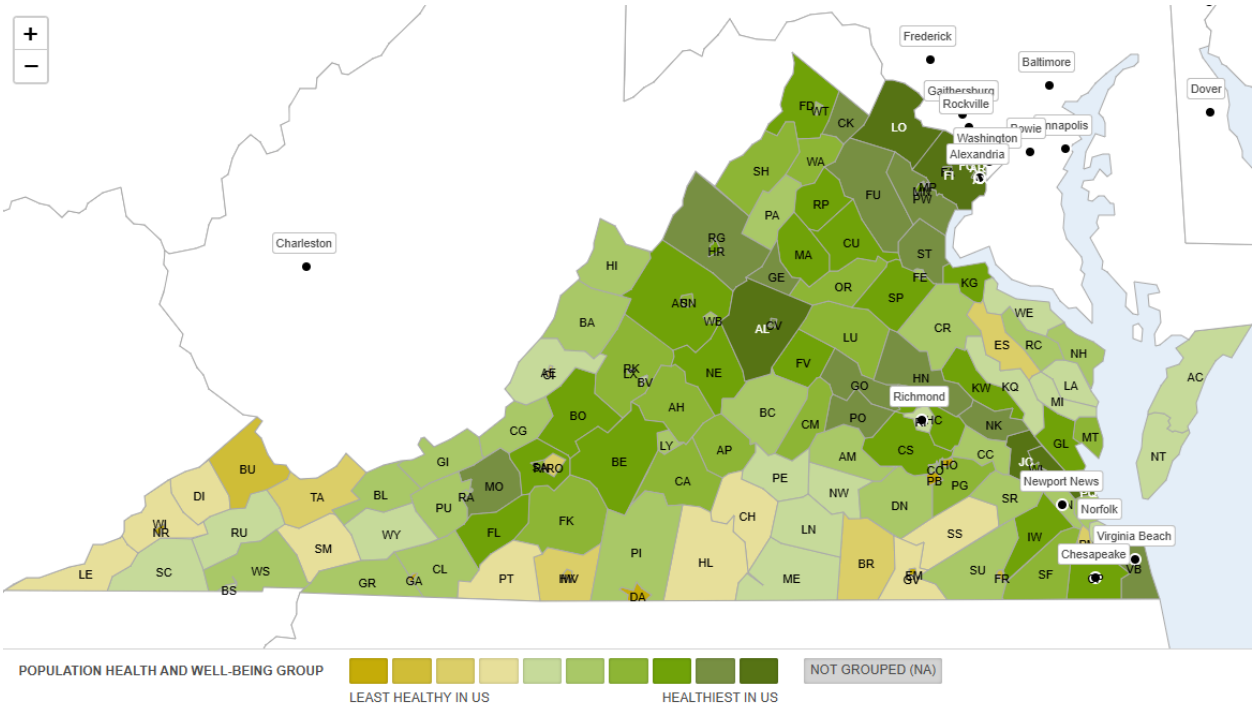
Virginia's Longevity Divide: Income, Obesity, and Exercise Access Across Counties

Team REGAN



Introduction

Although life expectancy is often regarded as a concise indicator of a community's overall well-being, notable disparities still exist even within a single state. Virginia's 133 counties and independent cities have an average life expectancy of just over seventy-seven years, according to the 2025 County Health Rankings report, highlighting disparities in the state's progress toward public health objectives (County Health Rankings & Roadmaps, 2025). Three recurrent factors that influence longevity are household economic resources, the burden of chronic diseases (obesity), and the characteristics of local environments (such as rural versus urban areas), according to decades of epidemiological research, including 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). Understanding how these factors manifest at the county level in Virginia can guide targeted interventions and equitable allocation of resources.

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?

Drawing on publicly available data from the County Health Rankings portal, we assembled a dataset in which each row represents a single jurisdiction, and the outcome of interest is the average life expectancy in years.

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 extract from the *County Health Rankings & Roadmaps program*, a collaborative effort between the University of Wisconsin Population Health Institute and the Robert Wood Johnson Foundation that gathers county-level health indicators nationwide each year. Before harmonizing the series to a single 2025 reference year, the institute gathers each metric directly from approved federal sources, including the Bureau of Labor Statistics, the American Community Survey, and the CDC WONDER database.

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.

