Predicting Solo Commuting Patterns in California using Demographics and Geography*

Terrell Enoru

December 9, 2025

Understanding commuting patterns is useful for transportation planning and environmental policy in California. This study examines whether ethnic and geographic factors can predict the proportion of workers who drive alone to work across different parts of California. The dataset records transportation methods across ethnicities and regions. Using weighted multiple linear regression , I analyze transportation mode choices across different regions and ethnicities. The results show that ethnicity,region, and population were found to be excellent at predicting the proportion of workers who drive alone ($R^2 = 0.96$) . These results suggest that commuting methods differ significantly between different regions and ethnicities.

1 Introduction

The motivation for this paper came from switching to driving to work after using BART for years. This prompted me to look into the most popular transportation methods in California. During this search I discovered that California has the second highest vehicle emissions in the United States (Fernandez 2024). One approach to reducing those emissions would be to reduce the number of solo commuters. So understanding the factors that influence whether workers drive alone to work is critical to changing policy to promote carpooling.

The knowledge gap being addressed is understanding how demographics and geography drive transportation choices in California. Most of the analysis on demographics and transportation focuses on age rather than ethnicity. This paper looks to focus on ethnicity instead with the hopes that it has a more substantial impact on transportation choice than age. In order to address the knowledge gap this paper uses weighted linear regression with log-transformed population data to model the relationship between regional characteristics (region, population, and ethnic composition) and the percentage of workers who drive alone to work.

*Project repository available at: https://github.com/tycebot/california-solo-driving-study

The structure of this paper is organized into five main sections: Data introduces the dataset and key variables. Methods explains the statistical methods used for the analysis. Results discusses the results of the analysis and their interpretation. Discussion provides an overview on the project before going into the results. Works Cited contains the citations for any external sources used in the paper.

2 Data

The California transportation data is provided by the California Open Data Portal and compiled from the American Community Survey(Health Equity 2025). Each row represents the transportation behavior for a specific demographic group within a Metropolitan Planning Organization (MPO) region during a particular reporting period. The dataset covers four reporting periods: 2000, 2005-2007, 2008-2010, and 2006-2010.

The key variables used in this analysis are percent which is the percentage of residents aged 16 and older who use a particular mode of transportation to work. Region_name and region_code are the Metropolitan Planning Organization (MPO)-based region name and numeric code, which identifies the geographic area. Pop_total is the total population of a demographic group in a region. This variable is log-transformed in the analysis to account for the nonlinear relationship between population size and transportation behavior. Race_eth_code is a numeric code representing different race/ethnicity groups (e.g., White, Hispanic/Latino, Asian, African American, etc.). Reportyear is the time period when data was collected. Mode is the transportation method used to get to work.

The data was collected by the American Community Survey, which uses a combination of mail, telephone, and in-person interviews to gather information about commuting patterns. The survey asks respondents about their usual mode of transportation to work during the week. One limitation is that the survey captures only the primary mode of transportation, so people with mixed commutes are only counted for one category. A key assumption of this analysis is that the self-reported transportation mode accurately reflects typical commuting behavior. Another key assumption is that the demographic and geographic classifications don't change significantly within reporting periods.

The data was filtered to 2006-2010 report years to reduce temporal effects. 2006-2010 was chosen as the time period because it was the second most recent but had 70000 more records than the most recent time period. The geotype was also filtered to region because we are interested looking at regional variation. Finally the mode was filtered to only solo driving because that is the focus of the analysis.

