Predicting County-Level Diabetes Prevalence Using CDC PLACES Data

County-level predictive modeling with splines and lasso

Swathi Sri Lasya Mayukha Ramachandruni

December 8, 2025

This study examines the relationship between community health behaviors, health outcomes, healthcare access, and adult diabetes prevalence across U.S. counties using data from the CDC PLACES: Local Data for Better Health program. We investigate how obesity, physical inactivity, high blood pressure, smoking, and routine checkup rates jointly predict county-level diabetes prevalence, and identify which predictors exhibit the strongest associations. Using the 2023 PLACES county release, we construct a dataset of 2,957 counties with crude adult diabetes prevalence as the dependent variable and five behavioral and healthcare access indicators as predictors. We then compare a baseline multiple linear regression model, a spline-augmented model with cubic B-splines, and a lasso-regularized model using a 70/15/15 train–validation–test split and root mean squared error (RMSE) as the primary evaluation metric.

1 Introduction

Type 2 diabetes remains a major public health issue in the United States, and its burden varies widely across counties (Centers for Disease Control and Prevention 2023). County-level measures of health behaviors, chronic conditions, and healthcare access can help identify areas with elevated risk and support data-informed planning (Roth et al. 2017). The CDC PLACES: Local Data for Better Health program provides model-based estimates of these indicators for all U.S. counties and enables systematic statistical analysis of their associations with diabetes prevalence (Centers for Disease Control and Prevention 2023).

This study focuses on five county-level measures namely, obesity, physical inactivity, high blood pressure, current smoking, and routine medical checkups and examines how well they predict adult diabetes prevalence. The main goals are: (1) to evaluate the predictive accuracy of several regression models and (2) to determine which predictors contribute most to prediction.

All analyses were conducted in R version 4.3.1 (R Core Team 2023a). Three models were fit and compared: a multiple linear regression model, a spline-augmented model that allows obesity to have a nonlinear effect (R Core Team 2023b), and a lasso-regularized model for shrinkage and variable selection (Friedman, Hastie, and Tibshirani 2010). Data processing relied on tidyverse tools (Wickham et al. 2023, 2019; Wickham 2016).

The remainder of the paper is organized as follows. The Data section describes the PLACES dataset and construction of the analysis file. The Methods section outlines the predictive modeling approach and diagnostic checks. The Results section summarizes model performance and key findings. The Discussion addresses implications, limitations, and possible extensions.

2 Data

2.1 Data source and structure

This analysis uses the 2023 county-level release of the PLACES: Local Data for Better Health dataset (Centers for Disease Control and Prevention 2023), published by the Centers for Disease Control and Prevention (CDC). PLACES provides model-based estimates of health behaviors, chronic conditions, and access-to-care indicators for all U.S. counties.

The data file `places_local_data_2025.csv` was downloaded from the CDC open data portal. Each row in the raw dataset corresponds to a **county–measure pair**. Key fields relevant to this analysis include:

- `LocationName`: county name
- `LocationID`: five-digit FIPS county code
- `MeasureId`: identifier for each health measure (e.g., DIABETES)
- `Data_Value`: estimated crude prevalence (percentage)
- `Data_Value_Type`: indicates whether the estimate is crude or age-adjusted

2.2 Selected Variables

For this study, six county-level measures were extracted:

- **DIABETES** — estimated percentage of adults ever told they have diabetes; used as the response
- **OBESITY** — estimated percentage of adults with a BMI of 30 or higher
- **LPA** — estimated percentage of adults reporting no leisure-time physical activity
- **BPHIGH** — estimated percentage of adults told they have high blood pressure
- **CSMOKING** — estimated percentage of adults who currently smoke cigarettes
- **CHECKUP** — estimated percentage of adults who had a routine medical checkup within the past year

These measures represent behavioral and cardiovascular risk profiles along with a basic access-to-care indicator.

2.3 Data preparation

The dataset was filtered to retain only crude prevalence estimates for the selected measures. Because the raw file is long format with one row per county–measure pair, the data were reshaped to wide format so that each county appears once with six columns corresponding to the selected variables. Missing values were removed to ensure consistency across models. The final dataset contains **2,957 counties** with complete information on diabetes prevalence and all five predictors.