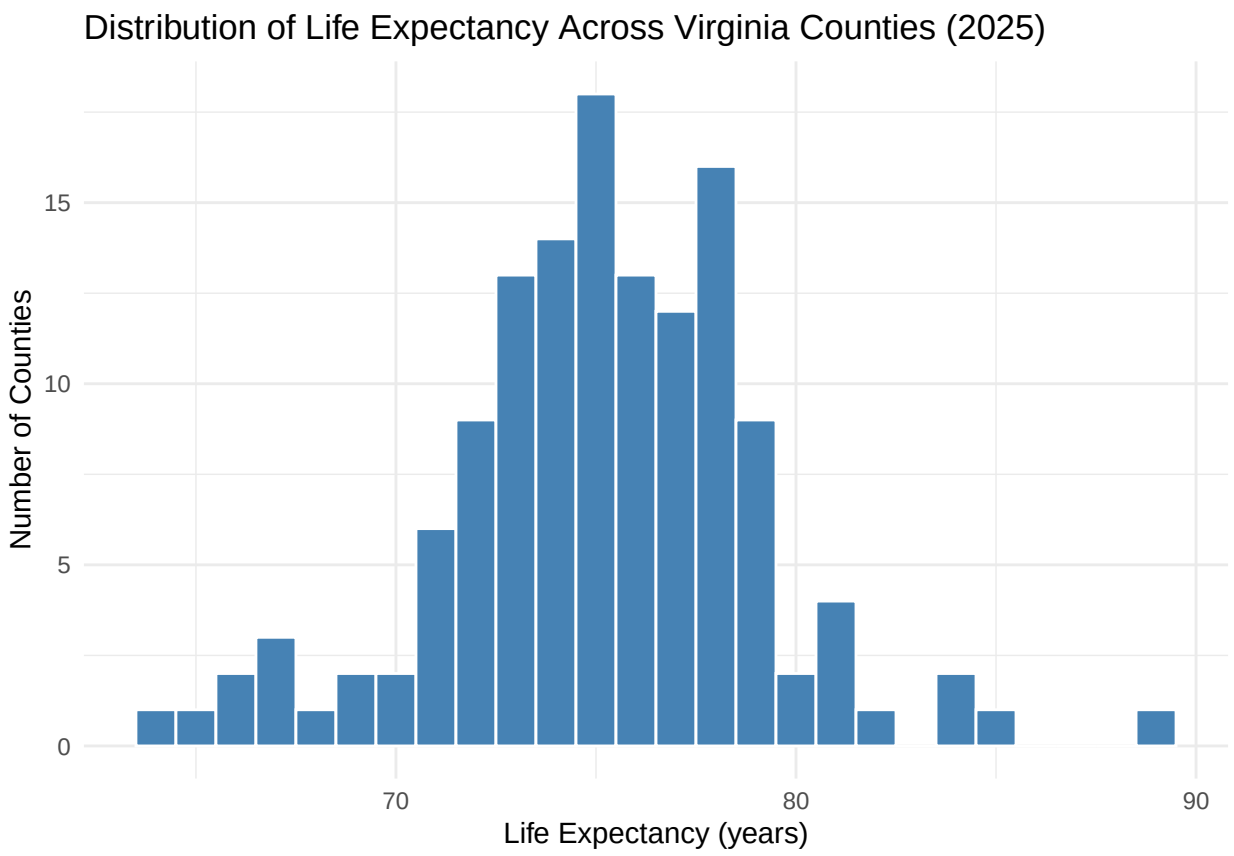
Several small, logically motivated changes were necessary. The ratio of primary care physicians was provided as a text string, such as "2,210:1." To ensure that lower figures indicate increased provider availability, it was transformed to a straightforward numerical count of residents per physician and then split into 5 bins "<=1k", "1-2k", "2-3k", "3-5k", ">=5k". 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 occasional entries with missing outcome values and one jurisdiction without a county name, 132 observations were included in the study sample.

The County Health Rankings have a high level of credibility, as they are frequently referenced in peer-reviewed health services research and utilize open documentation of data provenance and imputation procedures. However, three cautions are worth mentioning. First, particular data show suppressed numbers for a limited number of counties, which can contribute to the uncertainty. Second, there is no set time for retrieval because several focus areas of data are gathered in different years. Lastly, measurements are obtained from a wide range of reliable sources, each with its own unique data collection techniques, and then compiled into sheets. These restrictions will be reviewed when evaluating model findings; however, they do not

