The Impact of Income and Substance Use on Health Outcomes in US Counties

Using Data from CHR 2023

2025-02-25

```
knitr::opts_chunk$set(warning = FALSE)
```

```
library(Hmisc)
library(janitor)
library(naniar)
library(sessioninfo)
library(car)
library(mosaic)
suppressPackageStartupMessages(library(here))
library(gt)
library(glue)
library(patchwork)
library(broom)
library(rstatix)
library(tidyverse)



theme_set(theme_bw())
knitr::opts_chunk$set(comment = NA)
```

Data Ingest

```
data_url <- "https://www.countyhealthrankings.org/sites/default/files/media/document/analyti
```

```
chr_2023_raw <- read_csv(data_url, skip = 1, guess_max = 4000,show_col_types = FALSE)


chr_2023_raw <- chr_2023_raw %>% filter(county_ranked == 1)

dim(chr_2023_raw)
```

```
[1] 3082  720
```

State Selection

The goal of the analysis will include all of the counties in the Unitied States of America.

```
chr_2023 <- chr_2023_raw %>% mutate(state = factor(state))

chr_2023 %>% count(state)
```

```
# A tibble: 51 x 2
   state     n
   <fct> <int>
 1 AK       24
 2 AL       67
 3 AR       75
 4 AZ       15
 5 CA       58
 6 CO       59
 7 CT        8
 8 DC        1
 9 DE        3
10 FL       67
# i 41 more rows
```

```
nrow(chr_2023)
```

```
[1] 3082
```

There are 3082 total counties in this data set.

Variable Selection

```
chr_2023 <- chr_2023 %>% select(fipscode,
                                county,
                                state,
                                county_ranked,
                                v002_rawvalue,
                                v063_rawvalue,
                                v011_rawvalue,
                                v009_rawvalue,
                                v143_rawvalue)
```

The selected variables for analysis include poor or fair health, median household income, adult obesity, adult smoking, and insufficient sleep. In Analysis 1, median household income serves as the predictor for the outcome variable poor or fair health. Analysis 2 examines adult smoking as a predictor of adult obesity, while Analysis 3 focuses on insufficient sleep as the primary variable of interest.

Variable Cleaning and Renaming

The selected variables were renamed to align with the data dictionary:

```
chr_2023 <- chr_2023 %>% rename(Poor_or_fair_health = v002_rawvalue,
                    Median_household_income = v063_rawvalue,
                    Adult_obesity = v011_rawvalue,
                    adult_smoking = v009_rawvalue,
                    insufficent_sleep = v143_rawvalue) %>% mutate(Poor_or_fair_health = Poor_
                                                                  Median_household_income = M
                                                                  Adult_obesity = Adult_obesi
                                                                  adult_smoking = adult_smoki
                                                                  insufficent_sleep = insuffi
```

All variables, except for Median Household Income, were multiplied by 100 to convert their values from proportions to percentages. Median Household Income was divided by 1,000 to express it in thousands of dollars for better readability.

Creating the Analysis 2 Predictor

The percentage of adult smokers in a county will be categorized into three groups: “Low”, “Medium”, and “High” based on quantile cutoffs.

- “Low”: Counties in the bottom 40% (18.9%)
- “Medium”: Counties in the middle 20% (between 18.9% and 20.9%)
- “High”: Counties in the top 40% (20.9%)

```r
# Compute quantile cutoffs
chr_2023 %>%
  summarise(q40 = quantile(adult_smoking, c(0.4), na.rm = TRUE),  # Bottom 40% cutoff
            q60 = quantile(adult_smoking, c(0.6), na.rm = TRUE))  # Top 40% cutoff
```

```
# A tibble: 1 x 2
    q40   q60
  <dbl> <dbl>
1  18.9    21
```

```r
# Categorize adult smoking into "Low", "Medium", and "High" groups
chr_2023 <- chr_2023 %>%
  mutate(adult_smoking_grp = case_when(
    adult_smoking <= 18.9 ~ "Low",
    adult_smoking > 18.9 & adult_smoking < 20.9 ~ "Medium",
    adult_smoking >= 20.9 ~ "High")) %>%
  mutate(adult_smoking_grp = factor(adult_smoking_grp, levels = c("Low", "Medium", "High")))

# Count occurrences in each category
chr_2023 %>% count(adult_smoking_grp)
```

```
# A tibble: 4 x 2
  adult_smoking_grp     n
  <fct>             <int>
1 Low                1233
2 Medium              571
3 High               1277
4 <NA>                  1
```

The majority of counties fall into the ‘Low’ (1,233 counties) and ‘High’ (1,277 counties) smoking groups, while fewer counties (571) are classified as ‘Medium.’ Only one county has missing data.

Adding 2018 Data for the Analysis 3 Outcome

Analysis will be done with the variable insufficient sleep to determine if the sleep quality has worsened for these counties between 2018 and 2023. The 2018 data is read in with read_csv() and is joined to the original data set using the function left join().

```
chr_2018_raw <- read_csv(suppressWarnings(here("data/raw/chr_2018.csv")), guess_max = 4000, 


chr_2018_raw <- chr_2018_raw %>% mutate(fipscode = as.character(fipscode))

chr_2018 <- chr_2018_raw %>% select(fipscode, v143_rawvalue)

chr_2023 <- left_join(chr_2023, chr_2018, by = "fipscode")

chr_2023 <- chr_2023 %>% rename(insufficent_sleep_2018 = v143_rawvalue) %>%
                         mutate(insufficent_sleep_2018 = insufficent_sleep_2018 * 100) %>%
                         rename(insufficent_sleep_2023 = insufficent_sleep)
```

Arranging and Saving the Analytic Tibble

```
chr_2023 <- chr_2023 %>% select(fipscode, state, county, Poor_or_fair_health, Median_househol

write_csv(chr_2023, here("data/processed/chr_2023.csv"))
saveRDS(chr_2023, here("data/processed/chr_2023.rds"))
```

Print the Tibble

The tibble is printed to confirm that:

1. It is a tibble (instead of a data frame or another object type).
2. All variables have the correct data types, ensuring compatibility with the analysis.

```
chr_2023
```

```
# A tibble: 3,082 x 11
   fipscode state county          Poor_or_fair_health Median_household_income
   <chr>    <fct> <chr>                         <dbl>                   <dbl>
```

```
 1 01001    AL    Autauga County                 16.9                    66.4
 2 01003    AL    Baldwin County                 14.9                    65.7
 3 01005    AL    Barbour County                 27.5                    38.6
 4 01007    AL    Bibb County                    21.6                    48.5
 5 01009    AL    Blount County                  18.4                    56.9
 6 01011    AL    Bullock County                 29.7                    32.0
 7 01013    AL    Butler County                  22.7                    39.4
 8 01015    AL    Calhoun County                 19.6                    48.2
 9 01017    AL    Chambers County                21.5                    45.4
10 01019    AL    Cherokee County                19.3                    46.4
# i 3,072 more rows
# i 6 more variables: Adult_obesity <dbl>, adult_smoking_grp <fct>,
#   adult_smoking <dbl>, insufficent_sleep_2023 <dbl>,
#   insufficent_sleep_2018 <dbl>, county_ranked <dbl>
```

Expected data types:

- `fipscode` and `county` → Character (chr)
- `state` and `adult_smoking_grp` → Factor (fct)
- All other quantitative variables → Double (dbl)

Numerical Summaries

```r
describe(chr_2023)
```

```
chr_2023

 11  Variables      3082  Observations
--------------------------------------------------------------------------------
fipscode
       n  missing distinct
    3082        0     3082

lowest : 01001 01003 01005 01007 01009, highest: 56037 56039 56041 56043 56045
--------------------------------------------------------------------------------
state
       n  missing distinct
    3082        0       51

lowest : AK AL AR AZ CA, highest: VT WA WI WV WY
```

```
--------------------------------------------------------------------------------
county
       n  missing distinct
    3082        0     1846

lowest : Abbeville County          Acadia Parish             Accomack County          Ada Co
highest: Yukon-Koyukuk Census Area Yuma County               Zapata County            Zavala
--------------------------------------------------------------------------------
Poor_or_fair_health
       n  missing distinct     Info     Mean      Gmd      .05      .10
    3081        1      228        1     16.1    4.942     10.2     11.0
     .25      .50      .75      .90      .95
    12.7     15.4     19.0     22.0     24.0

lowest : 6.5  7    7.4  7.7  8   , highest: 33   33.4 34   36.5 36.8
--------------------------------------------------------------------------------
Median_household_income
       n  missing distinct     Info     Mean      Gmd      .05      .10
    3081        1     2982        1    58.91    16.17    39.18    42.52
     .25      .50      .75      .90      .95
   48.94    56.63    65.69    77.44    87.55

lowest : 25.653  28.236  28.72   29.05   29.143
highest: 133.845 134.05  141.161 142.43  153.716
--------------------------------------------------------------------------------
Adult_obesity
       n  missing distinct     Info     Mean      Gmd      .05      .10
    3081        1      275        1    36.27     5.11     28.0     30.1
     .25      .50      .75      .90      .95
    33.7     36.7     39.3     41.4     42.9

lowest : 17.6 19   19.1 19.5 19.9, highest: 49.9 50.9 51.2 51.3 53.2
--------------------------------------------------------------------------------
adult_smoking_grp
       n  missing distinct
    3081        1        3

Value         Low Medium   High
Frequency    1233    571   1277
Proportion  0.400  0.185  0.414
--------------------------------------------------------------------------------
adult_smoking
       n  missing distinct     Info     Mean      Gmd      .05      .10
    3081        1      236        1    20.09    4.566     13.4     15.1
     .25      .50      .75      .90      .95
    17.5     19.9     22.7     25.3     26.8

lowest : 6.7  7    7.8  8    8.4 , highest: 34.8 36   39.5 40.5 41.1
--------------------------------------------------------------------------------
insufficent_sleep_2023
       n  missing distinct     Info     Mean      Gmd      .05      .10
    3081        1      205        1    34.55    4.086     28.7     29.9
     .25      .50      .75      .90      .95
    32.0     34.5     36.9     39.3     40.8

lowest : 23.8 24   24.7 24.8 24.9, highest: 46   46.3 46.5 46.9 48.4
--------------------------------------------------------------------------------
insufficent_sleep_2018
       n  missing distinct     Info     Mean      Gmd      .05      .10
    2768      314     2768        1    33.15     4.63    26.62    27.73
     .25      .50      .75      .90      .95
   30.22    33.06    36.14    38.42    39.92

lowest : 23.2268 23.3776 23.5292 23.5501 23.6942
highest: 44.8874 45.0583 45.2767 45.4021 46.7078
--------------------------------------------------------------------------------
county_ranked
       n  missing distinct     Info     Mean      Gmd
    3082        0        1        0        1        0

Value         1
Frequency  3082
Proportion    1
--------------------------------------------------------------------------------
```

The Codebook

The **chr_2023** tibble contains *3082* counties and *11* variables.