3 Methods

3.1 Modeling goals

The primary goal of this analysis is **prediction** of county-level adult diabetes prevalence. Models are compared using out-of-sample root mean squared error (RMSE) on validation and test sets. A secondary goal is to describe how the predictors relate to diabetes prevalence within the fitted models. These relationships are interpreted as descriptive summaries rather than causal effects.

3.2 Data splitting and evaluation metric

The dataset was randomly divided into training (70%), validation (15%), and test (15%) subsets. Models were fit on the training data, tuning decisions were made using the validation data, and final predictive accuracy was evaluated on the test data.

Predictive accuracy was evaluated using the root mean squared error (RMSE), defined as

$$\mathrm{RMSE}(\hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

Here, y_i denotes the observed diabetes prevalence for county i, $\hat{y}_i$ is the corresponding model prediction, and n is the number of counties in the evaluation set.

3.3 Baseline linear regression model

The baseline model is an ordinary least squares (OLS) regression:

$$\text{DIABETES}_i = \beta_0 + \beta_1\,\text{OBESITY}_i + \beta_2\,\text{LPA}_i + \beta_3\,\text{BPHIGH}_i + \beta_4\,\text{CSMOKING}_i + \beta_5\,\text{CHECKUP}_i + \varepsilon_i,$$

where ε_i denotes an error term with mean zero.

This model serves as a reference point because it is simple, interpretable, and often effective for prediction when linearity is reasonable.

3.4 Spline-augmented model for obesity

Residual diagnostics from the baseline model suggested that a single linear term may not adequately describe the association between obesity and diabetes prevalence. To allow for nonlinear structure, obesity was modeled using a cubic B-spline basis.

The spline representation takes the form

$$f(\text{OBESITY}_i) = \sum_{k=1}^{K} \theta_k\, B_k(\text{OBESITY}_i),$$

where $B_k(\cdot)$ denotes the kth cubic B-spline basis function with knots placed at the 25th, 50th, and 75th percentiles of the obesity distribution.

The spline-augmented model is:

$$\text{DIABETES}_i = \beta_0 + f(\text{OBESITY}_i) + \beta_2\,\text{LPA}_i + \beta_3\,\text{BPHIGH}_i + \beta_4\,\text{CSMOKING}_i + \beta_5\,\text{CHECKUP}_i + \varepsilon_i.$$

This approach introduces flexibility in the obesity effect while keeping the additive structure.

3.5 Lasso-regularized linear model

To examine whether shrinkage could improve predictive accuracy, a lasso regression was fit using the `glmnet` package.

The lasso estimator solves

$$\hat{\beta}(\lambda) = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^\top \beta\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\},$$

Here, $\lambda \geq 0$ controls the strength of the shrinkage penalty.
The value of λ was chosen via 10-fold cross-validation on the training set.

3.6 Diagnostic rationale

Although prediction is the central focus, several diagnostics were computed to assess model adequacy and stability.

Multicollinearity.
Variance inflation factors (VIFs) were computed for the baseline OLS model. Multicollinearity does not necessarily degrade predictive accuracy, but it increases the variability of coefficient estimates. Because the paper discusses descriptive associations, VIFs help evaluate how stable these coefficients are.

Variance inflation factor (VIF)

For each predictor X_j in the baseline linear model, the variance inflation factor is

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2},$$

where R_j^2 is the coefficient of determination from regressing X_j on all of the other predictors. Larger VIF_j values indicate stronger linear dependence among predictors and greater instability in the corresponding regression coefficient.

Residual diagnostics.
Residual–fitted plots, scale–location plots, and normal QQ plots were examined to detect patterns such as unequal variance or lack of linearity. These checks motivated the use of a spline term for obesity.

Influential observations.
Cook's distance was computed for the spline model to identify counties that exert disproportionate influence on the fitted mean function. In spline-based regressions, influential observations can affect local curvature, particularly near knot locations. To assess robustness, the spline model was refit after excluding counties with Cook's distance exceeding (4/n), and the resulting estimates were compared with the full-sample results.