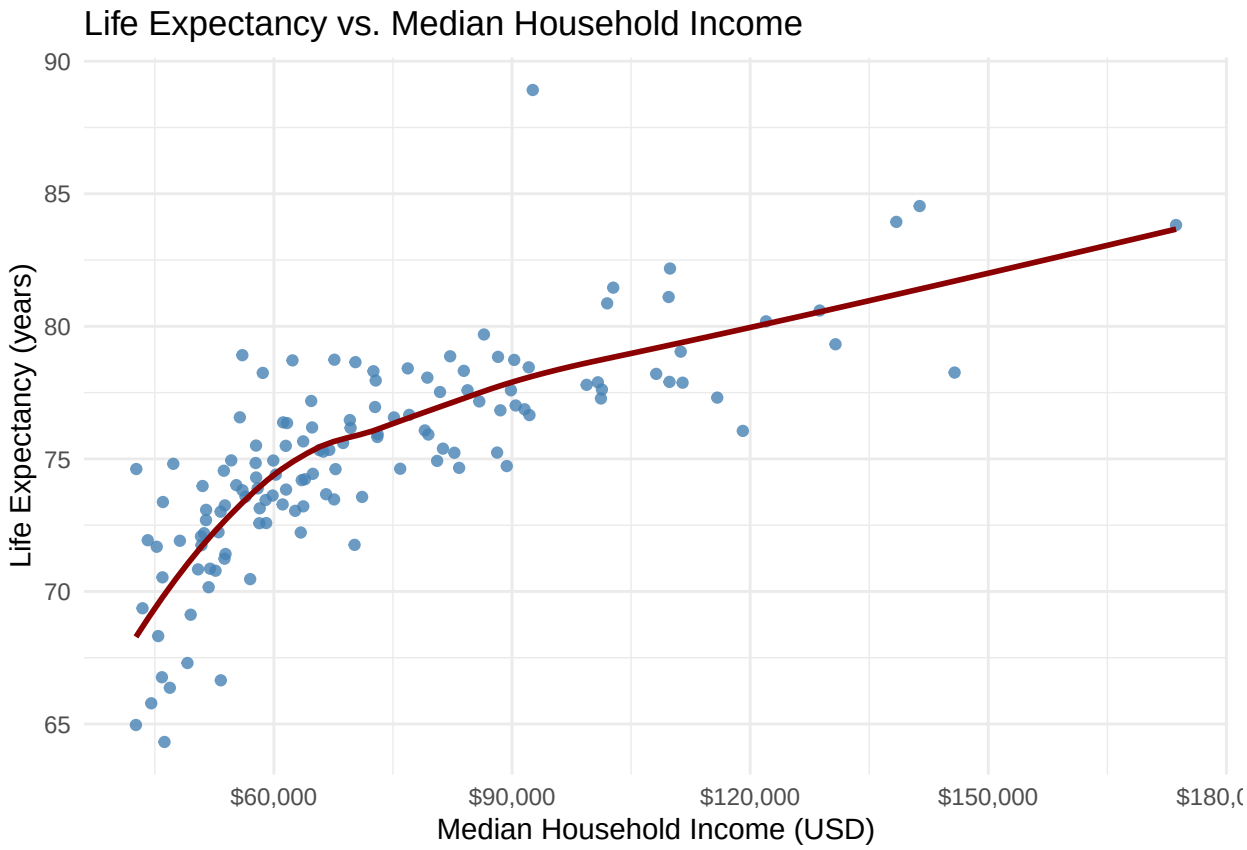
compromise the dataset's overall integrity.