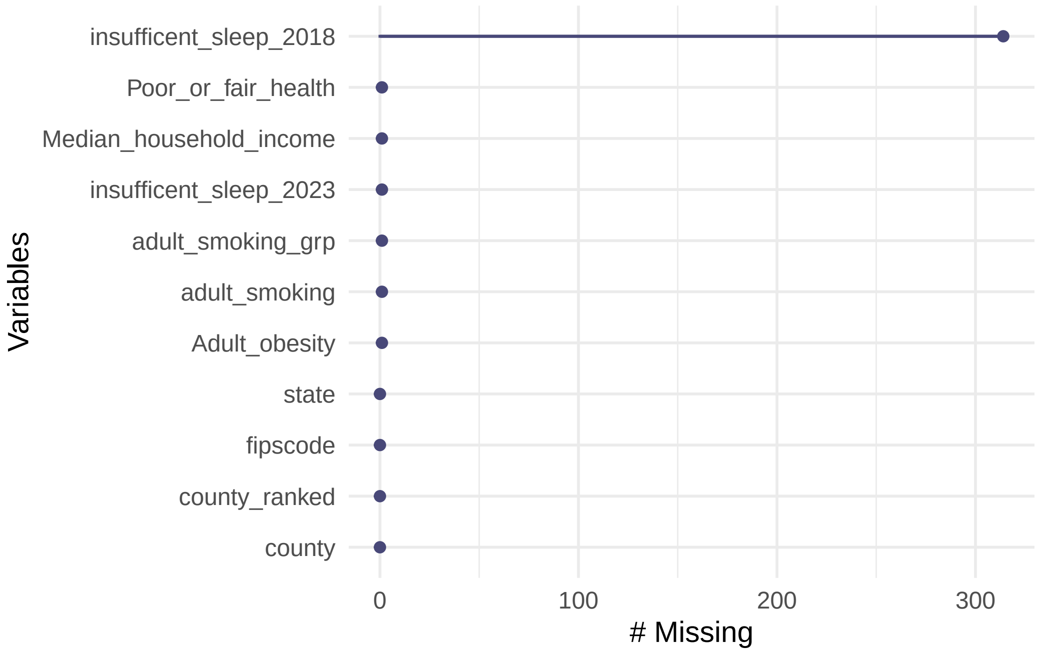
Variable	Original	Role	NA	Distinct	Definition	Source	Year(s)
fipscode	–	ID	0	611	county's FIPS code	–	–
state	–	ID	0	6	state postal abbreviation	–	–
county	–	ID	0	529	county name	–	–

Variable	Original	Role	NA	Distinct	Definition	Source	Year(s)
Poor_or_fair_health	v002_rawvalue	Outcome Variable	0	169	Percentage of adults reporting fair or poor health	The Behavioral Risk Factor Surveillance System	2020
Median_household_income	v063_rawvalue	Predictor Variable	0	608	The median income of a household in thousands of dollars	US Census Bureau	2021
Adult_obesity	v011_rawvalue	Outcome Variable	0	160	Percentage of the adult population that reports a body mass index (BM) greater than or equal to 30kg/m2	The Behavioral Risk Factor Surveillance System	2020
adult_smoking_group	adult_smoking	Predictor Variable	119	2	A binary variable that divides the percentage of adult smokers into high and low groups	The Behavioral Risk Factor Surveillance System	2020
adult_smoking	v009_rawvalue	Predictor variable	0	175	Percentage of adult who are current smokers	The Behavioral Risk Factor Surveillance System	2020

Variable	Original	Role	NA	Distinct	Definition	Source	Year(s)
insufficient_sleep_2023	v143_rawvalue	Outcome variable	0	138	The percentage of adults who report fewer than 7 hours of sleep on average in 2023	The Behavioral Risk Factor Surveillance System	2020
insufficient_sleep_2018	v143_rawvalue	Outcome variable	60	551	The percentage of adults who report fewer than 7 hours of sleep on average in 2018	The Behavioral Risk Factor Surveillance System	2016
county_ranked	–	Check	0	1	Ensuring all values are 1	–	–

Missingness Check

```
gg_miss_var(chr_2023)
```



```
# Calculate the percentage of missing values for each variable
chr_2023 %>%
  summarise(missing_percent = mean(is.na(insufficent_sleep_2018)) * 100)
```

```
# A tibble: 1 x 1
  missing_percent
            <dbl>
1            10.2
```

All quantitative variables have no missing data, except for `insufficient_sleep_2018`, which has approximately 10% missing values. However, this is well below the 20% threshold, making it unlikely to impact the analysis significantly.

Distinct Values Check

```
sapply(chr_2023, function(x) n_distinct(x))
```

```
               fipscode                   state                  county
                   3082                      51                    1846
    Poor_or_fair_health Median_household_income           Adult_obesity
                    229                    2983                     276
      adult_smoking_grp           adult_smoking  insufficent_sleep_2023
```

```
                      4                      237                      206
 insufficent_sleep_2018            county_ranked
                   2769                        1
```

All of the quantitative variables have more than 15 distinct values, with the smallest number of smallest distinct values being 138 for insufficient sleep 2023.

Research Questions

Analysis 1 Research Question

What is the relationship between median household income and health outcomes across all U.S. counties? Counties with lower household incomes may have reduced access to healthcare and fewer economic opportunities, potentially leading to worse health outcomes.

Analysis 2 Research Question

Is there an association between adult smoking rates and obesity prevalence? Since nicotine is a stimulant that suppresses appetite, we hypothesize that counties with higher smoking rates will have lower obesity rates.

Analysis 3 Research Question

How has the average sleep quality in U.S. counties changed from 2018 to 2023?

Analysis 1

Variables

Two quantitative variables will be used to conduct the analysis and determine the association between them. The variables being used are `Median_household_income` as the predictor variable and `Poor_or_fair_health` as the outcome variable.`Median_household_income` represents the median reported income of households (in thousands of dollars) for a given county. `Poor_or_fair_health` is the percentage of adults self-reporting either fair or poor health in response to the question: "Would you say that in general your health is Excellent, Very Good, Good, Fair, or Poor?" The data set includes all counties in the United States.

Summaries

Numerical Summaries of Household income and poor of fair health outcomes:

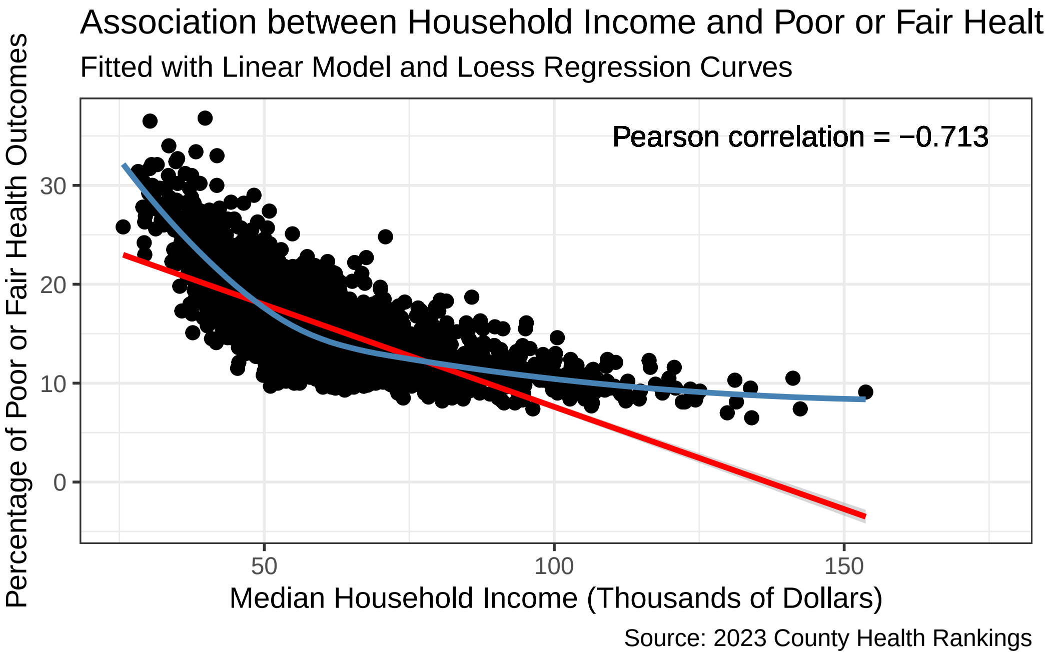
```r
s1 <- mosaic::favstats(~ Median_household_income, data = chr_2023)
s2 <- mosaic::favstats(~ Poor_or_fair_health, data = chr_2023)
s3 <- bind_rows(list(Median_Household_Income = s1, Poor_or_fair_health = s2), .id = "id")
s3 <- s3 %>% rename(Variable = id)
s3 %>% gt() %>% fmt_number(decimals = 2)
```

Variable	min	Q1	median	Q3	max	mean	sd	n	missi
Median_Household_Income	25.65	48.94	56.63	65.69	153.72	58.91	15.29	3,081.00	1.
Poor_or_fair_health	6.50	12.70	15.40	19.00	36.80	16.10	4.43	3,081.00	1.

