Union Membership Trends by Demographics and Industry: 1973–2024

Final Project Written Report

Neeka Lucas

Department of Statistics, University of California, Santa Cruz

STAT 155: Methods in Data Science

Professor Marcela Alfaro-Córdoba

June 9, 2025

Introduction

After spending the last four years at the University of California, Santa Cruz, I've become very familiar with labor strikes. These strikes occur nearly every year, and the presence of unions on campus is undeniable. While my understanding of labor and union rights began on campus, recent federal developments have shown broader, systemic concerns involving those rights. In April, a whistleblower revealed that DOGE, a Trump administration initiative led by Elon Musk, may have accessed and exfiltrated sensitive data from the National Labor Relations Board (McLaughlin, 2025). My project examines union membership differences between gender, sector, industry, and education from 1973 to 2024. Tracking trends in union membership data could reveal who is most at risk.

Data Wrangling

For this project, I used data that was originally collected by the U.S. Census Bureau and Bureau of Labor Statistics (BLS) through the Current Population Survey (CPS). For high earning individuals, researchers used Pareto distribution estimates (Hirsch et al., 2025). You can find the data at https://www.unionstats.com/ under "By Sector and Demographic Group: 1973-2024". The data came in 11 separate CSV files, each involving a different demographic variable: education level (less than college and college graduate), gender (female and male), industry (public administration, wholesale, and manufacturing), and sector (public and private). To clean the data, I skipped non-data header rows, renamed columns, removed copyright footer rows, added identifiers like “college” vs “less_college”, and converted percentage columns from text to numeric values. Then, I merged the datasets into three: education, sex, and industry. All of this was done in R, using these tidyverse packages: readr and tibble.

Exploratory Data Analysis (Project II)

The variables of interest I used to explore membership differences across demographics were: year, perc_mem, per_coverage, industry, sex, and education.

Figure 1 shows union membership by industry, sector, and sex. Public sector and public administration industries were the highest, and males had higher rates than females across all industries.

Figure 1

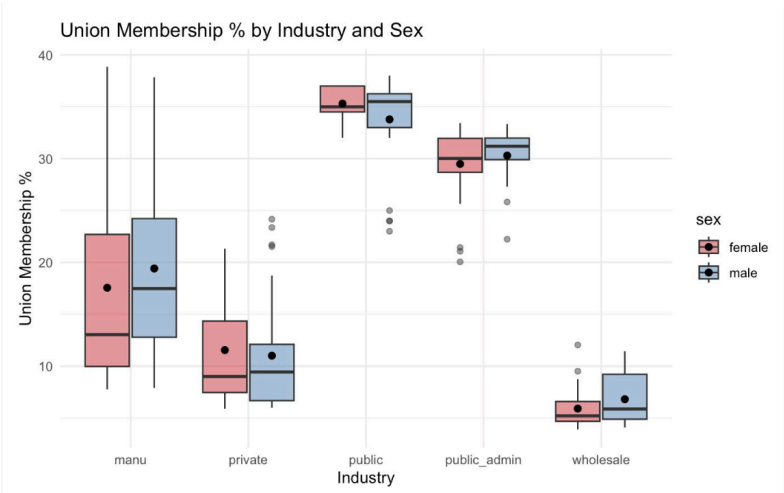
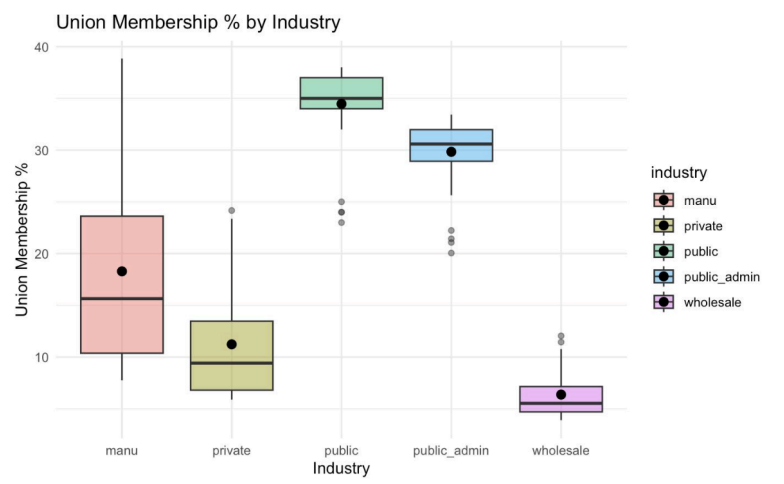