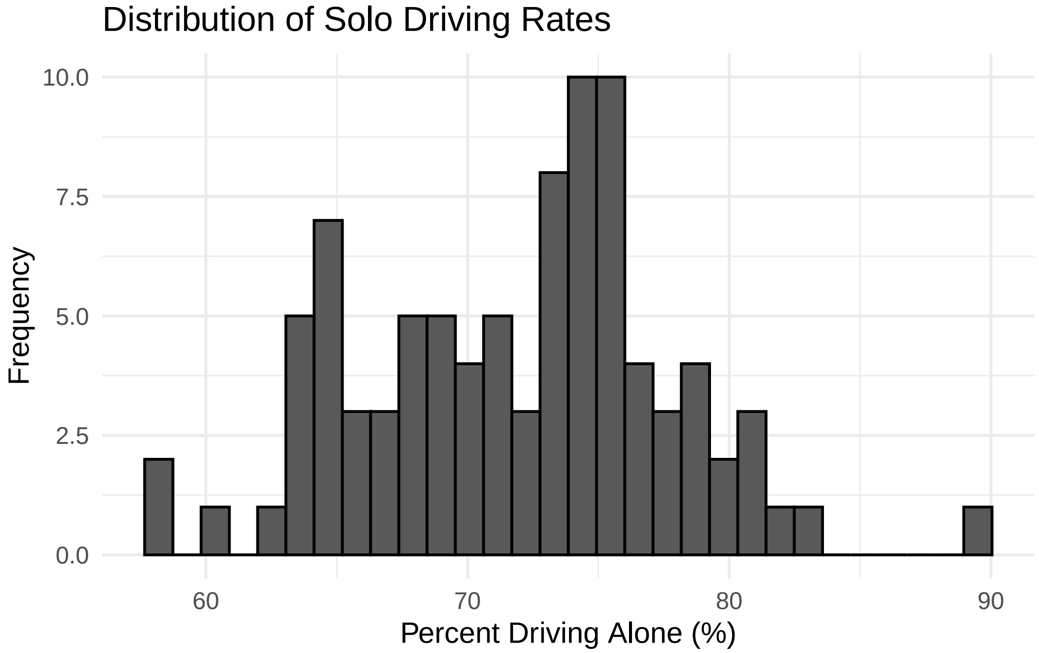

Figure 1: Distribution of the percentage of workers who drive alone across all regions and demographic groups. The distribution is left skewed with a peak around 75%, suggesting that driving alone is the most common commuting mode.

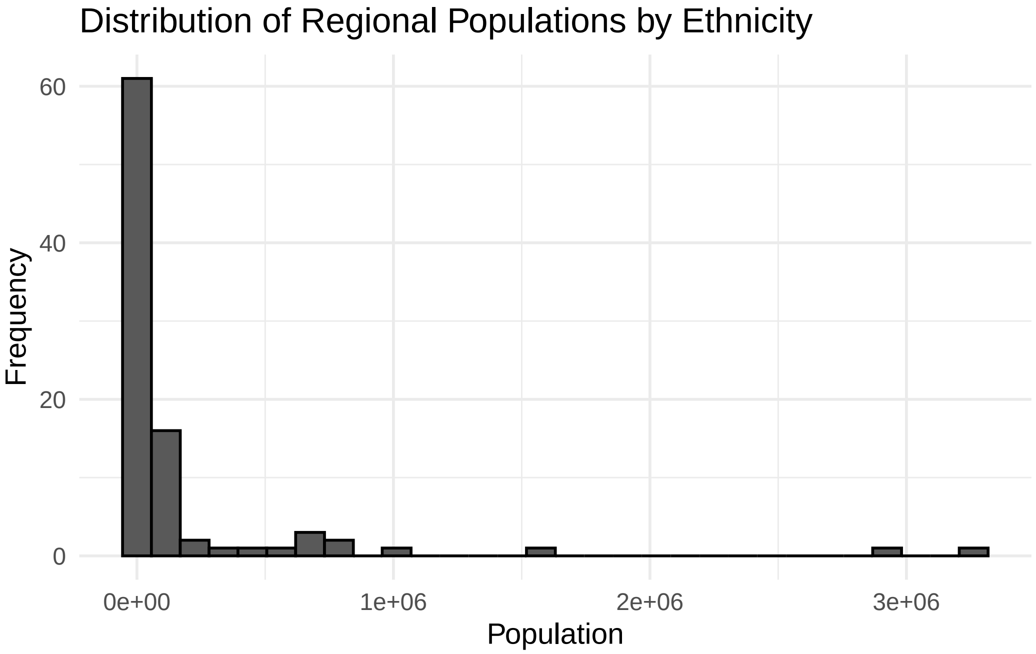

Figure 2: Distribution of the total population of a specific ethnicity in a specific region. The distribution is extremely right skewed with a peak around 0 because some regions have very small(few 100) population of certain ethnicities which is why weighted least squares was used.