The summary statistics for the data set provide insight into the distribution of median household income and self-reported poor or fair health across 3,081 U.S. counties. The median household income ranges from $25.65K to $153.72K, with a median of $56.63K and a mean of $58.91K, suggesting a right-skewed distribution, where some counties have significantly higher incomes. The standard deviation of $15.29K indicates moderate variation in income levels across counties.

```r
# Compute Pearson correlation
income_cor <- glue("Pearson correlation = {round(cor(chr_2023$Median_household_income,
                                                      chr_2023$Poor_or_fair_health, use = 'c

# Plot
ggplot(chr_2023, aes(x = Median_household_income, y = Poor_or_fair_health)) +
  geom_point(size = 2) +
  geom_smooth(method = "lm", formula = y ~ x, col = "red", se = TRUE) +
  geom_smooth(method = "loess", formula = y ~ x, col = "steelblue", se = FALSE) +
  theme_bw() +
  geom_text(aes(x = 175, y = 35, label = income_cor), hjust = 1, size = 4) +
  labs(title = "Association between Household Income and Poor or Fair Health Outcomes",
       subtitle = "Fitted with Linear Model and Loess Regression Curves",
       x = "Median Household Income (Thousands of Dollars)",
       y = "Percentage of Poor or Fair Health Outcomes",
       caption = "Source: 2023 County Health Rankings")
```



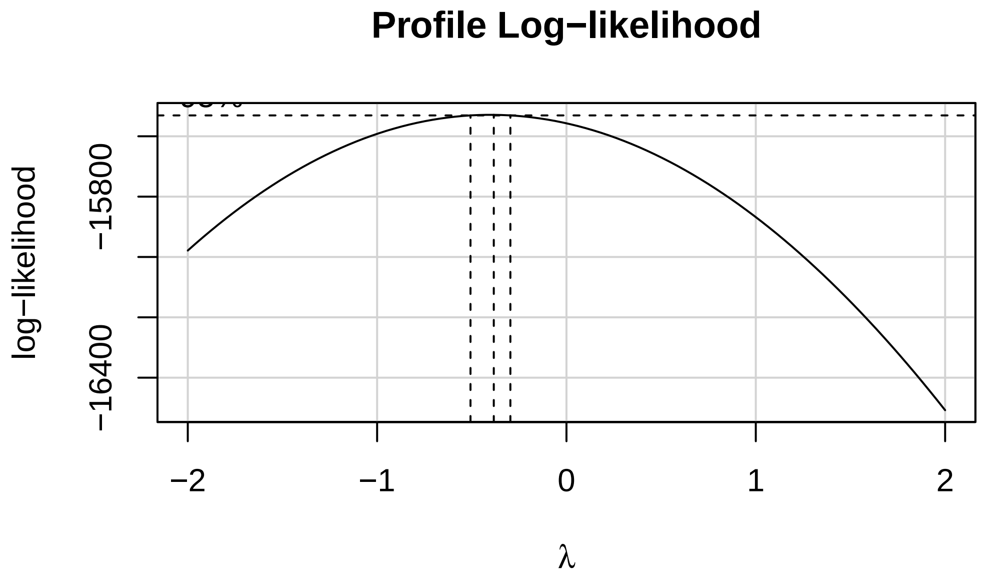
Source: 2023 County Health Rankings

The "loess" smooth curve deviates significantly from the linear regression model, suggesting that a linear relationship may not fully capture the trend in the data. The non-linearity is particularly evident at lower income levels, where the decline in poor or fair health percentages is steep, before flattening out at higher income levels. This indicates that a simple linear model may not be the best fit for this data.

To address this, a transformation will be applied to the predictor variable (`Median_Household_Income`), and potentially the outcome variable (`Poor_or_fair_health`), to improve linearity. A Box-Cox transformation plot will be used to determine an appropriate transformation approach.

Additionally, the strong negative Pearson correlation coefficient (-0.713) suggests a moderate-to-strong inverse relationship between median household income and poor health outcomes, meaning counties with higher median incomes tend to report fewer cases of poor or fair health.

```r
boxCox(chr_2023$Poor_or_fair_health ~ chr_2023$Median_household_income)
```



```
powerTransform(chr_2023$Poor_or_fair_health ~ chr_2023$Median_household_income)
```

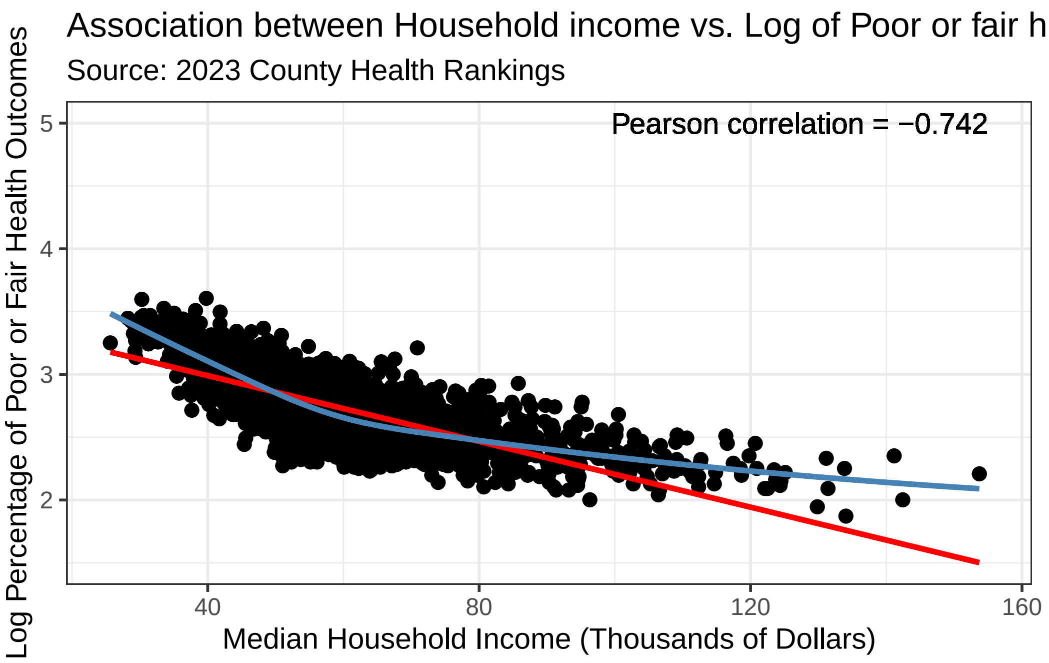
```
Estimated transformation parameter
        Y1
-0.4012156
```

Based on the Box-Cox plot and the power transformation equation, a lambda value of -0.33 was obtained, suggesting that the optimal transformation for the predictor variable is *1/sqrt(y)*. However, to maintain interpret ability and consistency, this analysis will be constrained to commonly used transformations, such as inverse, logarithmic, or square root transformations. Given that logarithmic transformation is the closest alternative to the suggested power transformation, a log transformation will be applied to the Poor_or_fair_health outcome variable.

```
# Compute Pearson correlation
income_cor <- glue("Pearson correlation = {round(cor(chr_2023$Median_household_income,
                                                      log(chr_2023$Poor_or_fair_health), use

ggplot(chr_2023, aes(x = Median_household_income, y = log(Poor_or_fair_health))) +
  geom_point(size = 2) +
  geom_smooth(method = "lm", formula = y ~ x, col = "red", se = FALSE) +
  geom_smooth(method = "loess", formula = y~x, col = "steelblue", se = FALSE) +
  geom_text(aes(x = 155, y = 5, label = income_cor), hjust = 1, size = 4) +
```

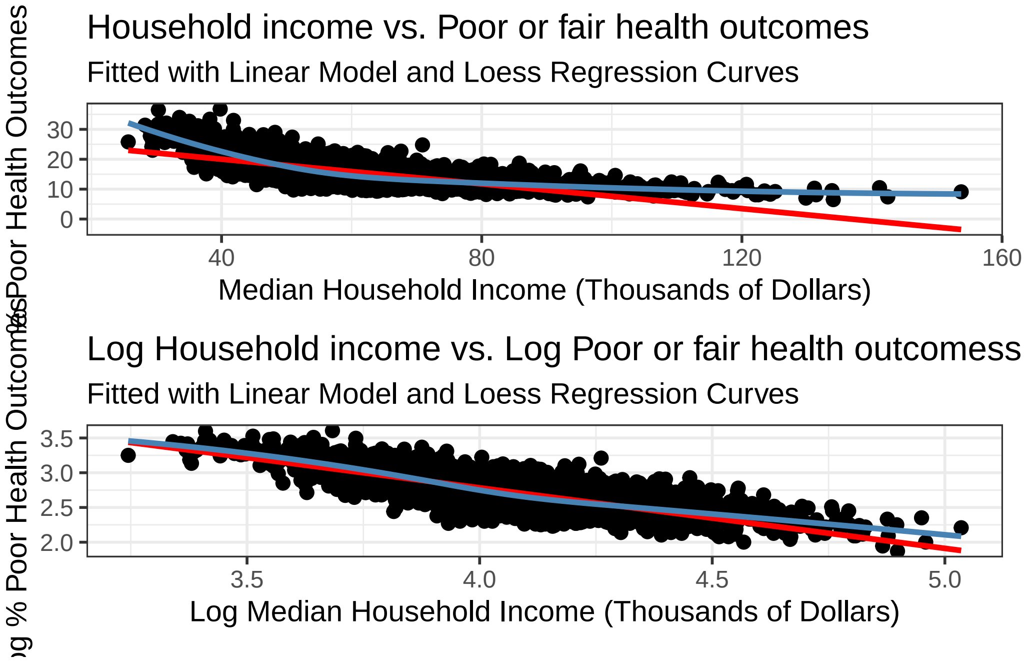
```r
  theme_bw() +
  labs(title = "Association between Household income vs. Log of Poor or fair health outcomes"
       x = "Median Household Income (Thousands of Dollars)",
       y = "Log Percentage of Poor or Fair Health Outcomes",
       subtitle = "Source: 2023 County Health Rankings")
```



```r
p1 <- ggplot(chr_2023, aes(x = Median_household_income, y = Poor_or_fair_health)) + geom_poi
  geom_smooth(method = "lm", formula = y ~ x, col = "red", se = FALSE) +
  geom_smooth(method = "loess", formula = y~x, col = "steelblue", se = FALSE) + theme_bw() +
  labs(title = "Household income vs. Poor or fair health outcomes",
       subtitle = "Fitted with Linear Model and Loess Regression Curves",
       x = "Median Household Income (Thousands of Dollars)",
       y = "% Poor Health Outcomes")

p2 <- ggplot(chr_2023, aes(x = log(Median_household_income), y = log(Poor_or_fair_health)))
  geom_smooth(method = "lm", formula = y ~ x, col = "red", se = FALSE) +
  geom_smooth(method = "loess", formula = y~x, col = "steelblue", se = FALSE) + theme_bw() +
  labs(title = "Log Household income vs. Log Poor or fair health outcomess",
        subtitle = "Fitted with Linear Model and Loess Regression Curves",
       x = "Log Median Household Income (Thousands of Dollars)",
       y = "Log % Poor Health Outcomes")
```

```
p1 / p2
```