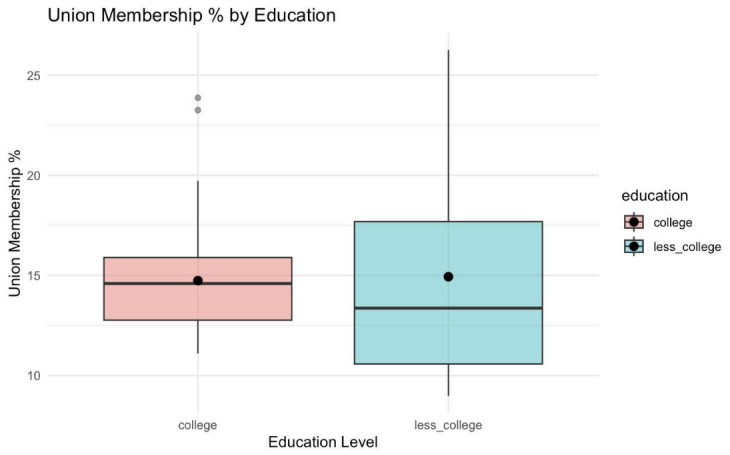
Figure 2 shows industry and sector data, with wholesale industry and private sector at the lowest union membership percentage.

Figure 2



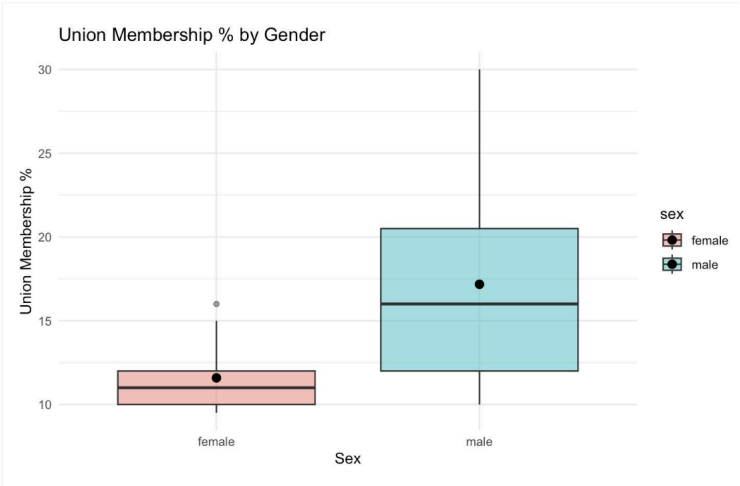
In Figure 3, comparing union membership by education, there was little difference between college and less than college.

Figure 3



In Figure 4, comparing union membership by gender, the union membership for males was higher.

Figure 4



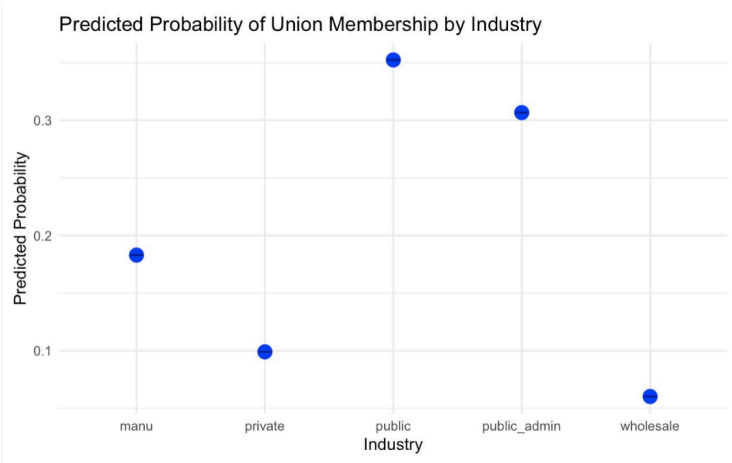
Modeling (Project III)

For this part of the project, I used logistic and linear regression models to track trends in union membership percentage across industry, sex, and education. Logistic regression modeled

the probability (0 to 1), of being a union member. While linear regression modeled the percentage of union membership.

