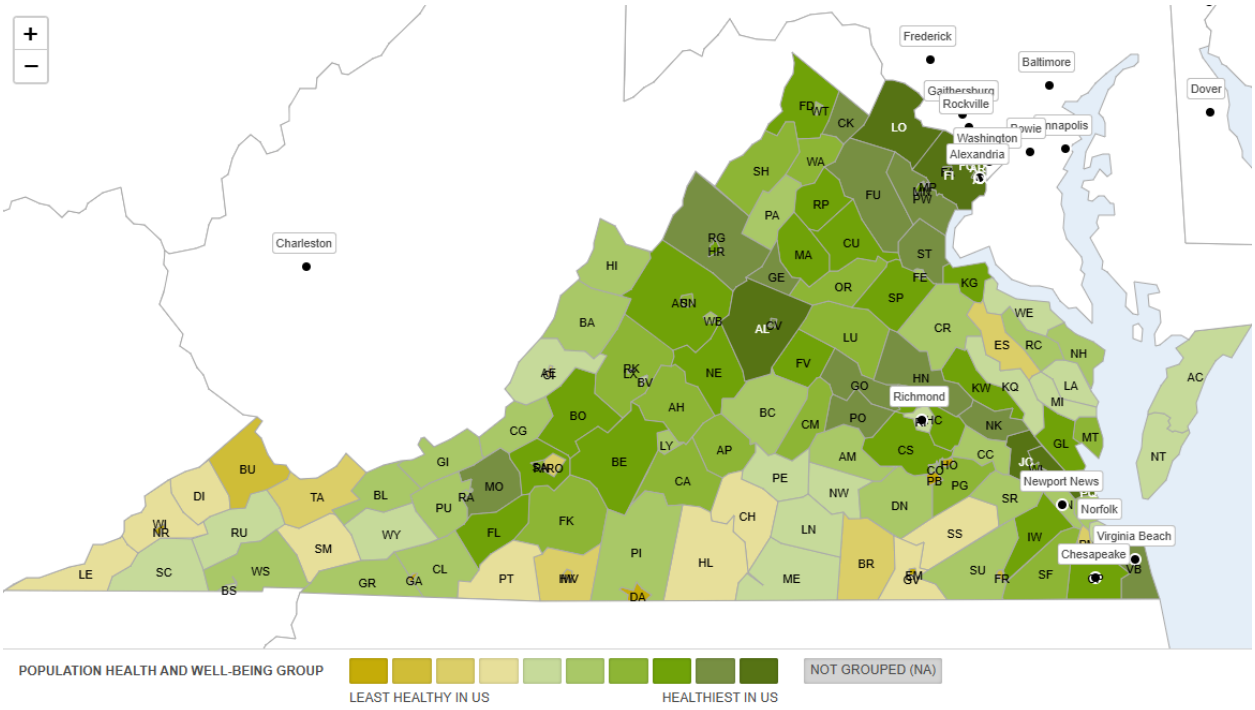
Virginia's Longevity Divide: Income, Obesity, and Exercise Access Across Counties

Team REGAN



Introduction

Despite the fact that life expectancy is frequently thought of as a succinct measure of a community's general well-being, significant differences still occur even within a single state. The 2025 County Health Rankings report, which highlights differences in the state's progress toward public health goals, found that the average life expectancy in Virginia's 133 counties and independent cities is just over seventy-seven years (County Health Rankings & Roadmaps, 2025). Decades of epidemiological research, such as the Centers for Disease Control and Prevention's findings on social determinants of health (Hacker, 2022) and the World Health Organization's analyses of obesity-related mortality (WHO, 2025), identified three recurring factors that influence longevity: household economic resources, the burden of chronic diseases (obesity), and the characteristics of local environments (such as rural versus urban areas). Knowing how these elements appear in Virginia counties can help direct focused interventions and fair resource distribution.

The overarching question guiding this study is:

Which socioeconomic, health-behavior, and environmental factors best explain the variation in 2025 life expectancy across Virginia's counties and independent cities?

The average life expectancy in years is the outcome of interest, and we created a dataset using publicly accessible data from the County Health Rankings portal, where each row represents a single jurisdiction.