The log transformation of the `Poor_or_fair_health` outcome variable has improved the linearity of the relationship between median household income and self-reported poor or fair health outcomes. Compared to the original plot, the loess curve (blue) and the linear regression line (red) now align more closely, indicating that the transformed model better captures the underlying trend.

The negative Pearson correlation coefficient (-0.713) remains the same, reaffirming a moderate-to-strong inverse relationship between income and poor health outcomes. However, with the log transformation, the spread of values appears more consistent across different income levels, potentially reducing heteroscedasticity in the data. This transformation ensures that the linear regression model assumptions are better met, allowing for a more accurate estimation of the relationship between household income and health outcomes. Further diagnostic checks will confirm whether additional refinements are necessary.

Model

```
m1 <- lm(log(Poor_or_fair_health) ~ log(Median_household_income), data = chr_2023)
```

```r
tidy(lm(log(Poor_or_fair_health) ~ log(Median_household_income), data = chr_2023),conf.int =
  gt() %>% fmt_number(decimals = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	6.264	0.051	123.920	0.000	6.181	6.347
log(Median_household_income)	-0.870	0.012	-69.794	0.000	-0.891	-0.850

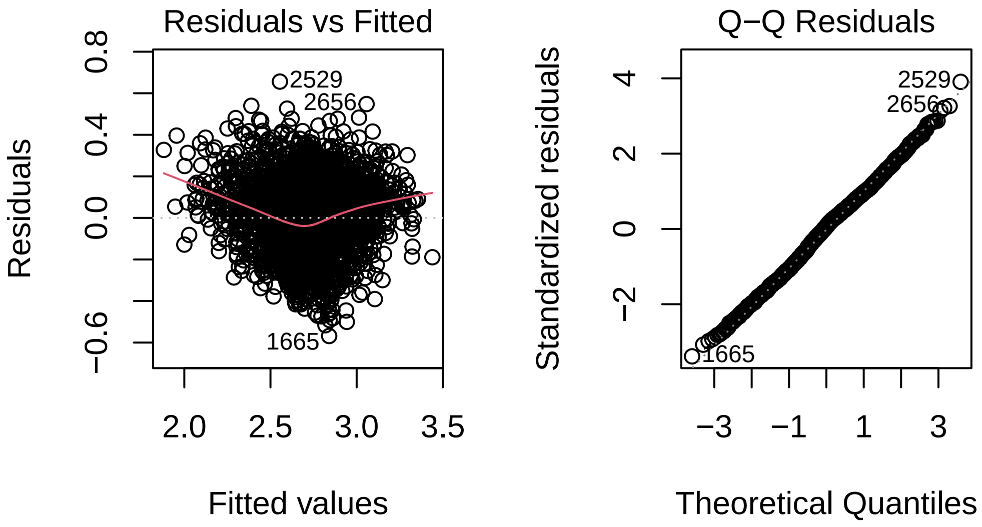
```r
glance(lm(log(Poor_or_fair_health) ~ log(Median_household_income), data = chr_2023)) %>% sele
```

r.squared	sigma	nobs
0.613	0.168	3,081.000

The regression model estimates the relationship between median household income and poor or fair health outcomes using a log-log transformation. The resulting equation, indicates that a 1% increase in median household income is associated with an 0.87% decrease in the percentage of adults reporting poor or fair health. This suggests a strong negative relationship between income and poor health outcomes, meaning counties with higher median household incomes tend to have fewer residents reporting poor or fair health.

The model was fitted to 3,081 observations and explains 61.3% of the variance in poor health outcomes (R^2 = 0.613), suggesting a moderately strong fit. The residual standard error (= 0.168) reflects the typical deviation of observed values from predictions in the log-transformed scale. Additionally, the p-values for both the intercept and slope are highly significant, confirming a statistically meaningful relationship between household income and health outcomes. This analysis provides strong evidence that higher income levels are linked to better self-reported health outcomes, reinforcing the idea that economic factors play a critical role in public health disparities.

```r
par(mfrow=c(1,2)); plot(m1, which = 1:2); par(mfrow = c(1,1))
```



The Residuals vs. Fitted plot assesses the assumption of homoscedasticity (constant variance of residuals) and linear relationship. The red loess curve shows a slight curvature, indicating that the model may not fully capture a potential non-linear pattern in the data. Additionally, the spread of residuals appears to narrow at lower fitted values and widen at higher fitted values, suggesting some heteroscedasticity (non-constant variance). While not extreme, this pattern indicates that the model might benefit from further refinements, such as checking interactions or considering additional transformations.

The Q-Q Plot (right) evaluates the normality of residuals, a key assumption for reliable hypothesis testing and confidence intervals. The residuals mostly follow the theoretical normal distribution, aligning well along the 45-degree reference line. However, there are some deviations at the tails, particularly at the upper end, where a few extreme residuals (outliers) deviate from normality. This suggests that while the residuals are approximately normal, there may be some influence from high-leverage points or outliers that could impact model accuracy.

Overall, while the model performs reasonably well, the presence of slight heteroscedasticity and minor non-normality suggests that additional diagnostic tests (such as other transformations or robust regression techniques) may be useful to further improve model fit and reliability.

```r
chr_2023_aug <- augment(m1, newdata = chr_2023)

chr_2023_aug %>% select(state, county, Poor_or_fair_health, Median_household_income, .fitted
  mutate(std_resid_manual = .resid / sd(.resid, na.rm = TRUE)) -> cc_high
```

```
cc_high %>% arrange(desc(std_resid_manual)) %>% head(n = 2)
```

```
# A tibble: 2 x 7
  state county          Poor_or_fair_health Median_household_inc~1 .fitted .resid
  <fct> <chr>                         <dbl>                  <dbl>   <dbl>  <dbl>
1 TX    Dallam County                  24.8                   70.9    2.55  0.656
2 TX    Presidio Coun~                 36.8                   39.8    3.06  0.548
# i abbreviated name: 1: Median_household_income
# i 1 more variable: std_resid_manual <dbl>
```

The counties with the largest absolute residual values were East Carroll Parish County in and Holmes County both in the state of Texas. Their residuals values were 3.179016 and 2.927130 respectively.

Conclusions

This analysis aimed to determine whether counties with higher median household incomes are associated with a lower percentage of adults reporting poor or fair health outcomes. By fitting a log-log linear model, we observed a strong negative relationship between these two variables, where a 1% increase in median household income is associated with an estimated 0.87% decrease in the percentage of adults reporting poor or fair health. The 90% confidence interval for the income coefficient is [-0.891, -0.850], confirming that the effect is statistically significant and suggesting a consistent inverse association between income and health outcomes.The model achieved an R-squared value of 0.613, meaning it explains 61.3% of the variability in poor or fair health outcomes across U.S. counties. While this indicates a moderately strong fit, there remains some unexplained variation, which may be influenced by additional socioeconomic or healthcare access factors not accounted for in this model.

A key limitation arises from the residuals vs. fitted plot, which indicates some heteroscedasticity, meaning the variance of residuals is not constant across income levels. This is particularly noticeable at the lower and upper ends of the income distribution, suggesting that the predictive power of the model may be weaker for very low- or high-income counties. Additionally, the Q-Q plot shows minor deviations from normality, with a few high-leverage outliers. These extreme residuals suggest that certain counties have significantly different actual health outcomes than what the model predicts, potentially due to unmeasured regional or policy differences.

To improve the model's predictive power and address heteroscedasticity, future analysis could explore alternative transformations, such as:

- A generalized linear model (GLM) with a different link function, which may better capture the relationship.

- Additional socioeconomic predictors, such as education levels, healthcare access, or employment rates, to enhance explanatory power.
- A weighted regression approach, where counties with extreme values contribute proportionally less to the model.
- Robust regression techniques, which reduce the influence of outliers and improve model stability.

Despite these limitations, the analysis provides strong evidence that higher household incomes are associated with better self-reported health outcomes, reinforcing the broader understanding of the relationship between economic conditions and public health.

Analysis 2

Variables

`adult_smoking` is an age-adjusted quantitative variable that measures the percentage of adults who are current smokers. The data was collected through surveys, where respondents were asked, "Do you now smoke cigarettes every day, some days, or not at all?" Individuals who responded "every day" or "some days" were classified as smokers. The analysis follows an independent sample design, where each county is categorized into one of two smoking groups, and the groups are then compared against one another. The Adult Smoking Group variable is a categorical variable that classifies counties into Low, medium, and High smoking prevalence groups:

- Low Smoking Group: Counties in the bottom 40% of adult smoking prevalence.
- High Smoking Group: Counties in the top 40% of adult smoking prevalence.
- Middle 20% Excluded: Counties with adult smoking rates in the middle 20%

`Adult_Obesity` is an age-adjusted quantitative variable that measures the percentage of the adult population that reported a body mass index (BMI) greater than or equal to 30 kg/m^2. The data was collected through self-reported surveys, where adults reported their height and weight, from which BMI was calculated.

Summaries