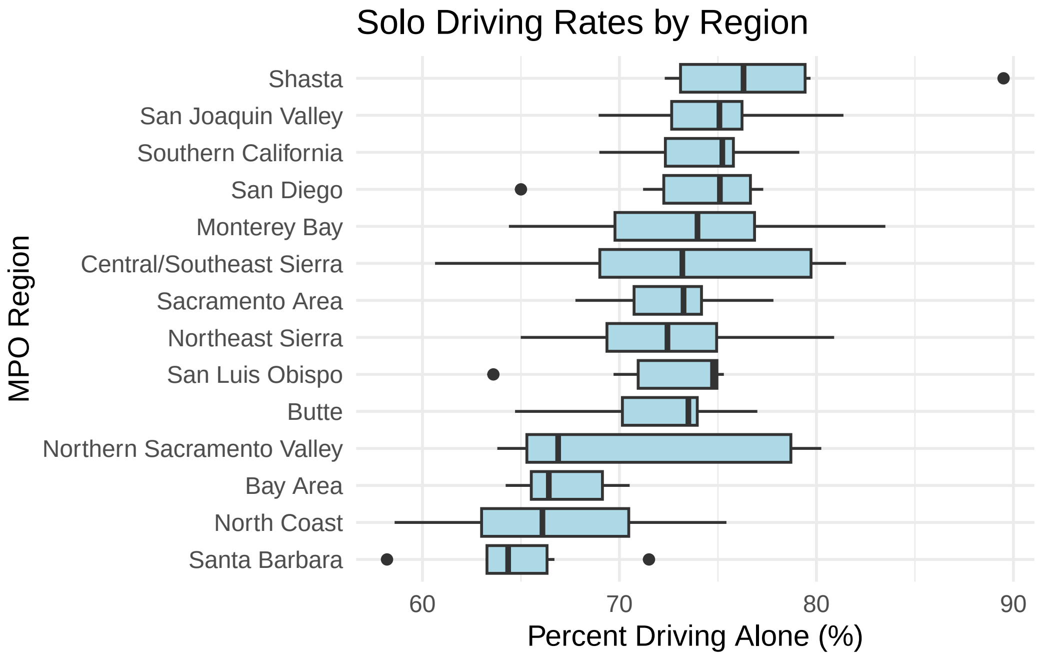

Figure 3: Comparison of solo driving rates across different MPO regions, showing substantial variation. Some regions (such as Santa Barbara, North Coast, Bay Area, and North Sacramento Valley) show lower median rates of solo driving, suggesting that geographic and infrastructure factors play an important role.

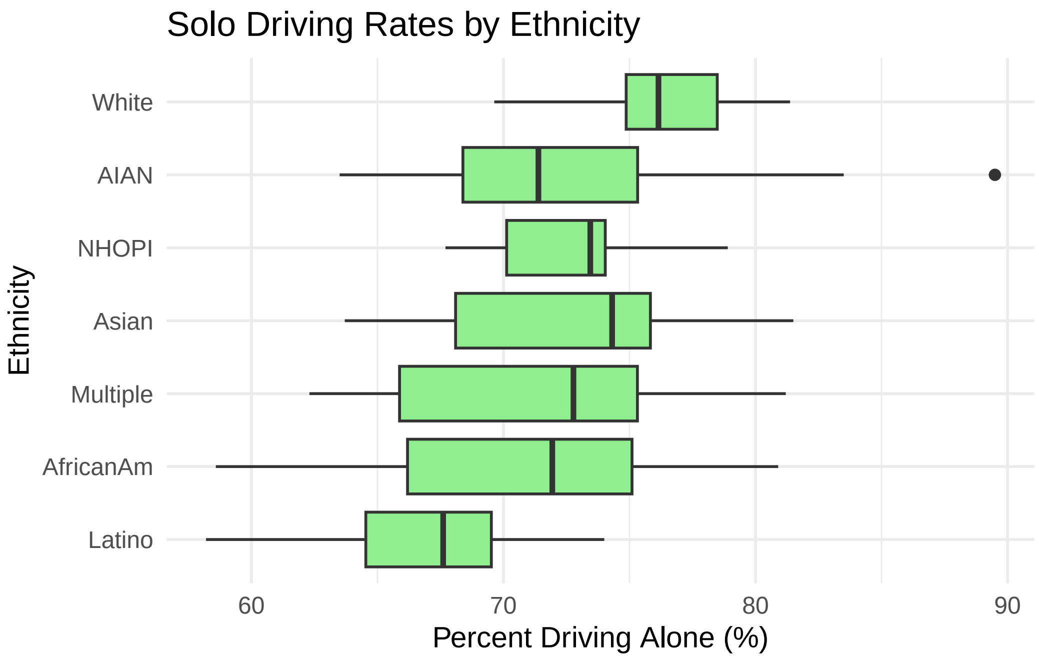

Figure 4: Comparison of solo driving rates across different ethnicities, showing similar driving rates across ethinicities with two exceptions. Whites have higher median rates of solo driving and Latinos have lower median rates of solo driving than other ethnicities.