Exploratory Data Analysis

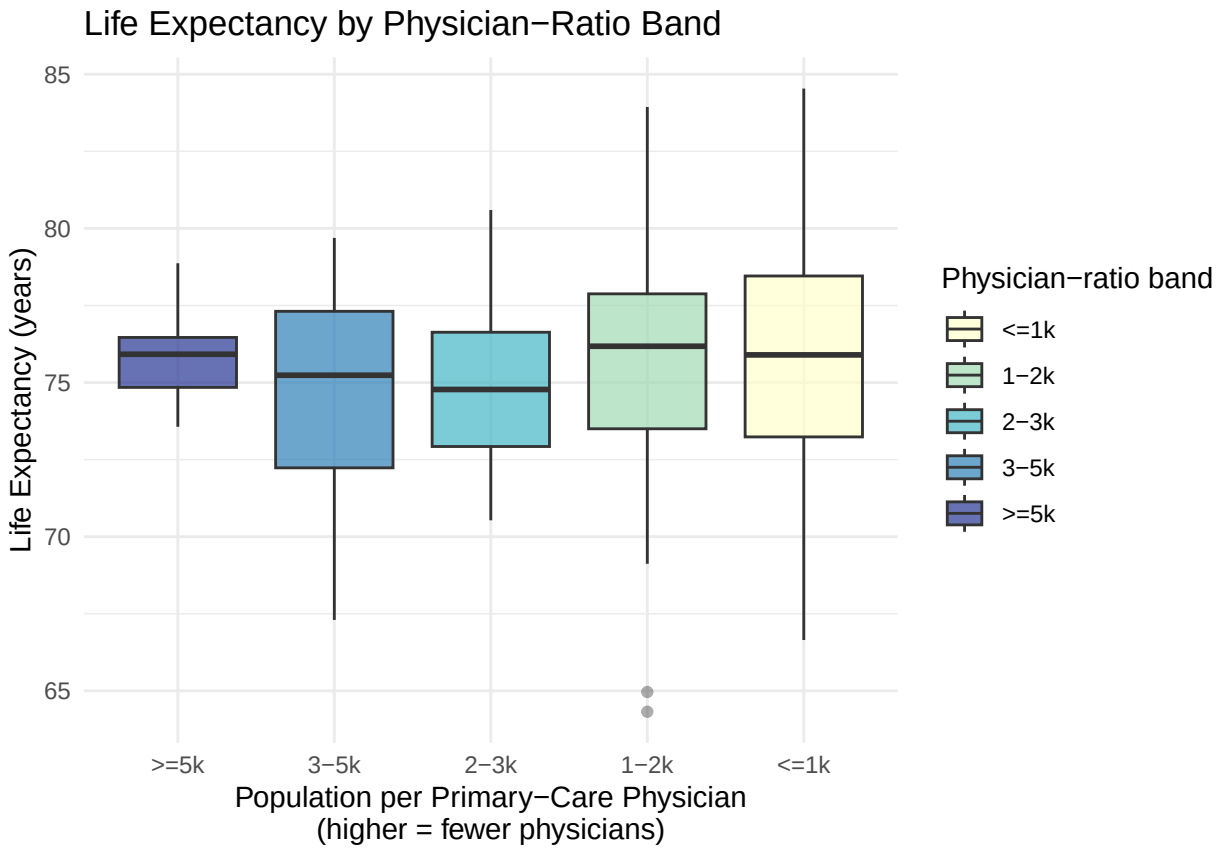
```
# For this section, i was unsure whether I should add in my summary (and detailed)
#statistics code I generated or leave it out. If you are wondering where my values
#come from in the next section, just know I generated them along with each plot
#and saved them in .csv files.

# 1. Histogram of Life Expectancy Counts (Counties)
addl_df <- addl_df %>%
  filter(!is.na(county), county != "NA")

p_hist <- ggplot(addl_df, aes(x = life_expectancy)) +
  geom_histogram(binwidth = 1,
                 fill     = "steelblue",
                 colour   = "white") +
  labs(title = "Distribution of Life Expectancy Across Virginia Counties (2025)",
       x     = "Life Expectancy (years)",
       y     = "Number of Counties") +
  theme_minimal()
print(p_hist)
```