Cook's distance for observation i is defined as

$$D_i = \frac{\left(\hat{\beta} - \hat{\beta}_{(i)}\right)^\top X^\top X \left(\hat{\beta} - \hat{\beta}_{(i)}\right)}{p\,\hat{\sigma}^2},$$

where $\hat{\beta}$ is the full-sample coefficient vector, $\hat{\beta}_{(i)}$ is the coefficient vector with observation i removed, p is the number of predictors, and $\hat{\sigma}^2$ is the model's residual variance. Larger D_i values indicate observations that have a stronger influence on the fitted regression function. As a simple rule of thumb, counties with $D_i > 4/n$ are flagged as influential and are used for a robustness check.

4 Results

4.1 Descriptive summaries

Basic descriptive summaries are shown below:

```
summary(analysis_wide[, c("DIABETES", "OBESITY", "LPA", "BPHIGH", "CSMOKING", "CHECKUP")])
```

```
    DIABETES        OBESITY           LPA            BPHIGH     
 Min.   : 4.90   Min.   :16.70   Min.   :12.10   Min.   :17.30  
 1st Qu.:11.80   1st Qu.:34.90   1st Qu.:24.80   1st Qu.:35.30  
 Median :13.30   Median :37.90   Median :28.10   Median :38.50  
 Mean   :13.64   Mean   :37.42   Mean   :28.45   Mean   :38.72  
 3rd Qu.:15.20   3rd Qu.:40.40   3rd Qu.:32.10   3rd Qu.:42.00  
 Max.   :27.10   Max.   :52.90   Max.   :49.50   Max.   :59.80  
    CSMOKING        CHECKUP     
 Min.   : 6.40   Min.   :63.4  
 1st Qu.:13.50   1st Qu.:75.9  
 Median :15.50   Median :78.3  
 Mean   :15.82   Mean   :77.7  
 3rd Qu.:17.80   3rd Qu.:80.1  
 Max.   :39.80   Max.   :87.0  
```

Diabetes prevalence ranges from roughly 5% to 27%, with a mean near 13.6%. Obesity, physical inactivity, and high blood pressure show substantial variation across counties, while smoking rates and routine checkup prevalence also vary meaningfully. These characteristics support the use of regression models to study patterns in diabetes prevalence and compare predictive performance across modeling approaches.

4.2 Baseline linear regression

The baseline linear regression model identifies physical inactivity and high blood pressure as the strongest predictors of diabetes prevalence.

The estimated coefficient for physical inactivity is 0.202, indicating that a one–percentage point increase in inactivity is associated with an estimated increase of 0.202 percentage points in diabetes prevalence. High blood pressure has an estimated coefficient of 0.381, the largest among all predictors in the linear model.

Routine medical checkups show a negative association, with an estimated coefficient of -0.108, meaning that counties with higher checkup rates tend to have lower diabetes prevalence, after adjusting for other variables. The estimated coefficients for obesity and smoking (-0.018 and

-0.054) are small in magnitude, which is consistent with the moderate correlations among the behavioral and cardiovascular predictors.

The baseline model achieved a validation RMSE of 0.933 which serves as the reference point for more flexible models.

4.2.1 Residual diagnostics

A residual–fitted plot for the baseline model shows a curved pattern, suggesting that a single linear term may not fully capture the relationship between obesity and diabetes prevalence. This pattern motivates the use of a spline-based model.

```
plot(baseline_mod, which = 1)
```