3 Methods

We adopt a weighted multiple linear regression model with percent of worker's who are solo commuters as the response variable. The predictors are ethnicity,region, and the log transformation of population. Let Y_i denote the percentage of workers driving alone for observation i. Let $\log(X_{1i})$ denote the natural logarithm of the total population for observation i. Let X_{2i} denote a categorical indicator variable for region code. Finally X_{3i} denotes categorical indicator variables for race/ethnicity code. The model can be written as:

$$Y_i = \beta_0 + \beta_1 \log(X_{1i}) + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i$$

where β_0 represents the baseline percentage of solo drivers for the reference group. β_1 represents the expected change in solo driving percentage for a one-unit increase in log population. β_2 represents the expected change in solo driving percentage for a specific region over the baseline. β_3 represents the expected change in solo driving percentage for a specific ethnicity over the baseline. ϵ_i represents the error not explained by the model.

The log transformation of population is used to account for the expected nonlinear relationship between population and transportation method. As populations grow, the marginal impact of additional people on transportation patterns likely shrinks. Which a logarithmic transformation models better than a linear relationship. Weighted least squares is used to account for large differences in population across regions and ethnicity. Specifically the inverse of the squared standard error is used as the weight.

$$w_i = \frac{1}{\mathrm{SE}(Y_i)^2}$$

The analysis relies on four key assumptions about the error terms. The assumptions are that after log transformation the relationship between predictors and response is linear, observations are independent of one another, the variance is constant $Var(\epsilon_i) = \sigma^2$ for all i, and error terms are normally distributed $\epsilon_i \sim N(0, \sigma^2)$. These assumptions allow for inference on the coefficients.

To determine whether ethnicity and geography have statistically significant effects on solo driving behavior, I conduct the following F-tests.

For ethnicity effect:

$$H_0 : \beta_{\text{race}} = 0 \quad \text{vs.} \quad H_A : \beta_{\text{race}} \neq 0$$

where $\beta_{\text{race}} = (\beta_{\text{race}_1}, \beta_{\text{race}_2}, \dots, \beta_{\text{race}_k})^T$ is the vector of ethnicity coefficients.

For region effect:

$$H_0 : \beta_{\text{region}} = 0 \quad \text{vs.} \quad H_A : \beta_{\text{region}} \neq 0$$

where $\beta_{\text{region}} = (\beta_{\text{region}_1}, \beta_{\text{region}_2}, \dots, \beta_{\text{region}_m})^T$ is the vector of region coefficients.

To determine whether log(pop_total) has a statistically significant effect on solo driving behavior I conduct a one sided t-test.

For population effect:

$$H_0 : \beta_{\log(\text{pop})} = 0 \quad \text{vs.} \quad H_A : \beta_{\log(\text{pop})} \neq 0$$

The significance level is set at $\alpha = 0.05$ for all tests.

I implemented this analysis using the R programming language(R Core Team 2025) and the tidyverse package(Wickham et al. 2019) for data cleaning and plot creation.

4 Results

The model's weighted R-squared value is 0.966, indicating that approximately 96.6% of the variation in drive-alone percentages is explained by race/ethnicity, region, and population size.The estimated intercept is $b_0 = 84.579$. In other words the expected solo commuting rate of the reference group(african americans in the bay area) if log(pop_total) = 0 is 84.6%. The log population coefficient is $\hat{\beta}_1$ = -1.49. This indicates that for a 1% increase in population, the expected percentage of workers driving alone changes by approximately -0.0149 percentage points.

The F test on ethnicity returned a p value of 5.85×10^{-41}. Since the p value is less than the significance level(a=0.05) we reject the null hypothesis that all the ethnicity coefficients are zero. Similarly the p value for region 6.62×10^{-37} is less than the significance level so we reject the null hypothesis that all the region coefficients are zero.