Three specific research questions are presented to focus the general investigation. First, *is the average life expectancy longer in jurisdictions with greater median household incomes?* This tackles the quantifiable relationship that previous national studies have proposed between longevity and economic prosperity. Second, *is life expectancy lower in counties with the highest tertile of adult obesity rates than in those with the lowest tertile?* This investigates the relationship between mortality risk and a quantifiably modifiable health behavior variable. Third, *is there a difference in mean life expectancy between Virginia's primarily rural and urban jurisdictions?* This question investigates whether geographic context alone confers a longevity advantage or disadvantage by contrasting a qualitative classification of place.

Data Summary

Data Sources

This analysis is based on the 2025 Virginia excerpt from the *County Health Rankings & Roadmaps program*, which collects county-level health indicators across the country annually through a partnership between the Robert Wood Johnson Foundation and the University of Wisconsin Population Health Institute. The institute collects each metric directly from authorized federal sources, such as the CDC WONDER database, the American Community Survey, and the Bureau of Labor Statistics, before harmonizing the series to a single 2025 reference year.

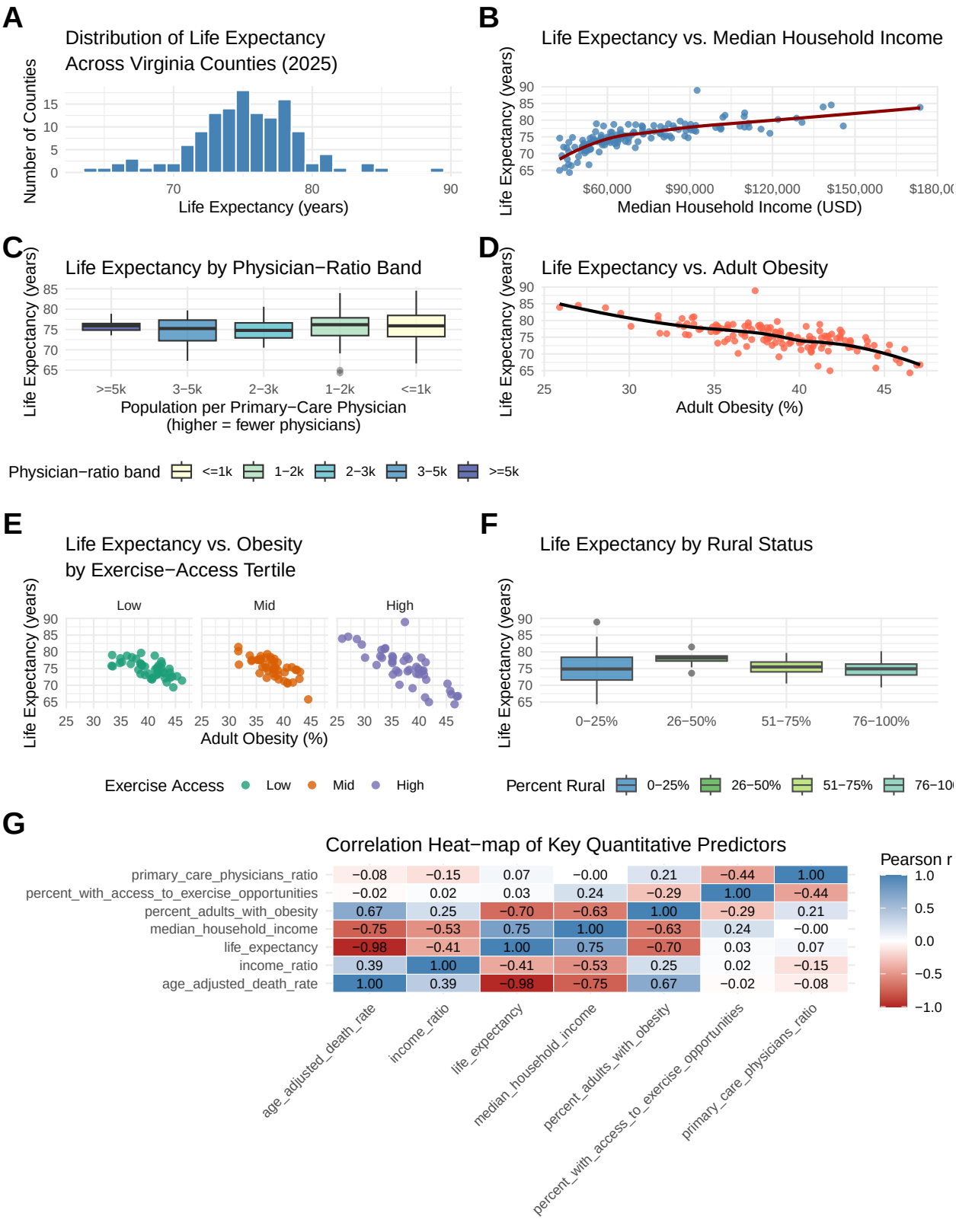
For this study, the population is the complete set of 133 Virginia counties and independent cities. Two companion tables provided by the Rankings, Select Measure Data and Additional Measure Data, contain (i) core health outcomes and (ii) socioeconomic and health-behavior covariates. These tables were combined on each county's five-digit FIPS code to create a single cross-sectional data frame. A continuous response variable—life expectancy at birth (years)—was retained exactly as reported, ensuring comparability with CDC methodology.