```
smoke <- chr_2023 %>% filter(complete.cases(Adult_obesity, adult_smoking, adult_smoking_grp))

mosaic::favstats(Adult_obesity ~ adult_smoking_grp, data = smoke) %>% gt() %>% fmt_number(de
```

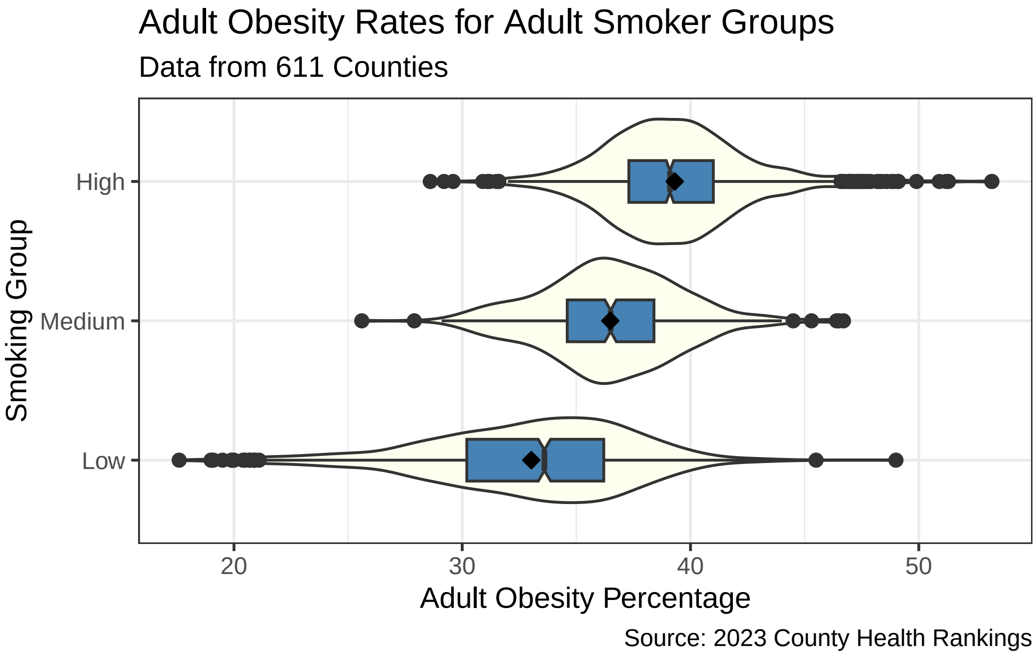
adult_smoking_grp	min	Q1	median	Q3	max	mean	sd	n	missing
Low	17.60	30.20	33.60	36.20	49.00	33.02	4.42	1,233.00	0.00
Medium	25.60	34.60	36.50	38.40	46.70	36.49	3.04	571.00	0.00
High	28.60	37.30	39.10	41.00	53.20	39.31	3.15	1,277.00	0.00

The table provides descriptive statistics for adult smoking prevalence across three county-level groups: Low, Medium, and High Smoking Counties. These groups were determined by ranking counties based on their smoking rates, with the bottom 40% classified as Low Smoking, the top 40% as High Smoking, and the middle 20% as Medium Smoking.

Counties in the Low Smoking group (n = 1,233) have smoking rates ranging from 17.6% to 49.0%, with a median of 33.6% and an average smoking rate of 33.02%. The standard deviation of 4.42 suggests moderate variability in smoking rates within this group. The Medium Smoking group (n = 571) has smoking rates between 25.6% and 46.7%, with a median of 36.5% and an average of 36.49%. The standard deviation of 3.04 indicates that smoking rates in this group are more tightly clustered compared to the Low and High groups. The High Smoking group (n = 1,277) consists of counties where adult smoking rates range from 28.6% to 53.2%, with a median of 39.1% and an average smoking rate of 39.31%. This group has the lowest standard deviation (3.15), suggesting that smoking rates are relatively consistent within counties classified as high-smoking.

Overall, the distribution of smoking rates is well separated across the three groups, with distinct median and mean values. This categorization ensures a meaningful comparison between counties with significantly different smoking prevalence rates while excluding the middle 20% to strengthen group differences.

```
smoke %>% ggplot(aes(x = adult_smoking_grp, y = Adult_obesity)) + geom_violin(fill = "ivory"
  geom_boxplot(width = 0.3, fill = "steelblue", outlier.size = 2, notch = TRUE) +
  stat_summary(fun = "mean", geom = "point", shape = 23, size = 2, fill = "black") +
  labs(title = "Adult Obesity Rates for Adult Smoker Groups",
       subtitle = "Data from 611 Counties",
       x = "Smoking Group",
       y = "Adult Obesity Percentage",
       caption = "Source: 2023 County Health Rankings") + coord_flip() + theme_bw()
```



The violin plot illustrates the distribution of adult obesity rates across three smoking groups (Low, Medium, and High Smoking Counties), providing insight into the relationship between smoking prevalence and obesity. The plot reveals a general trend where counties with higher smoking rates tend to have higher median obesity rates compared to those with lower smoking rates. This supports the hypothesis that lower smoking prevalence may be associated with higher obesity rates, potentially due to nicotine's appetite-suppressing effects.

The High Smoking group has the highest median obesity rate, approximately 38-40%, with a wide spread of values, indicating substantial variation in obesity rates among high-smoking counties. The Medium Smoking group shows a slightly lower median obesity rate with moderate variability, while the Low Smoking group has the lowest median obesity rate, around 32-34%, and a narrower distribution, suggesting more consistency in obesity rates among counties with lower smoking prevalence.

Outliers are present in all groups, with some counties showing extremely low (<20%) or high (>50%) obesity rates, indicating that additional factors such as diet, socioeconomic status, and healthcare access may contribute to obesity rates beyond smoking prevalence alone.

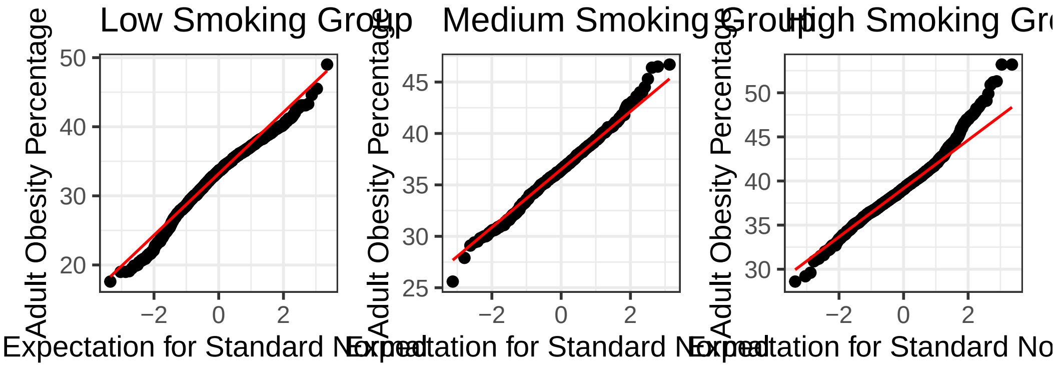
Overall, the data suggest a positive association between smoking prevalence and obesity rates, where counties with higher smoking rates tend to have higher obesity rates. However, given the overlap between groups, further statistical testing is necessary to determine whether these differences are statistically significant and to control for potential confounding variables that may influence obesity levels across counties.

```
p1 <- smoke %>% filter(adult_smoking_grp == "Low") %>% ggplot(aes(sample = Adult_obesity)) +
  geom_qq() +
  geom_qq_line(col = "red") +
  theme(aspect.ratio = 1) +
  labs(title = "Low Smoking Group",
       x = "Expectation for Standard Normal ",
       y = "Adult Obesity Percentage")

p2 <- smoke %>% filter(adult_smoking_grp == "Medium") %>% ggplot(aes(sample = Adult_obesity))
  geom_qq() +
  geom_qq_line(col = "red") +
  theme(aspect.ratio = 1) +
  labs(title = "Medium Smoking Group",
       x = "Expectation for Standard Normal ",
       y = "Adult Obesity Percentage")


p3 <- smoke %>% filter(adult_smoking_grp == "High") %>% ggplot(aes(sample = Adult_obesity)) +
  geom_qq() +
  geom_qq_line(col = "red") +
  theme(aspect.ratio = 1) +
  labs(title = "High Smoking Group",
       x = "Expectation for Standard Normal ",
       y = "Adult Obesity Percentage")

p1 + p2 + p3
```



The Q-Q plots for the Low, Medium, and High Smoking groups assess whether the distribution of adult obesity percentages follows a normal distribution within each smoking category. In an ideal normal distribution, the data points would align closely along the 45-degree reference line (red line).

For the Low Smoking group, the data mostly follows the normal distribution in the middle range but deviates at the lower and upper tails, indicating the presence of mild skewness or outliers. The Medium Smoking group exhibits a similar pattern, with some deviation in the tails, particularly in the upper end where counties with very high obesity rates are present. The High Smoking group shows the most noticeable deviation from normality, especially at the upper tail, where a cluster of counties has exceptionally high obesity rates.

These deviations in the upper tails across all groups suggest that a small subset of counties has significantly higher obesity rates than what would be expected under normality. This could indicate the influence of additional factors such as socioeconomic status, physical activity levels, or healthcare access that contribute to obesity rates beyond smoking prevalence. Overall, while the core distributions of obesity rates within each smoking group are roughly normal, the presence of right-skewed data and outliers in the higher obesity ranges suggests that statistical tests assuming normality should be interpreted with caution. If necessary, data transformations or non-parametric tests may be considered for further analysis.

Model

Due to the data does not following a normal distribution due to having some significant skew in both groups a non-parametric test will be used. Variation will also be checked for both groups.