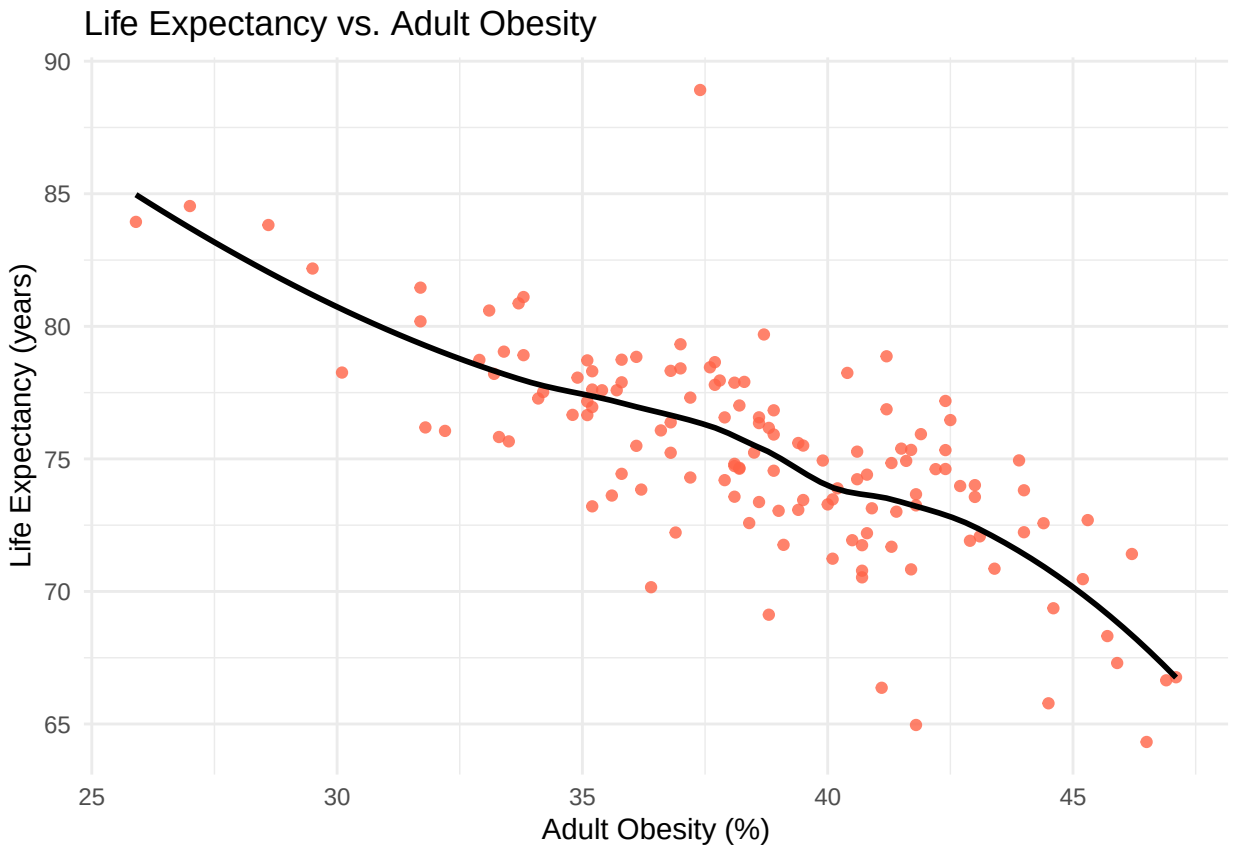

```
# 2. Scatterplot: Life Expectancy vs. Median Income (+ LOESS)
p_scatter <- ggplot(full_df,
                    aes(x = median_household_income,
                        y = life_expectancy)) +
  geom_point(colour = "steelblue", alpha = 0.8) +
  geom_smooth(method = "loess", se = FALSE, colour = "darkred") +
  scale_x_continuous(labels = dollar_format()) +
  labs(title = "Life Expectancy vs. Median Household Income",
       x = "Median Household Income (USD)",
       y = "Life Expectancy (years)") +
  theme_minimal()
print(p_scatter)
```