Several small, logically motivated changes were necessary. The ratio of primary care physicians was provided as a text string, such as "2,210:1." To ensure that lower figures indicate increased provider availability, it was transformed to a straightforward numerical count of residents per physician. To draw attention to non-linear gradients in built-environment resources, the percentage of people with access to exercise opportunities was recoded into an ordered three-level factor (Low, Mid, and High). Ultimately, the rural population percentage was divided into four equal-width groups, referred to as rural bands: 0–25%, 26–50%, 51–75%, and 76–100%. This discretization facilitates the understanding of regional differences while maintaining a monotonic ordering. After removing entries with missing values and one jurisdiction without a county name, 132 observations were included in the study sample.

Because the County Health Rankings use open documentation of data provenance and imputation procedures and are regularly cited in peer-reviewed health services research, they are highly credible. Three warnings, though, are essential to note. First, specific data may add to the uncertainty by displaying suppressed numbers for a small number of counties. Second, since multiple focus areas of data are collected in various years, there is no set time for retrieval. Last but not least, measurements are gathered from numerous trustworthy sources, each using a different method for collecting data, and then combined into sheets. The overall integrity of the dataset is unaffected by these limitations, which will be examined when assessing model results.

Exploratory Data Analysis

Key Exploratory Plots



EDA Summary

The response variable, life expectancy at birth, has a bell-shaped distribution with a slight skew (–0.04), a mean of 75.15 years, and a standard deviation of just 3.89 years. Since the variable covers more than twenty-four years (64.3–88.9), it satisfies the normality and homoscedasticity assumptions usually needed for multiple linear regression; however, most counties cluster firmly between 73 and 78 years. Therefore, it appears that no transformation is necessary.

A substantial, positive correlation, which steadily plateaus above USD 90,000, is evident when life expectancy is plotted against median household income. Together, the LOESS fit (residual SE = 2.33 years) and the Pearson correlation of 0.75 suggest that economic improvement in lower-income areas may result in significant longevity benefits. Extremely wealthy regions, on the other hand, already seem to be close to a ceiling. On the other hand, the adult obesity rate exhibits a linear, detrimental effect: a 0.68-year drop in life expectancy is predicted for every percentage point rise, and the correlation of –0.72 accounts for more than half of the variation (R2 = 0.52). When obesity is re-examined within tertiles of access to exercise opportunities, its relationship with longevity steepens in high-access environments (r = –0.79, slope = –0.84), suggesting that behavioral choices, rather than structural constraints, exacerbate health disparities in environments with plentiful facilities.

Provider availability also matters, though more subtly. Counties with fewer than 1,000 residents per primary-care physician record a median expectancy of 75.9 years, while mid-access bands (2,000–5,000 residents per doctor) drop to 74.8–75.2 years before a slight rebound in the scarcest group. This pattern suggests that physician density exerts incremental benefits only up to a threshold, beyond which contextual factors—such as wealth, built environment, or rurality—likely dominate. The geographic context itself is significant. The most extended lifespans (median = 78.2 years) are found in jurisdictions with 26–50% rural populations, outliving both highly urban counties (74.8 years) and the most rural groups (= 75.5 years). While larger dispersion within metropolitan areas (SD = 5.65) indicates a mix of impoverished inner-city populations and wealthy suburbs, semi-rural regions tend to be more socioeconomically homogeneous and generally healthier.