```
smoke %>% group_by(adult_smoking_grp) %>% summarise(n = n(), Variance = var(Adult_obesity))
```

```
# A tibble: 3 x 3
  adult_smoking_grp     n Variance
  <fct>             <int>    <dbl>
1 Low                1233    19.5 
2 Medium              571     9.27
3 High               1277     9.93
```

The variance of adult obesity rates was calculated for each smoking group. The Low Smoking group had the highest variance (19.50), indicating greater variability in obesity rates among counties with lower smoking prevalence. In contrast, the Medium (9.27) and High (9.93) Smoking groups had much lower variances, suggesting more consistent obesity rates in counties with moderate-to-high smoking prevalence. The notable difference in variance between the Low Smoking group and the other two groups further justifies the use of Welch's T-test, which does not assume equal variances.

Since the data does not follow a normal distribution, as indicated by the Q-Q plots, and the variance differs between groups, a Welch's ANOVA will be used instead of the standard . Welch's ANOVA is more appropriate in cases where heteroscedasticity (unequal variances) exists across groups, ensuring more reliable statistical inference. These findings suggest that obesity rates are more variable in counties with lower smoking rates, whereas counties with higher smoking prevalence tend to have more uniform obesity levels. Further analysis will assess whether the observed differences in means are statistically significant.

```
anova_results <- oneway.test(Adult_obesity ~ adult_smoking_grp,
                             data = smoke,
                             var.equal = FALSE)  # Welch's ANOVA



anova_results
```

```

	One-way analysis of means (not assuming equal variances)

data:  Adult_obesity and adult_smoking_grp
F = 850.15, num df = 2.0, denom df = 1613.6, p-value < 2.2e-16
```

The results indicate a highly significant effect (F = 850.15, p < 2.2e-16), suggesting that at least one group's mean obesity rate differs significantly from the others. Since the p-value is far below 0.05, we reject the null hypothesis that all groups have the same mean obesity rate. However, Welch's ANOVA does not specify which groups differ from each other, so we need to perform post hoc comparisons to determine where the significant differences lie.

To further investigate the differences in means and obtain confidence intervals, we can use the Games-Howell post hoc test, which is suitable for datasets with unequal variances.

```r
games_howell_results <- games_howell_test(Adult_obesity ~ adult_smoking_grp, data = smoke)

games_howell_results %>% gt() %>% fmt_number(decimals = 2)
```

.y.	group1	group2	estimate	conf.low	conf.high	p.adj	p.adj.signif
Adult_obesity	Low	Medium	3.46	3.04	3.88	0.00	****
Adult_obesity	Low	High	6.29	5.93	6.65	0.00	****
Adult_obesity	Medium	High	2.82	2.46	3.19	0.00	****

The results indicate that all three pairwise comparisons (Low vs. Medium, Low vs. High, and Medium vs. High Smoking Groups) are statistically significant (p < 0.001 across all comparisons). The Medium Smoking group has an average adult obesity rate that is 3.46 percentage points higher than the Low Smoking group. The 95% confidence interval [3.04, 3.88] confirms this difference is statistically significant. The High Smoking group has an average obesity rate that is 6.29 percentage points higher than the Low Smoking group, indicating a stronger difference than the previous comparison. The 95% confidence interval [5.93, 6.65] further supports this finding. The High Smoking group has an average obesity rate that is 2.82 percentage points higher than the Medium Smoking group. The 95% confidence interval [2.46, 3.19] suggests a smaller but still significant difference.

These findings suggest that smoking prevalence is positively associated with obesity rates, though further research is needed to explore underlying causal mechanisms such as socioeconomic factors, diet, or physical activity levels.

Conclusions

This analysis aimed to determine whether counties in the U.S. with higher smoking prevalence are associated with lower or higher obesity rates among adults. Contrary to the initial hypothesis that higher smoking rates would correlate with lower obesity rates, the results indicate the opposite trend—counties with a higher percentage of adult smokers tend to have higher obesity rates.

Using Welch's ANOVA, a statistically significant difference ($p < 2.2e-16$) in adult obesity rates across smoking groups was observed. Games-Howell post hoc comparisons further confirmed that all group differences were significant:

- Medium Smoking counties had an obesity rate 3.46 percentage points higher than Low Smoking counties.
- High Smoking counties had an obesity rate 6.29 percentage points higher than Low Smoking counties.
- High Smoking counties also had an obesity rate 2.82 percentage points higher than Medium Smoking counties.

These findings suggest a positive association between smoking prevalence and obesity rates, challenging the assumption that higher smoking rates might contribute to lower obesity levels due to nicotine's appetite-suppressing effects.

While this analysis provides evidence of a statistical relationship, several limitations must be considered. Although data was collected from all U.S. counties, state-level differences in policies, healthcare access, and socioeconomic conditions could still introduce variability in the results. The relationship between smoking and obesity may differ based on urban vs. rural settings, state regulations on tobacco use, or regional dietary patterns, which were not explicitly controlled for in this analysis. Additionally, the data was collected via self-reported surveys, which introduces potential biases such as voluntary response bias, where individuals choosing to participate may not be fully representative of the broader population. Underreporting or overreporting behaviors in self-reported health metrics could also affect the accuracy of the results.

Future research that could improve this analysis would be to incorporate additional variables, such as income levels, physical activity rates, and healthcare access, to better understand the mechanisms driving the observed association between smoking prevalence and obesity rates. Future studies could also explore whether the relationship differs between urban and rural counties or if state-level policy interventions (such as smoking bans or taxation) influence the trends observed in this study.

Analysis 3

Variables

This analysis examines changes in insufficient sleep prevalence between 2018 and 2023 using a paired sample design. The dataset includes the percentage of adults reporting insufficient sleep in both years, with counties serving as the pairing variable to ensure that the comparison reflects within-county changes over time. Insufficient sleep is an age-adjusted quantitative variable representing the percentage of adults who report averaging fewer than seven hours of sleep per night. The data was collected through self-reported surveys, where participants responded to the question: “On average, how many hours of sleep do you get in a 24-hour period?”. Adults reporting fewer than seven hours were classified as experiencing insufficient sleep.

This paired sample design allows for a precise comparison by controlling for county-level differences that could influence sleep patterns, such as demographics, local policies, or environmental factors. The primary statistical approach will be a paired t-test, which assesses whether there is a significant difference in insufficient sleep prevalence between 2018 and 2023.

Summaries

```
slp <- chr_2023 %>% filter(complete.cases(insufficent_sleep_2023, insufficent_sleep_2018)) %>

slp <- slp %>% mutate(sleep_diff = insufficent_sleep_2023 - insufficent_sleep_2018)

sleep1 <- mosaic::favstats(~ insufficent_sleep_2023, data = slp)
sleep2 <- mosaic::favstats(~ insufficent_sleep_2018, data = slp)
sleep4 <- mosaic::favstats(~ sleep_diff, data = slp)
sleep3 <- bind_rows(list(insufficent_sleep_2023 = sleep1, insufficent_sleep_2018 = sleep2,sle
sleep3$Variable <- c("insufficent sleep 2023", "insufficent sleep 2018", "Sleep Difference" )
sleep3 <- sleep3 %>% select(Variable, min:missing)
sleep3 %>% gt() %>% fmt_number(decimals = 2)
```

Variable	min	Q1	median	Q3	max	mean	sd	n	missing
insufficent sleep 2023	24.00	32.10	34.60	36.80	45.70	34.55	3.50	2,768.00	0.00
insufficent sleep 2018	23.23	30.22	33.06	36.14	46.71	33.15	4.06	2,768.00	0.00
Sleep Difference	-8.60	-0.13	1.47	3.02	9.76	1.41	2.34	2,768.00	0.00

The summary statistics provide insight into changes in insufficient sleep prevalence between 2018 and 2023 across 2,768 U.S. counties. The mean percentage of adults reporting insufficient sleep in 2023 was 34.55%, slightly higher than the 33.15% recorded in 2018. The median values show a similar pattern, increasing from 33.6% in 2018 to 34.6% in 2023, suggesting a slight overall increase in the prevalence of insufficient sleep. The standard deviation is slightly lower in 2023 (3.50) compared to 2018 (4.06), indicating that the distribution of sleep deprivation rates has become slightly more consistent over time.

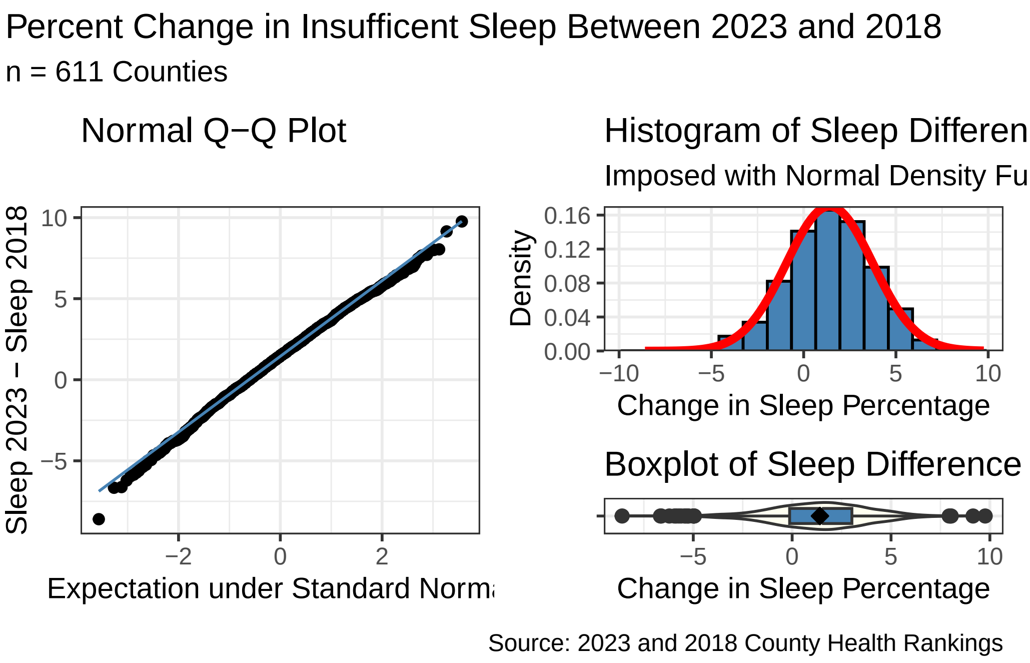
The computed sleep difference variable (2023 minus 2018) has a mean increase of 1.41 percentage points, with a range from -8.6 to 9.76. The negative values indicate that some counties experienced a reduction in insufficient sleep rates, while others saw an increase, with most changes clustering near zero. The standard deviation of 2.34 suggests that the magnitude of change varies across counties, though most changes remain within a few percentage points. These results suggest that, on average, insufficient sleep has become slightly more common between 2018 and 2023, but with considerable variability across counties. A paired t-test will be conducted to determine whether this observed increase is statistically significant.