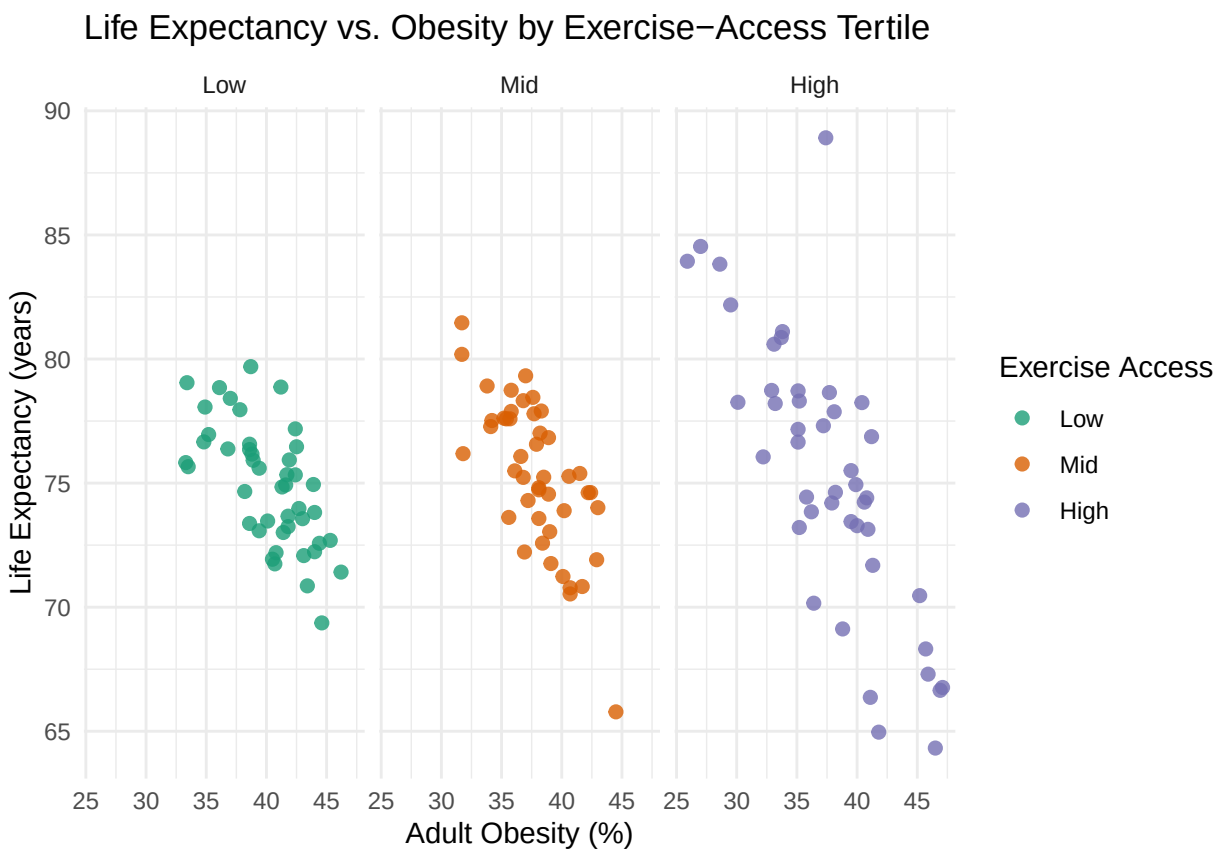

```
# 3. Scatterplot: Life Expectancy vs. Primary-Care Physician Ratio
plot_df <- full_df %>%
           filter(!is.na(phys_ratio_band))

p_phys <- ggplot(plot_df,
                 aes(x    = phys_ratio_band,
                     y    = life_expectancy,
                     fill = phys_ratio_band)) +
  geom_boxplot(alpha = 0.7, outlier.alpha = 0.4) +
  scale_x_discrete(limits = rev(levels(plot_df$phys_ratio_band))) +
  scale_fill_brewer(palette = "YlGnBu",
                    name     = "Physician-ratio band") +
  labs(title = "Life Expectancy by Physician-Ratio Band",
       x = "Population per Primary-Care Physician\n(higher = fewer physicians)",
       y = "Life Expectancy (years)") +
  theme_minimal()
plot(p_phys)
```




```
# 4. Scatterplot: Life Expectancy vs. Adult Obesity (%)
p_obesity <- ggplot(full_df,
                    aes(x = percent_adults_with_obesity,
                        y = life_expectancy)) +
  geom_point(colour = "tomato", alpha = 0.8) +
  geom_smooth(method = "loess", se = FALSE, colour = "black") +
  labs(title = "Life Expectancy vs. Adult Obesity",
       x = "Adult Obesity (%)",
       y = "Life Expectancy (years)") +
  theme_minimal()
print(p_obesity)
```