The model and hypothesis test rely on four key assumptions which are that the regression function is linear, the error terms have equal variance, the error terms are independent, and the error terms are normally distributed. The equal variance assumption seems to hold after weighting based on the variance remaining roughly constant across fitted values.There some extra variation around the fitted value of 73 but that is normal for real world data. The normality assumptions seems to hold for most values however there is a clear deviation at the tails which given I used weighted least squares makes sense and fulfills the assumption for most of the points of interest. Independence of errors is likely violated due to different ethnicities within the same region likely being influenced by unobserved regional factors(eg transit infrastructure).

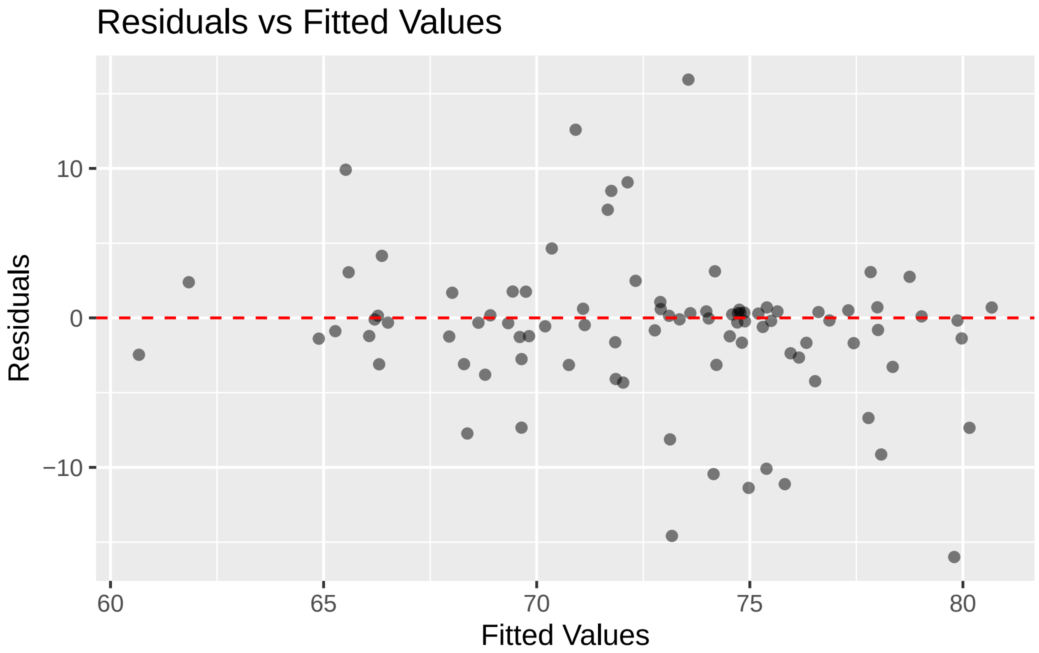

Figure 5: Residual vs fitted values plot showing generally random scatter around zero.