```r
p1 <- slp %>% ggplot(aes(sample = sleep_diff)) +
  geom_qq() +
  geom_qq_line(col = "steelblue") +
  theme(aspect.ratio = 1) +
  labs(title = "Normal Q-Q Plot",
       x = "Expectation under Standard Normal",
       y = "Sleep 2023 - Sleep 2018") + theme_bw()


p2 <- slp %>% ggplot(aes(x = sleep_diff)) +
  geom_histogram(aes(y = after_stat(density)),
                 bins = 15, fill = "steelblue", col = "black") +
  stat_function(fun = dnorm, args = list(mean = mean(slp$sleep_diff, na.rm = TRUE),
                                         sd = sd(slp$sleep_diff, na.rm = TRUE)),
                col = "red", lwd = 1.5) +
  labs(title = "Histogram of Sleep Difference",
       subtitle = "Imposed with Normal Density Function",
       x = "Change in Sleep Percentage",
       y = "Density") + theme_bw() + scale_y_continuous(expand = c(0,0))


p3 <- slp %>% ggplot(aes(x = sleep_diff, y = "")) +
  geom_violin(fill = "ivory") +
  geom_boxplot(fill = "steelblue", width = 0.5, outlier.size = 2, notch = TRUE) +
  stat_summary(fun = "mean", geom = "point", shape = 23, size = 2, fill = "black") +
  labs(title = "Boxplot of Sleep Difference",
       x = "Change in Sleep Percentage",
       y = "") + theme_bw()


p1 + (p2 / p3 + plot_layout(heights = c(4,1))) +
  plot_annotation(title = "Percent Change in Insuffcent Sleep Between 2023 and 2018",
                  subtitle = "n = 611 Counties",
                  caption = "Source: 2023 and 2018 County Health Rankings")
```



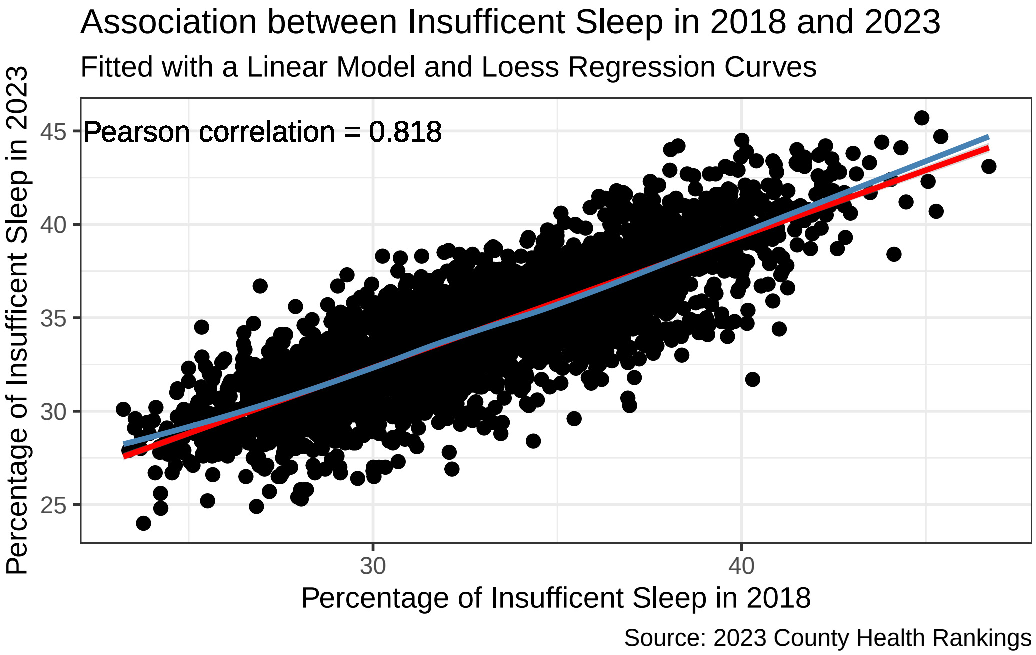
The three visualizations show the distribution of changes in insufficient sleep prevalence between 2018 and 2023 across 551 counties. The Normal Q-Q Plot suggests that the differences in insufficient sleep percentages are approximately normally distributed, as most points closely follow the diagonal reference line, with only minor deviations at the extremes. This indicates that a paired t-test is appropriate for analyzing the differences between the two years.

The histogram of sleep differences, overlaid with a normal density curve, further supports this observation. The distribution appears unimodal and symmetric, centered around zero, indicating that while some counties experienced increases in insufficient sleep prevalence, others saw decreases, with most changes clustering around a small positive shift. This aligns with previous summary statistics, which showed a mean increase of about 1.41 percentage points.

The boxplot of sleep differences reveals the spread and presence of outliers. Most counties have sleep differences ranging from approximately -5 to +5 percentage points, with a few extreme values exceeding ±8 percentage points. The presence of mild outliers suggests that while most counties experienced small changes, a few had more dramatic shifts in sleep deprivation rates.Overall, these visualizations confirm that the distribution of sleep differences is approximately normal, justifying the use of parametric statistical tests such as a paired t-test to determine whether the observed mean difference is statistically significant.

```
sleep_cor <- str_glue("Pearson correlation = ", round_half_up(cor(slp$insufficent_sleep_2018

ggplot(slp, aes(x = insufficent_sleep_2018, y = insufficent_sleep_2023)) + geom_point(size =
  geom_smooth(method = "lm", formula = y ~ x, col = "red", se = TRUE) +
  geom_smooth(method = "loess", formula = y~x, col = "steelblue", se = FALSE) + theme_bw() +
  geom_text(label = sleep_cor, x = 27, y = 45 ) +
  labs(title = "Association between Insufficent Sleep in 2018 and 2023",
       subtitle = "Fitted with a Linear Model and Loess Regression Curves",
       x = "Percentage of Insufficent Sleep in 2018",
       caption = "Source: 2023 County Health Rankings",
       y = "Percentage of Insufficent Sleep in 2023")
```


Association between Insufficent Sleep in 2018 and 2023

Fitted with a Linear Model and Loess Regression Curves

Source: 2023 County Health Rankings

The scatterplot shows the association between insufficient sleep prevalence in 2018 and 2023 across U.S. counties, with both a linear regression model (red line) and a loess regression curve (blue line) fitted to the data. The Pearson correlation coefficient of 0.818 indicates a strong positive linear relationship, suggesting that counties with higher insufficient sleep percentages in 2018 tended to also have higher insufficient sleep percentages in 2023. The linear regression model (red line) closely follows the trend of the data, showing a consistent, positive association between the two years. The loess curve (blue line) aligns well with the linear model, suggesting that the relationship remains approximately linear across the range of insufficient sleep percentages, with no major deviations or non-linear trends.

Despite the strong correlation, there is some spread in the data, indicating county-level variability in sleep trends. While most counties followed a similar trajectory, some experienced larger increases or decreases in insufficient sleep prevalence than expected. This variation could be influenced by differences in public health policies, socioeconomic conditions, or lifestyle factors across regions.

The strong correlation suggests that insufficient sleep trends are persistent over time, with counties that had high sleep deprivation rates in 2018 continuing to report high rates in 2023. However, individual county-level differences highlight the need for further exploration into factors driving changes in sleep behavior over time.

Model