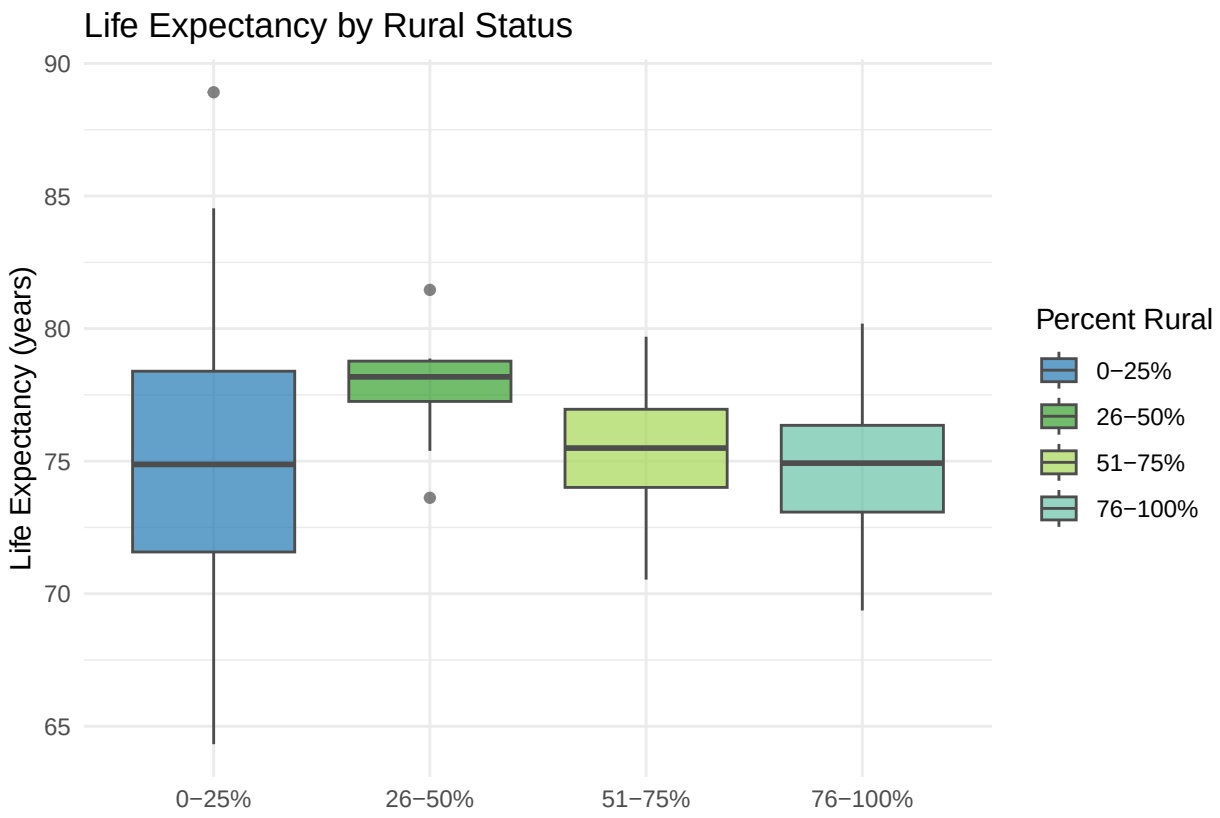



```
# 5. Faceted scatter: Life Expectancy ~ Adult Obesity, coloured by Exercise-Access Terti
p_exercise <- full_df %>%
  filter(!is.na(exercise_tertile)) %>%
  ggplot(aes(x = percent_adults_with_obesity,
             y = life_expectancy,
             colour = exercise_tertile)) +
  geom_point(size = 2, alpha = 0.8) +
  scale_colour_brewer(palette = "Dark2",
                      name = "Exercise Access",
                      na.translate = FALSE) +
  facet_wrap(~ exercise_tertile, nrow = 1, drop = TRUE) +
  labs(title = "Life Expectancy vs. Obesity by Exercise-Access Tertile",
       x = "Adult Obesity (%)",
       y = "Life Expectancy (years)") +
  theme_minimal()
print(p_exercise)
```