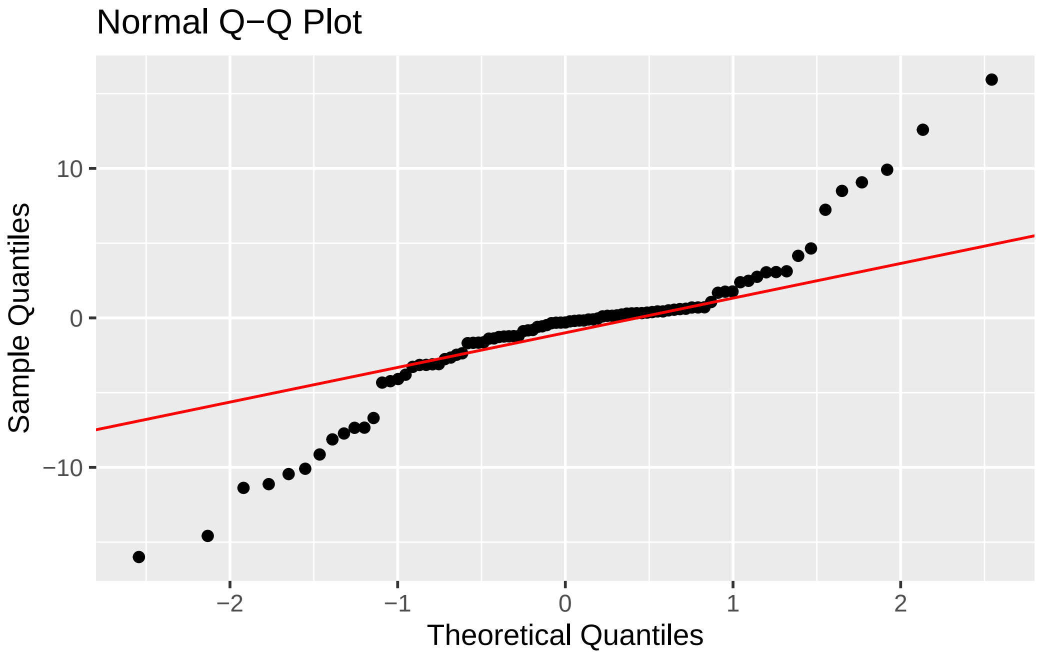

Figure 6: Q-Q plot of residuals showing the data mostly follows a normal distribution, except at the tails.

5 Discussion

In conclusion this paper was looking to examine whether ethnic and geographic factors can predict the proportion of workers who drive alone to work in California. It did this by looking at a data set of transportation method by ethnicity and geography from 2006-2010. A weighted regression model was then fit to this data which yielded a weighted R^2 of 0.96. Also two F-tests at a = 0.05 were conducted to assess whether ethnicity and region significantly impact drive-alone rates. The ethnicity hypothesis test yielded a p value of 5.85×10^{-41} and the region hypothesis test yielded a p value of 6.62×10^{-37}. From the results of the regression model and hypothesis test we can conclude there is enough evidence to conclude that both ethnicity and geography are significant predictors of solo commuting.

This analysis examined whether ethnic and geographic factors can predict the proportion of workers who drive alone to work in California. Using multiple linear regression with log-transformed population data, I found that both demographic composition and geographic location are statistically significant predictors of solo driving behavior. This remains true even after controlling for population size and temporal trends. The negative coefficient on log population(-1.49) suggests that a doubling of the population is associated with an approximate 1.49 percentage point decrease in the solo driving rate. However the small effect suggests that increasing population density may not shift transportation patterns significantly without changes to transportation infrastructure.

The regional variation in solo driving rates has important implications for transportation planning. Some MPO regions show persistently higher solo driving rates regardless of demographics, suggesting that infrastructure and local policies play crucial roles. So to reduce vehicle emissions policymakers cannot just address demographic shifts but must also address the transportation options available in specific regions. Ethnic differences in transportation behavior, while statistically significant, should be interpreted carefully. These differences may reflect not only cultural preferences but also the geographic distribution of ethnic groups, as some regions having better access to public transportation than others. Future research could disentangle these factors by incorporating data on transit accessibility.

There are a three key weaknesses to this study. First aggregating data at the MPO region level potentially hides important within-region variation. Solo driving rates in downtown San Francisco probably differ significantly from suburban areas of the same MPO region. However this analysis cannot capture these differences. Second, temporal coverage is limited to 2006-2010, missing more recent trends such as the rise of ride-sharing services and increased remote work. The COVID-19 pandemic also altered commuting behavior, making these historical patterns potentially less relevant for current policy. Lastly, the independence assumption may be violated due to nearby regions likely have similar transportation patterns due to shared infrastructure and economic conditions.

Potential improvements on this study could be incorporating data on public transit infrastructure. This would help explain the reason behind regional differences and yield more immediately actionable results for transportation planning. Additionally expanding the time range to include more recent data would reveal whether the relationships observed in 2006-2010 remain stable or have shifted due to technological and social changes. Finally, extending this framework to analyze other transportation modes could reveal whether the same ethnic and geographic factors influence different commuting choices.

Works Cited

Fernandez, Lucia. 2024. "US Transportation CO2 Emissions by State." https://www.statista.com/statistics/1100175/transportation-co2-emissions-in-the-us-by-state.

Health Equity, Office of. 2025. "Transportation to Work - Dataset - California Open Data." *California Open Data Portal*. https://data.ca.gov/dataset/transportation-to-work.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.