudio is an open-source integrated development environment specifically designed for R programming. RStudio's script editor provides a convenient interface for writing and editing R scripts with useful features such as code highlighting and auto-completion. RStudio has an embedded R console which enables users to run code and view results within the application. RStudio also provides a workspace and data viewer allowing users to view initialized data frames and variables.

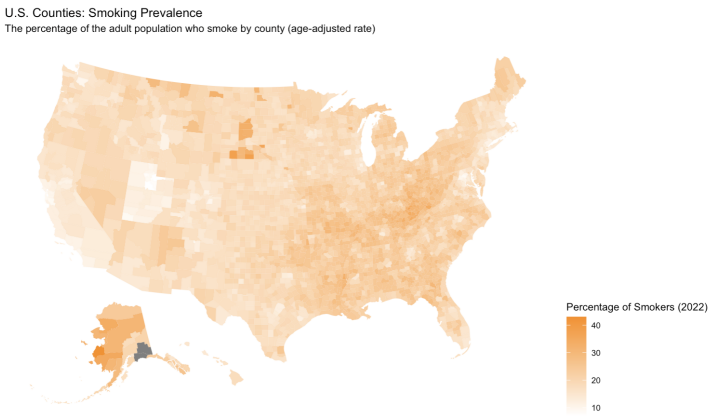
## Appendix B: U.S. County Maps

R package, usmap, allows for implementation of our U.S. County Maps. The usmap package has convenience functions for plotting choropleths and compatibility for working with FIPS codes. plot_usmap(...) is the function we used to implement the choropleths. One of the arguments of plot_usmap(...) is “regions”. This argument takes the region breakdown for the map, which in our case is "counties”. The argument “values” takes the name of the column that contains the values to be associated with a given region.

```
plot_usmap(regions = "counties") +
labs(title = "US Counties", subtitle =  "This is a
blank map of the counties of the United States.")
+ theme(panel.background = element_rect(color =
"black", fill = "white"))
```



```
plot_usmap(data = df_merge, values =
"Percentage.of.Smokers", color =
yarrr::transparent("orange", trans.val = .99)) +
scale_fill_continuous(low = "white", high =
"darkorange", name = "Percentage of Smokers
(2022)", label = scales::comma) +
theme(legend.position = "right") + labs(title =
"U.S. Counties: Smoking Prevalence", subtitle =
"The percentage of the adult population who smoke
by county (age-adjusted rate)")
```




## Appendix C: Correlation Coefficient

The correlation coefficient quantifies the strength of the linear relationship between a pair of variables. The correlation coefficient measures the degree of association ranging from -1 to 1. A correlation coefficient with a value of -1 indicates an inverse correlation between the variables, with the increase in one correlating with a decrease in the other. A positive correlation coefficient indicates that an increase in one of the variables is associated with an increase in the other. We can implement a correlation coefficient analysis in R. R calculates the correlation coefficient with the function cor(). In its most basic form, R's correlation coefficient function takes two arguments, the dependent variable and independent variable. It also has the argument use which takes an optional character string for computing correlation in the presence of missing values. We used the method = "pearson" argument to use the pearson coefficient, which is a commonly used correlation statistic to measure the degree of the relationship between linearly related variables.

```
cor(x, y = NULL, use = "complete.obs",  method = "pearson")
```

When investigating the correlation coefficients, we determined the significance of the correlation coefficient with the p-value. Through testing the probability that the true value of the coefficient is equal to zero, we can determine the significance of the strength of the linear relationship. If the reported p-value is less than 0.05, we can reject the null hypothesis and conclude a linear correlation between the variables.

## Appendix D: Simple Linear Regression

Regression expresses the relationship between the independent variable and the dependent variable in the form of an equation. The function used for building linear models in R is lm(). In simple linear regression, the lm() function takes two arguments, the formula and the data.

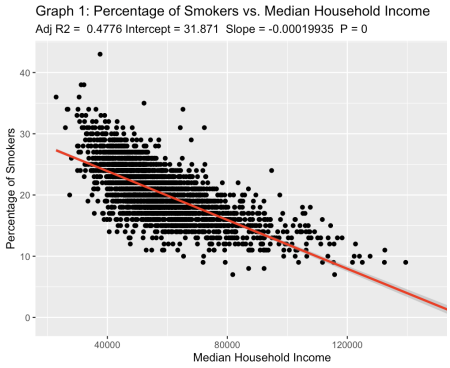
```
incomeModel <- lm(Percentage.of.Smokers ~ Median.Household.Income,
data=final_df)
```

We used p-value to interpret the significance of the coefficients in linear regression. A p-value less than 0.05 will indicate that we can reject the null hypothesis, that the coefficient is equal to 0, and that there is a statistically significant relationship between the independent variables and dependent variable.

The graphs visualizing the simple linear regression model were implemented using a function that takes a fitted model and uses the r package, ggplot, to graph the model.

```
incomeModel<- lm(Percentage.of.Smokers ~
Median.Household.Income, data=final_df)
ggplotRegression <- function (fit) {
ggplot(fit$model, aes_string(x = names(fit$model)[2], y =
names(fit$model)[1])) + geom_point()+ stat_smooth(method = "lm",
col = "red") + labs(title = "Graph 1: Percentage of Smokers vs.
Median Household Income", subtitle=paste("Adj R2 = ",
signif(summary(fit)$adj.r.squared, 5), "Intercept
=",signif(fit$coef[[1]],5 )," Slope =",signif(fit$coef[[2]], 5),
" P = " , signif(summary(fit)$coef[2,4], 5)))+ xlab("Median
Household Income") +
ylab("Percentage of Smokers")
}

plot <- ggplotRegression(lm(Percentage.of.Smokers ~
Median.Household.Income, data = final_df))
plot
```




## Appendix E: Multiple Linear Regression

Multiple regression allows us to investigate the relationship between a single dependent variable and several independent variables. Multiple regression aids in determining the importance of each of the predictors to the relationship, with the effect of the other predictors statistically eliminated. In order to perform multiple regression in r, we can use the same lm() function and expand upon the formula by providing the other two independent variables for one model.

```
allModels <- lm(Percentage.of.Smokers ~ Median.Household.Income +
Percent.of.adults.with.a.bachelors.degree.or.higher +
COVID19.death.rate, data=final_df)
```

## Appendix F: Multiple Linear Regression with Robust Standard Errors

Heteroscedasticity, which is the violation of homoscedasticity, occurs when the size of the error term differs across values of an independent variable. Incorrect conclusions can be made about the significance of the regression coefficients when standard errors are biased. We calculated robust standard errors to correct this bias and interpret resulting p-values to determine significance of the regression coefficients.

```
# Calculate robust standard errors
robust_se <- sqrt(diag(vcovHC(allModels)))
robust_se
# Create a data frame to store the results
results <- data.frame(
  Coefficients = coef(allModels),
  Robust_SE = robust_se,
  t_value = coef(allModels) / robust_se,
  p_value = sprintf((2 * (1 - pt(abs(coef(allModels) / robust_se),
df=df.residual(allModels)))), fmt = '%#.4f'))
```

## Appendix G: Correlation Matrix

```
library(metan)
corr_df <- na.pass(final_df)
keeps <- c("Median.Household.Income",

"Percent.of.adults.with.a.bachelors.degree.or.higher",
           "COVID19.death.rate")
corr_df <- corr_df[keeps]
ALL <- corr_coef(corr_df)
plot(ALL)
```