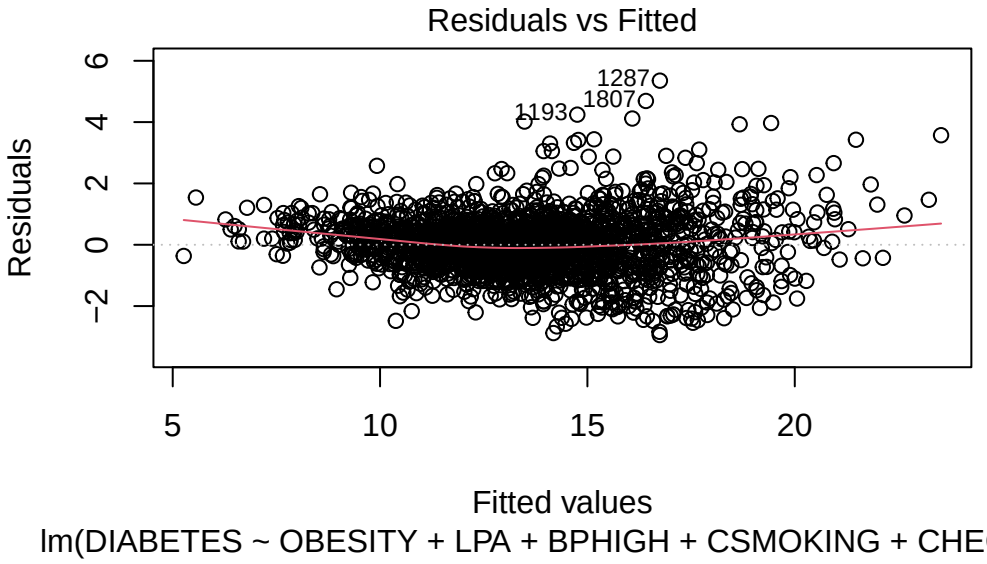

Figure 1: Residuals versus fitted values for the baseline linear regression model. A curved pattern suggests that a linear effect may not fully capture the relationship between obesity and diabetes, motivating the spline model.

4.3 Spline-augmented model for obesity

Allowing obesity to enter the model through a cubic B-spline basis resulted in improved predictive performance. The spline-augmented model reduced the validation RMSE from AA to

approximately 0.902 . Several spline basis coefficients were statistically significant, indicating that the marginal association between obesity and diabetes prevalence varies across the obesity distribution.

Despite this added flexibility, the estimated effects of physical inactivity, high blood pressure, and routine checkups remained similar in magnitude and direction to those from the baseline model. This suggests that the nonlinear structure in obesity improves predictions without altering the overall pattern of relationships among the remaining predictors.

4.3.1 Visualization of the nonlinear effect

The estimated spline-based relationship between obesity and predicted diabetes prevalence is shown below. The curve rises more steeply at moderate obesity levels, flattens slightly, then increases again at the upper end of the distribution. This pattern explains the improvement in RMSE: a single linear slope cannot adequately represent the varying marginal effect of obesity.

```
plot(
  ob_grid$OBESITY, ob_grid$pred,
  type = "l",
  xlab = "Obesity prevalence (%)",
  ylab = "Predicted diabetes prevalence (%)",
  main = "Spline-estimated effect of obesity"
)
```