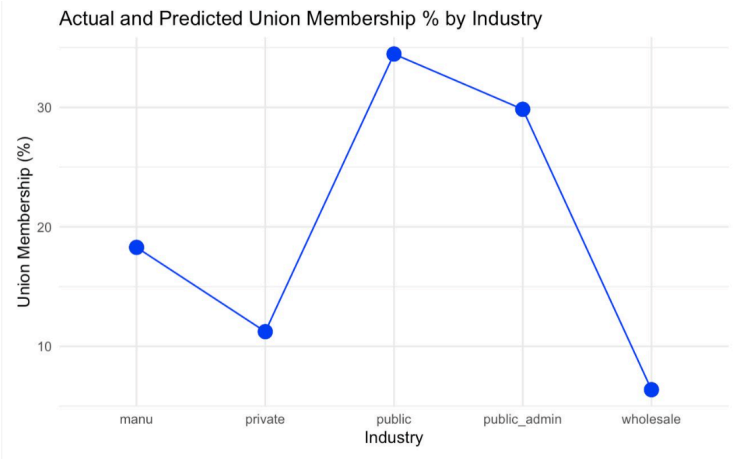
In Figure 5, the predicted probability of union membership by industry is shown. Public sector and public administration industries had the highest predicted probabilities of over 30%. Wholesale industries and the private sector had the lowest of under 10%.

Figure 5



In Figure 6, the actual and predicted union membership percentages by industry and sector are shown. Similarly, public sector and public administration had the highest percentages while wholesale industries and private sector had the lowest.

Figure 6



Across both models, the findings were consistent with industry having the largest effect.

Monte Carlo Simulation (Design)

I tested how different types of statistical distributions, used to produce simulated data, affect the results of linear and logistic regression models for predicting union membership. The data was generated through a simulation of the following distributions: gamma, beta, normal, binomial, poisson, exponential, chi-squared, t, F, and uniform. I used the following mathematical notations:

- linear regression: $y = B_0 + B_1*x_1 + B_2*x_2 + ... + e$
- logistic regression: $\ln (p/1-p) = B_0 + B_1*x_1 + B_2*x_2 + ... + B_k*x_k$

The assumptions are that whichever statistical distribution simulates data that is closest to the actual data has the best effect on results of linear and logistic regression models for predicting union membership. We can assume that all values are scaled and numeric, and that the sample size remains consistent throughout all simulations. For each distribution, there was a simulated sample size of $n = 100$. I used linear and logistic regression models to evaluate estimates of B_1, B_2, Type I error, and MSE. I hypothesized that distributions with outliers and skewness will produce less accurate estimates for bias of B_1, B_2, and MSE, and produce low performing linear and logistic models. Bias of B_1 and B_2 will be used to evaluate distance between the true values and the estimated coefficients. MSE of predicted union membership percentages for linear regression and predicted probabilities (0-1) for logistic regression. Some anticipated challenges or limitations may include: the actual sample of data being relatively small, and since there will be a total of 1,000 simulations (100 simulations per 10 distributions) the run time may be very long.

Monte Carlo Simulation (Results)

For linear regression, the best performing distributions were normal, Poisson, and uniform because of how low the bias and MSE were. Skewed distributions like t, chi-squared, and beta had more error. For logistic regression, accuracy was highest under normal and exponential distributions and was lowest under F and chi-squared distributions. The key findings from the Monte Carlo Simulation was that model performance depended heavily on the shape of the distribution, and normality mattered with small samples. Figure 7 illustrates a summary of the results.

Figure 7

distribution <chr>	lin_bias_b1 <dbl>	lin_bias_b2 <dbl>	lin_mse <dbl>	lin_type1_error <dbl>	lin_type1_error <dbl>	log_bias_b1 <dbl>	log_bias_b2 <dbl>	log_mse <dbl>	log_type1_error <dbl>
gamma	0.0105855475	−0.0191502994	0.9435197	0.04000000	0.04000000	0.20050971	−0.2338297	0.01137612	0.04333333
beta	0.0248115564	−0.0955204447	0.9491660	0.04666667	0.04666667	0.08826203	−0.3534083	0.01391430	0.05666667
normal	−0.0004480955	−0.0291448730	0.9370706	0.06000000	0.06000000	0.20009309	−0.2893938	0.01064469	0.05000000
binomial	−0.0102854024	0.0006369595	0.9219547	0.04333333	0.04333333	0.14573958	−0.1601078	0.01202230	0.04333333
poisson	−0.0016080369	0.0036393141	0.9408217	0.02666667	0.02666667	0.63108816	−0.7368962	0.01000439	0.06333333
exponential	0.0091110254	−0.0028943028	0.9425673	0.06000000	0.06000000	0.12822021	−0.3199380	0.01125477	0.04666667
chi_squared	0.0066217889	0.0044802072	0.9444686	0.03666667	0.03666667	1.16922703	−1.4223077	0.01040168	0.05666667
t	0.0141838883	0.0124566731	0.9559187	0.05333333	0.05333333	0.23472089	−0.2821159	0.01133908	0.05666667
F	0.0010898876	−0.0004642307	0.9403978	0.04666667	0.04666667	0.48201704	−0.7295598	0.01101808	0.03000000
uniform	0.0069140243	−0.0319284449	0.9251926	0.06666667	0.06666667	0.12896599	−0.1732095	0.01426070	0.04666667