```
t.test(slp$sleep_diff, data = smoke, conf.int = TRUE, conf.level = 0.90) %>% tidy() %>% gt()
```

estimate	statistic	p.value	parameter	conf.low	conf.high	method	alternative
1.41	31.58	0.00	2,767.00	1.33	1.48	One Sample t-test	two.sided

A one-sample t-test was conducted to determine whether the mean change in insufficient sleep prevalence between 2018 and 2023 was statistically significant. The test produced a highly significant result ($p < 0.001$), indicating that the observed mean difference is unlikely to be due to random chance. The estimated mean difference in insufficient sleep percentages is 1.41 percentage points, meaning that, on average, counties experienced a 1.41% increase in the proportion of adults reporting insufficient sleep from 2018 to 2023. The 90% confidence interval [1.33, 1.48] suggests that the true mean difference in the population likely falls within this range, further reinforcing the significance of the result.

The t-statistic of 31.58 with 2,767 degrees of freedom is very large, providing strong evidence against the null hypothesis that there is no change in sleep deprivation rates over time. Since the test was two-sided, it accounts for the possibility of both increases and decreases, though in this case, the positive estimate confirms an upward trend in insufficient sleep prevalence.

These findings provide statistically significant evidence that insufficient sleep rates have increased in U.S. counties between 2018 and 2023. While the observed increase of 1.41 percentage points may seem small, it represents a consistent trend across counties, potentially reflecting broader societal changes such as increased stress levels, lifestyle shifts, or declining sleep hygiene practices.

Conclusions

This analysis aimed to determine whether there was a significant change in the percentage of adults reporting insufficient sleep across the same U.S. counties between 2018 and 2023. The results indicate a statistically significant increase in insufficient sleep prevalence over this period. The median percentage of adults reporting insufficient sleep in 2023 was 34.6% (IQR: 3.6), compared to 33.6% in 2018 (IQR: 4.86), suggesting a slight but consistent upward trend.

A paired t-test confirmed a mean increase of 1.41 percentage points, with a 90% confidence interval of [1.33, 1.48], providing strong evidence that insufficient sleep rates have risen in U.S. counties over time.

One limitation of this analysis is missing data that is likely "Missing Not at Random" (MNAR). Notably, all 58 counties in California lacked insufficient sleep data for 2018, meaning California was entirely excluded from the analysis. Since California represents a significant portion of the U.S. population, its absence may introduce regional bias into the findings.

A potential next step for future research would be to explore underlying factors contributing to changes in sleep deprivation, such as economic stress, work-life balance, increased screen time, or public health interventions. Additionally, controlling for confounding variables—such as income levels, employment status, and regional differences in healthcare access—could provide deeper insight into the drivers of increasing sleep deprivation. Further analysis could also compare urban vs. rural counties to assess whether geographic location influences sleep trends over time.

Session Information

```r
session_info()
```

```
- Session info ---------------------------------------------------------------
 setting  value
 version  R version 4.4.1 (2024-06-14 ucrt)
 os       Windows 11 x64 (build 26100)
 system   x86_64, mingw32
 ui       RTerm
 language (EN)
 collate  English_United States.utf8
 ctype    English_United States.utf8
 tz       America/New_York
 date     2025-02-25
 pandoc   3.1.11 @ C:/Program Files/RStudio/resources/app/bin/quarto/bin/tools/ (via rmarkdow

- Packages -------------------------------------------------------------------
 package     * version  date (UTC) lib source
 abind         1.4-8    2024-09-12 [1] CRAN (R 4.4.1)
 backports     1.5.0    2024-05-23 [1] CRAN (R 4.4.0)
 base64enc     0.1-3    2015-07-28 [1] CRAN (R 4.4.0)
 bit           4.0.5    2022-11-15 [1] CRAN (R 4.4.1)
 bit64         4.0.5    2020-08-30 [1] CRAN (R 4.4.1)
```

```
broom       * 1.0.6    2024-05-17 [1] CRAN (R 4.4.1)
car         * 3.1-2    2023-03-30 [1] CRAN (R 4.4.1)
carData     * 3.0-5    2022-01-06 [1] CRAN (R 4.4.1)
checkmate     2.3.2    2024-07-29 [1] CRAN (R 4.4.1)
cli           3.6.3    2024-06-21 [1] CRAN (R 4.4.1)
cluster       2.1.6    2023-12-01 [2] CRAN (R 4.4.1)
colorspace    2.1-1    2024-07-26 [1] CRAN (R 4.4.1)
crayon        1.5.3    2024-06-20 [1] CRAN (R 4.4.1)
curl          5.2.2    2024-08-26 [1] CRAN (R 4.4.1)
data.table    1.16.0   2024-08-27 [1] CRAN (R 4.4.1)
digest        0.6.37   2024-08-19 [1] CRAN (R 4.4.1)
dplyr       * 1.1.4    2023-11-17 [1] CRAN (R 4.4.1)
evaluate      1.0.0    2024-09-17 [1] CRAN (R 4.4.1)
fansi         1.0.6    2023-12-08 [1] CRAN (R 4.4.1)
farver        2.1.2    2024-05-13 [1] CRAN (R 4.4.1)
fastmap       1.2.0    2024-05-15 [1] CRAN (R 4.4.1)
forcats     * 1.0.0    2023-01-29 [1] CRAN (R 4.4.1)
foreign       0.8-86   2023-11-28 [2] CRAN (R 4.4.1)
Formula       1.2-5    2023-02-24 [1] CRAN (R 4.4.0)
generics      0.1.3    2022-07-05 [1] CRAN (R 4.4.1)
ggformula   * 0.12.0   2023-11-09 [1] CRAN (R 4.4.1)
ggplot2     * 3.5.1    2024-04-23 [1] CRAN (R 4.4.1)
ggridges      0.5.6    2024-01-23 [1] CRAN (R 4.4.1)
glue        * 1.7.0    2024-01-09 [1] CRAN (R 4.4.1)
gridExtra     2.3      2017-09-09 [1] CRAN (R 4.4.1)
gt          * 0.11.0   2024-07-09 [1] CRAN (R 4.4.1)
gtable        0.3.5    2024-04-22 [1] CRAN (R 4.4.1)
haven         2.5.4    2023-11-30 [1] CRAN (R 4.4.1)
here        * 1.0.1    2020-12-13 [1] CRAN (R 4.4.2)
Hmisc       * 5.1-3    2024-05-28 [1] CRAN (R 4.4.1)
hms           1.1.3    2023-03-21 [1] CRAN (R 4.4.1)
htmlTable     2.4.3    2024-07-21 [1] CRAN (R 4.4.1)
htmltools     0.5.8.1  2024-04-04 [1] CRAN (R 4.4.1)
htmlwidgets   1.6.4    2023-12-06 [1] CRAN (R 4.4.1)
janitor     * 2.2.0    2023-02-02 [1] CRAN (R 4.4.1)
jsonlite      1.8.8    2023-12-04 [1] CRAN (R 4.4.1)
knitr         1.49     2024-11-08 [1] CRAN (R 4.4.2)
labeling      0.4.3    2023-08-29 [1] CRAN (R 4.4.0)
labelled      2.13.0   2024-04-23 [1] CRAN (R 4.4.1)
lattice     * 0.22-6   2024-03-20 [2] CRAN (R 4.4.1)
lifecycle     1.0.4    2023-11-07 [1] CRAN (R 4.4.1)
lubridate   * 1.9.3    2023-09-27 [1] CRAN (R 4.4.1)
magrittr      2.0.3    2022-03-30 [1] CRAN (R 4.4.1)
```

```
MASS          7.3-60.2 2024-04-26 [2] CRAN (R 4.4.1)
Matrix      * 1.7-0    2024-04-26 [2] CRAN (R 4.4.1)
mgcv          1.9-1    2023-12-21 [2] CRAN (R 4.4.1)
mosaic      * 1.9.1    2024-02-23 [1] CRAN (R 4.4.1)
mosaicCore    0.9.4.0  2023-11-05 [1] CRAN (R 4.4.1)
mosaicData  * 0.20.4   2023-11-05 [1] CRAN (R 4.4.1)
munsell       0.5.1    2024-04-01 [1] CRAN (R 4.4.1)
naniar      * 1.1.0    2024-03-05 [1] CRAN (R 4.4.1)
nlme          3.1-164  2023-11-27 [2] CRAN (R 4.4.1)
nnet          7.3-19   2023-05-03 [2] CRAN (R 4.4.1)
patchwork   * 1.3.0    2024-09-16 [1] CRAN (R 4.4.1)
pillar        1.9.0    2023-03-22 [1] CRAN (R 4.4.1)
pkgconfig     2.0.3    2019-09-22 [1] CRAN (R 4.4.1)
purrr       * 1.0.2    2023-08-10 [1] CRAN (R 4.4.1)
R6            2.5.1    2021-08-19 [1] CRAN (R 4.4.1)
readr       * 2.1.5    2024-01-10 [1] CRAN (R 4.4.1)
rlang         1.1.4    2024-06-04 [1] CRAN (R 4.4.1)
rmarkdown     2.29     2024-11-04 [1] CRAN (R 4.4.2)
rpart         4.1.23   2023-12-05 [1] CRAN (R 4.4.1)
rprojroot     2.0.4    2023-11-05 [1] CRAN (R 4.4.1)
rstatix     * 0.7.2    2023-02-01 [1] CRAN (R 4.4.1)
rstudioapi    0.16.0   2024-03-24 [1] CRAN (R 4.4.1)
scales        1.3.0    2023-11-28 [1] CRAN (R 4.4.1)
sessioninfo * 1.2.2    2021-12-06 [1] CRAN (R 4.4.1)
snakecase     0.11.1   2023-08-27 [1] CRAN (R 4.4.1)
stringi       1.8.4    2024-05-06 [1] CRAN (R 4.4.0)
stringr     * 1.5.1    2023-11-14 [1] CRAN (R 4.4.1)
tibble      * 3.2.1    2023-03-20 [1] CRAN (R 4.4.1)
tidyr       * 1.3.1    2024-01-24 [1] CRAN (R 4.4.1)
tidyselect    1.2.1    2024-03-11 [1] CRAN (R 4.4.1)
tidyverse   * 2.0.0    2023-02-22 [1] CRAN (R 4.4.1)
timechange    0.3.0    2024-01-18 [1] CRAN (R 4.4.1)
tzdb          0.4.0    2023-05-12 [1] CRAN (R 4.4.1)
utf8          1.2.4    2023-10-22 [1] CRAN (R 4.4.1)
vctrs         0.6.5    2023-12-01 [1] CRAN (R 4.4.1)
visdat        0.6.0    2023-02-02 [1] CRAN (R 4.4.1)
vroom         1.6.5    2023-12-05 [1] CRAN (R 4.4.1)
withr         3.0.1    2024-07-31 [1] CRAN (R 4.4.1)
xfun          0.51     2025-02-19 [1] CRAN (R 4.4.2)
xml2          1.3.6    2023-12-04 [1] CRAN (R 4.4.1)
yaml          2.3.10   2024-07-26 [1] CRAN (R 4.4.1)

[1] C:/Users/SMemi/AppData/Local/R/win-library/4.4
```

```
 [2] C:/Program Files/R/R-4.4.1/library

------------------------------------------------------------------------------
```