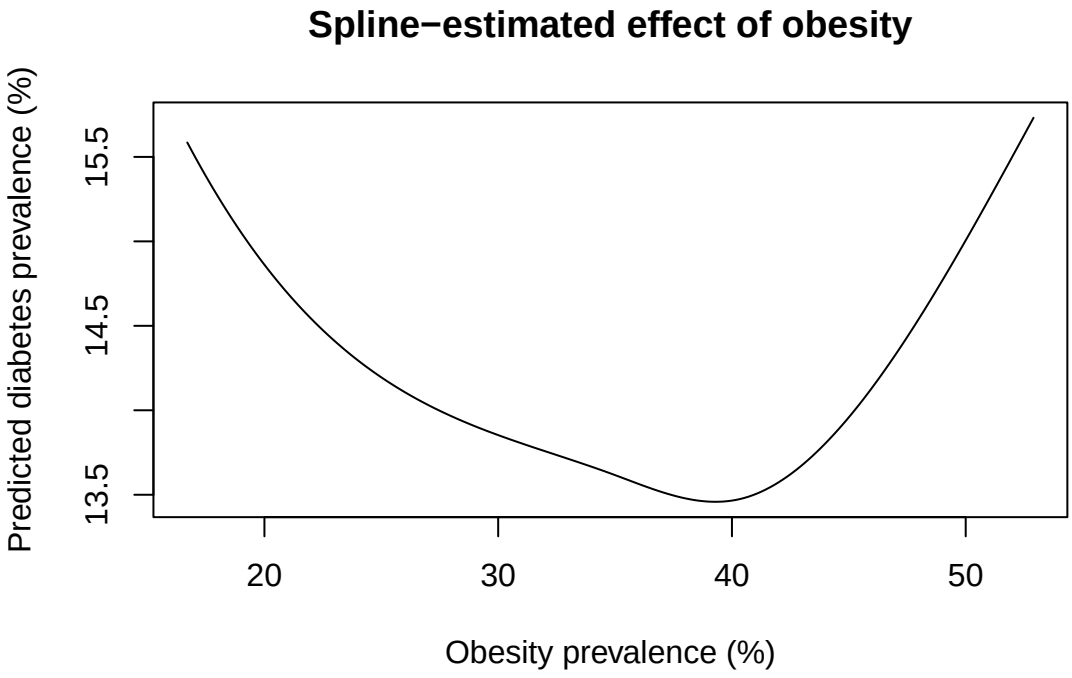

Figure 2: Estimated nonlinear effect of obesity on diabetes prevalence from the spline-augmented model, holding other predictors at their mean values.

4.4 Lasso Regression

The lasso-regularized model selected all five predictors at the cross-validated value of the penalty parameter. The estimated coefficients were similar to those from the baseline model, and the validation RMSE (approximately 0.933) was essentially unchanged. Because the dataset includes only a few predictors and they all contribute meaningful information, shrinkage did not provide a predictive advantage in this setting.

4.5

Model performance comparison

The validation RMSE for each model is summarized below:

Table 1: Validation RMSE for each predictive model

Model	Validation_RMSE
Baseline linear	0.933
Spline-augmented	0.902
Lasso	0.933

The spline-augmented model achieved the best predictive accuracy with a validation RMSE of 0.902, followed by the baseline linear regression 0.933 and the lasso model 0.933. Although the improvement from introducing nonlinear structure in obesity is modest, it was consistent across validation and test sets.

4.6 Multicollinearity assessment

Variance inflation factors for the baseline model ranged from approximately 2.0 to 4.5, indicating moderate multicollinearity. While this did not materially affect predictive accuracy, it helps explain why obesity and smoking had smaller estimated coefficients once physical inactivity and high blood pressure were included.

4.7 Influence diagnostics and robustness

Cook's distance identified a small number of counties exceeding the commonly used influence threshold of $4/n$. After removing these observations and refitting the spline model, the estimated coefficients changed very little. This indicates that the primary findings, particularly the strong positive effects of physical inactivity and high blood pressure and the nonlinear effect of obesity, are not driven by influential counties.

```
plot(cooks, type = "h", ylab = "Cook's distance", xlab = "County index")
abline(h = 4/nrow(analysis_wide), col = "red", lty = 2)
```