Summary and Reflection

To wrap everything up, let's return to the research question: How has union membership differed across gender, sector, industry, and education from 1973-2024? Public sectors had the highest union membership, and private sector and wholesale industries had the lowest. Gender differences were consistent, as males had higher union membership than females. Education level had very little influence on union membership. From this project, I learned how to clean and merge large, complex datasets across multiple demographics. I also learned how to create a Monte Carlo Simulation, gaining a deeper understanding on how to interpret bias, MSE, and type 1 error. Through the reproducibility test, I learned the importance of writing clean and executable code that others can follow. Finally, this project has given me more insights on the ethical

implications of modeling. The DOGE breach is a reminder of how data is handled unethically, as it puts entire communities at risk.

References

Hirsch, Barry T., David A. Macpherson, and William E. Even (2025). Union Membership, Coverage, and Earnings from the CPS. https://unionstats.com

Hirsch, B. T., Macpherson, D. A., & Even, W. E. (2025). Union Membership, Coverage, Density, and Employment: Manufacturing Workers[Data set]. Copyright by Barry T. Hirsch, David A. Macpherson, and William E. Even

Hirsch, B. T., Macpherson, D. A., & Even, W. E. (2025). Union Membership, Coverage, Density, and Employment: Wholesale and Retail Trade Workers[Data set]. Copyright by Barry T. Hirsch, David A. Macpherson, and William E. Even

Hirsch, B. T., Macpherson, D. A., & Even, W. E. (2025). Union Membership, Coverage, Density, and Employment: Public Administration[Data set]. Copyright by Barry T. Hirsch, David A. Macpherson, and William E. Even

Hirsch, B. T., Macpherson, D. A., & Even, W. E. (2025). Union Membership, Coverage, Density, and Employment: Private Sector Workers[Data set]. Copyright by Barry T. Hirsch, David A. Macpherson, and William E. Even

Hirsch, B. T., Macpherson, D. A., & Even, W. E. (2025). Union Membership, Coverage, Density, and Employment: Public Sector Workers[Data set]. Copyright by Barry T. Hirsch, David A. Macpherson, and William E. Even

Hirsch, B. T., Macpherson, D. A., & Even, W. E. (2025). Union Membership, Coverage, Density, and Employment: Workers with less than Bachelor's Degree[Data set]. Copyright by Barry T. Hirsch, David A. Macpherson, and William E. Even

Hirsch, B. T., Macpherson, D. A., & Even, W. E. (2025). Union Membership, Coverage, Density, and Employment: Workers with Bachelor's Degree or More[Data set]. Copyright by Barry T. Hirsch, David A. Macpherson, and William E. Even

Hirsch, B. T., Macpherson, D. A., & Even, W. E. (2025). Union Membership, Coverage, Density, and Employment: Male Workers [Data set]. Copyright by Barry T. Hirsch, David A. Macpherson, and William E. Even

Hirsch, B. T., Macpherson, D. A., & Even, W. E. (2025). Union Membership, Coverage, Density, and Employment: Female Workers[Data set]. Copyright by Barry T. Hirsch, David A. Macpherson, and William E. Even

Macpherson, D. A., & Hirsch, B. T. (2023). Five decades of CPS wages, methods, and union‐nonunion wage gaps at unionstats.com. Industrial Relations: A Journal of Economy and Society, 62(4), 439–452. https://doi.org/10.1111/irel.12330

McLaughlin, J. (2025, April 15). A whistleblower’s disclosure details how Doge may have taken sensitive labor data. NPR. https://www.npr.org/2025/04/15/nx-s1-5355896/doge-nlrb-elon-musk-spacex-security