The correlation heat-map clarifies potential collinearity. Age-adjusted death rate is nearly a mirror image of the outcome itself (r = –0.98); therefore, it will be excluded from formal modelling to avoid redundancy. Income and obesity correlate at –0.63, a level that may inflate variance inflation factors but not so high as to demand outright removal; instead, diagnostics will verify tolerable VIF values. Other predictors—physician ratio, rurality, and

exercise access—display only modest intercorrelations (|r| </= 0.29), indicating that they each convey largely distinct information about county environments.

All things considered, the exploratory study shows that it is appropriate to use multiple linear regression with life expectancy as the continuous response. It also supports a principled variable-screening approach that maintains income, obesity, exercise access, rural band, and physician ratio as core predictors, monitors multicollinearity primarily between income and obesity, and limits the age-adjusted death rate for conceptual and statistical reasons. Because the final model will account for both quantitative gradients and categorical contexts, it will be in a good position to explain why some Virginians live noticeably longer than others.

Methods and Analysis

The multiple-linear-regression concept map's suggested sequence was adhered to by our analytical approach. Using "na.exclude" to preserve row indices, we re-fit an ordinary least-squares (OLS) model on the 126 complete cases after exploratory analysis revealed five conceptually relevant predictors: median household income, adult obesity prevalence, access to exercise opportunities, rurality band (orthogonally coded into three contrasts), and the primary-care-physician ratio. The overall F-test was highly significant ($F(7, 118) = 46.7$, $p < 0.001$); income and obesity were strongly significant ($|t| > 2.9$), access to exercise met the 0.05 threshold ($t = -2.02$, $p = 0.046$). In contrast, physician density was only marginal ($t = 1.73$, $p = 0.086$). We regressed the absolute OLS residuals on the fitted values to look for possible heteroscedasticity; the positive slope showed that variance rose as the mean increased. Thus, we estimated a weighted least-squares (WLS) model with the precise specification after modeling the residual variance and computing observation-specific weights as the inverse squared fitted values.

The WLS fit enhanced precision and goodness of fit: the adjusted R2 increased from 0.719 to 0.741, and the residual standard error decreased from 1.87 to 1.35 years. Multicollinearity is minimal, according to variance-inflation diagnostics, which showed that all VIFs for the three-df rurality factor were </= 1.81, well below the traditional concern threshold of 10. As an alternative remedy for non-normality and scale issues, we obtained a data-driven Box–Cox transformation of the response (lambda = 0.13) and refit an unweighted model on the transformed scale. Although the Box–Cox model (lambda = 0.13) preserved the same sign pattern, its fit was markedly worse (AIC = 1608, BIC = 1633) than both the OLS (AIC = 525, BIC = 550) and WLS (AIC = 523, BIC = 548) models, so it was not pursued further.

The WLS specification's model diagnostics were produced. A Shapiro–Wilk test of standard-

ized residuals supports approximate normality ($W = 0.985$, $p = 0.166$), the scale-location plot validates variance stabilization, the residual-versus-fitted panel displays no systematic pattern, and the Q–Q plot closely aligns with the 45-degree line. Influence measures reveal eight observations with Cook's distances exceeding the $4/n$ threshold ($= 0.032$). The largest Cook's distance is 0.62—slightly above the informal 0.5 flag—yet leverage values remain moderate, so no single county dominates the estimates, although two observations warrant closer substantive review (diagnostic plots are provided in the appendix.) External validation was impl emented in two ways. First, a 70 / 30 hold-out split yielded a root-mean-square prediction error of 1.94 years on the test set, consistent with in-sample residual dispersion. Second, ten-fold cross-validation on the complete, weight-augmented data produced an average RMSE of 1.94 and an average R^2 of 0.712, indicating stable generalization performance.

Comparative evidence from fit statistics, residual behavior, and validation errors all favored the weighted least-squares specification. Accordingly, the WLS model was adopted as the final model for subsequent interpretation, while acknowledging the presence of a small cluster of moderately influential counties that merit substantive examination rather than statistical exclusion.

Results

Conclusions

Appendix A: Data Dictionary

Variable Name	Abbreviated Name	Description

Appendix B: Data Rows

Appendix C: Final Model Output and Plots

Appendix D: References

Background

Data Sources

Additional Help