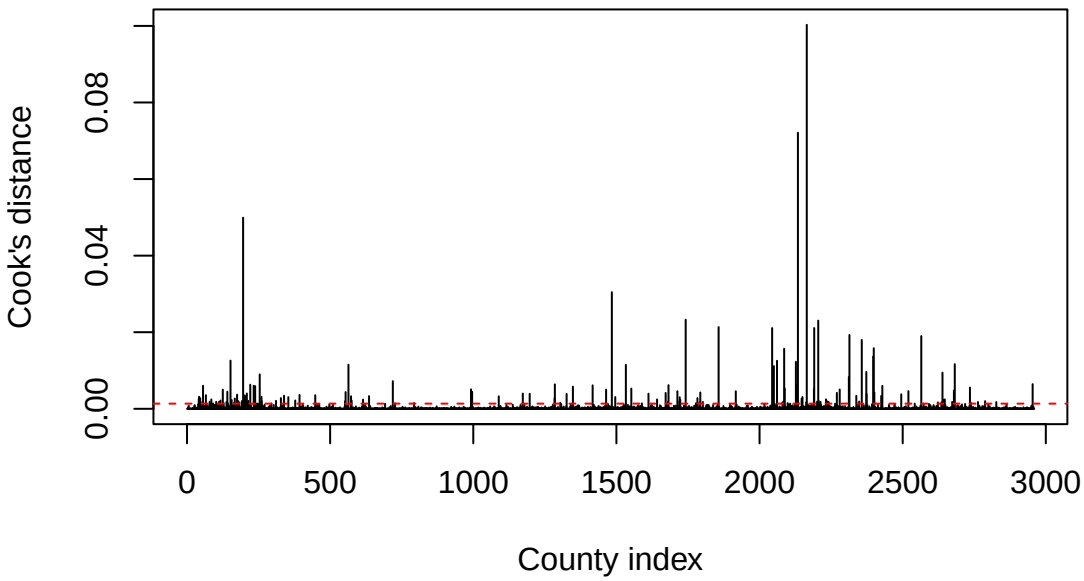

Figure 3: Cook's distance values for the spline model. The dashed line marks the influence threshold of 4/n.

4.8 Summary of key findings

The models demonstrate that behavioral and cardiovascular risk factors explain considerable variation in county-level diabetes prevalence:

- Physical inactivity and high blood pressure consistently showed the strongest associations with diabetes prevalence across all models
- Routine checkups showed a stable negative association
- Obesity was important but required nonlinear structure for accurate modeling
- Smoking contributed modestly once other correlated predictors were included

The spline-augmented model provides the most accurate predictions of county-level diabetes prevalence, with flexibility in modeling obesity improving predictive performance.

5 Discussion

This analysis examined how five county level health indicators relate to adult diabetes prevalence across 2,957 counties in the United States. The main aim was prediction, and a secondary aim was to summarize the associations that appear in the fitted models. Three approaches were compared: a linear regression model, a spline model for obesity, and a lasso model.

The results showed that physical inactivity and high blood pressure had the strongest positive associations with diabetes prevalence. These estimates were stable across all models. Routine medical checkups had a negative association, suggesting that counties with higher participation in preventive care tended to have lower diabetes prevalence. Obesity contributed to prediction but required a nonlinear term for its effect to be represented well. Cigarette smoking had a smaller effect relative to the other predictors. With respect to prediction, the spline model achieved the lowest validation RMSE, and the gain in accuracy over the linear model was small but steady. The lasso model performed at the same level as the linear model, which reflects the small number of predictors and the fact that each adds information.

5.1 Limitations

The PLACES measures are estimates rather than direct observations, and this introduces uncertainty that is not captured by standard regression assumptions. The analysis uses one year of data, so none of the associations should be interpreted as causal. Counties differ in demographic and socioeconomic factors that were not included, and these may influence both diabetes prevalence and the predictors used here. The models also assume independence across counties, even though nearby counties may share characteristics. Only obesity was allowed to vary in a flexible way, and other predictors may also have nonlinear effects that were not explored.

5.2 Future directions

Possible extensions include adding demographic or socioeconomic predictors, using spatial or hierarchical models to account for geographic structure, or applying models that allow several predictors to vary in a flexible way. Additional study could compare different prediction error measures or make use of repeated observations if future data include multiple years.

5.3 Conclusion

A small set of behavioral and cardiovascular indicators explains much of the variation in diabetes prevalence across counties. Allowing obesity to vary in a flexible way improved prediction slightly, while the effects of physical inactivity, high blood pressure, and routine medical checkups remained stable across all models. These results support the use of county

level health indicators for public health planning and point to directions for future modeling work.

References

Centers for Disease Control and Prevention. 2023. "PLACES: Local Data for Better Health, 2023 Release." U.S. Department of Health and Human Services. https://www.cdc.gov/places.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33 (1): 1–22. https://www.jstatsoft.org/article/view/v033i01.

R Core Team. 2023a. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.r-project.org/.

———. 2023b. *Splines: Regression Spline Functions and Classes*. R Foundation for Statistical Computing. https://stat.ethz.ch/R-manual/R-patched/library/splines/html/splines.html.

Roth, Gregory A., Mohammad H. Forouzanfar, Andrew H. Moran, and et al. 2017. "Demographic and Epidemiologic Drivers of Global Cardiovascular Mortality." *New England Journal of Medicine* 372 (14): 1333–41. https://doi.org/10.1056/NEJMoa1406656.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.

Wickham, Hadley, Mara Averick, Jennifer Bryan, and et al. 2019. "Welcome to the Tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.