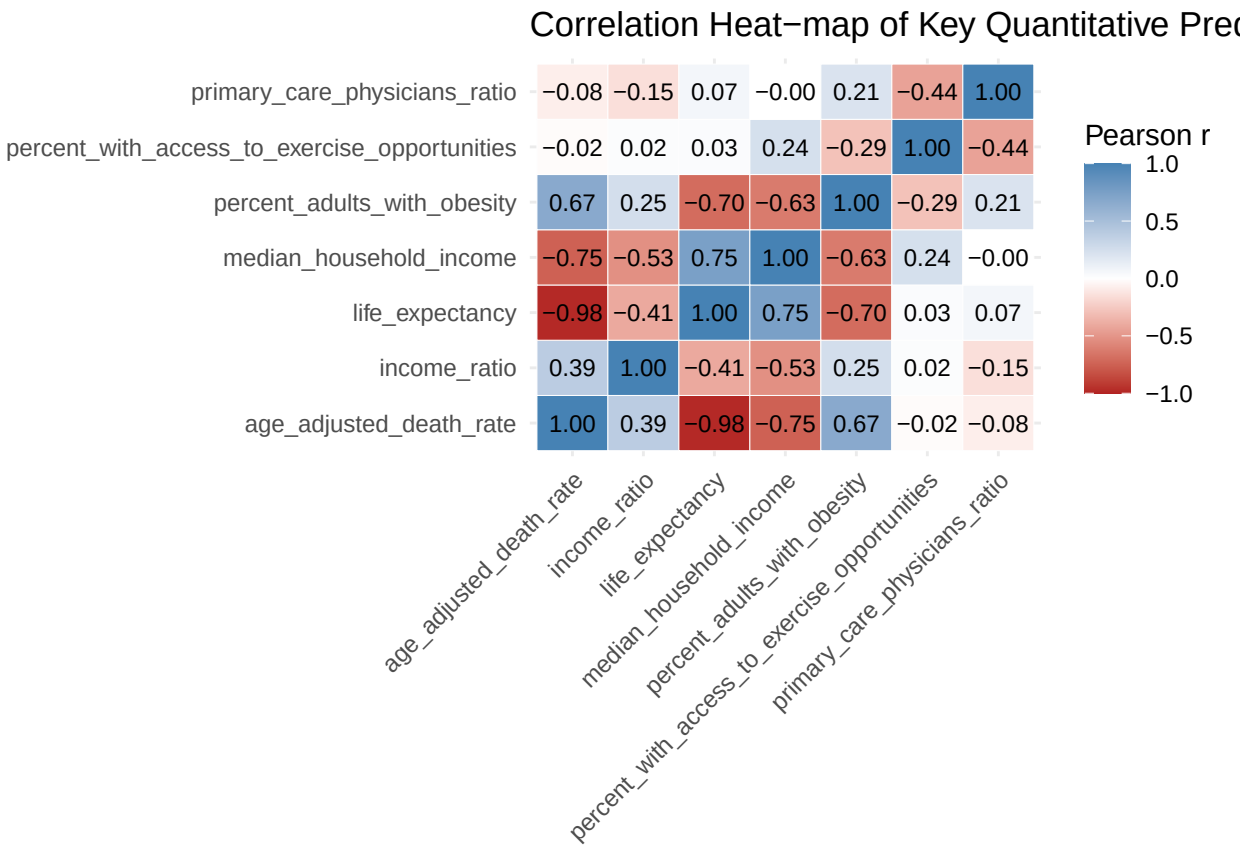

```
# 6. Boxplot: Life Expectancy by Rural
rural_cols <- c(
  "0-25%"   = "#1f78b4",
  "26-50%"  = "#33a02c",
  "51-75%"  = "#a6d854",
  "76-100%" = "#66c2a5"
)

df_rural <- full_df %>% filter(!is.na(rural_band))

p_rural <- ggplot(df_rural,
                  aes(x = rural_band,
                      y = life_expectancy,
                      fill = rural_band)) +
  geom_boxplot(alpha = 0.7, colour = "grey30") +
  scale_fill_manual(values = rural_cols, name = "Percent Rural") +
  labs(title = "Life Expectancy by Rural Status",
```

```
       x = "",
       y = "Life Expectancy (years)") +
  theme_minimal()
print(p_rural)
```




```
# 7. Correlation Heat-map of Key Quantitative Predictors
num_vars <- full_df %>%
  select(life_expectancy,
         median_household_income,
         percent_adults_with_obesity,
         primary_care_physicians_ratio,
         age_adjusted_death_rate,
         income_ratio,
         percent_with_access_to_exercise_opportunities) %>%
  drop_na()

corr_mat  <- round(cor(num_vars), 2)
```

```
corr_long <- corr_mat %>%
  as.data.frame() %>%
  rownames_to_column("Var1") %>%
  pivot_longer(-Var1, names_to = "Var2", values_to = "Pearson_r") %>%
  arrange(Var1, Var2)

p_corr <- ggplot(corr_long, aes(Var1, Var2, fill = Pearson_r)) +
  geom_tile(colour = "white") +
  geom_text(aes(label = sprintf("%.2f", Pearson_r)), size = 3) +
  scale_fill_gradient2(low = "firebrick", mid = "white", high = "steelblue",
                       midpoint = 0, limits = c(-1, 1), name = "Pearson r") +
  labs(title = "Correlation Heat-map of Key Quantitative Predictors",
       x = NULL, y = NULL) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
print(p_corr)
```



EDA Summary

The response variable, life expectancy at birth, has a mean of 75.15 years, a standard deviation of only 3.89 years, and a bell-shaped distribution that is almost symmetric with a slight skew (–0.04). The variable fulfills the normality and homoscedasticity assumptions typically required for multiple linear regression, as it spans more than twenty-four years (64.3–88.9); however, the majority of counties cluster firmly between 73 and 78 years. As a result, no transformation seems to be required.

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.

When considered collectively, the exploratory study demonstrates that multiple linear regression with life expectancy as the continuous response is appropriate. Additionally, it supports a principled variable-screening approach that limits the age-adjusted death rate for conceptual and statistical reasons, monitors multicollinearity primarily between income and obesity, and keeps income, obesity, exercise access, rural band, and physician ratio as core predictors. The finished model will be well-positioned to explain why some Virginians live noticeably longer than others, as it will capture both quantitative gradients and categorical contexts.

Methods and Analysis

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