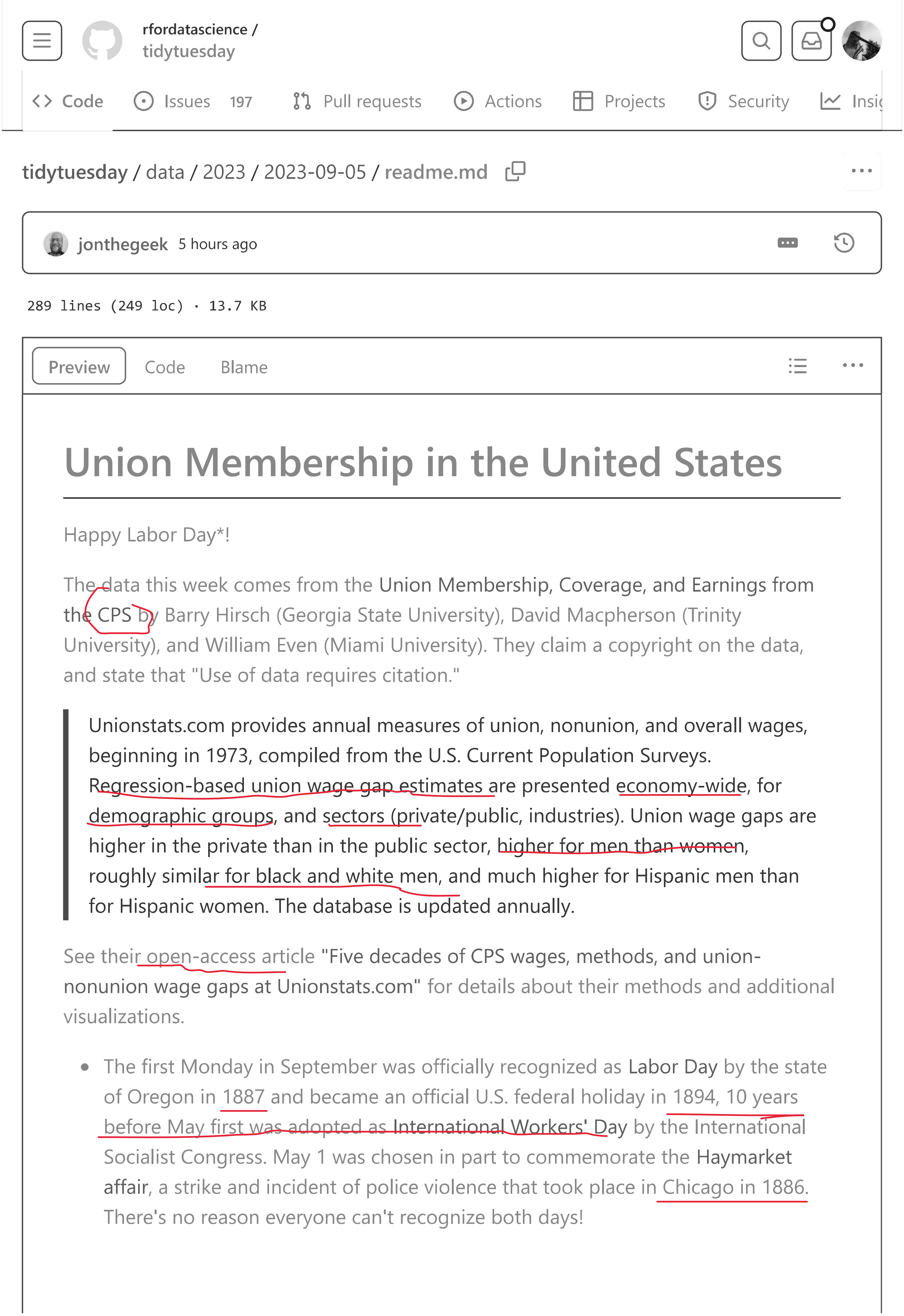

tidytuesday / data / 2023 / 2023-09-05 / readme.md

jonthegeek 5 hours ago

289 lines (249 loc) · 13.7 KB

# Union Membership in the United States

Happy Labor Day*!

The data this week comes from the Union Membership, Coverage, and Earnings from the CPS by Barry Hirsch (Georgia State University), David Macpherson (Trinity University), and William Even (Miami University). They claim a copyright on the data, and state that "Use of data requires citation."

> Unionstats.com provides annual measures of union, nonunion, and overall wages, beginning in 1973, compiled from the U.S. Current Population Surveys. Regression-based union wage gap estimates are presented economy-wide, for demographic groups, and sectors (private/public, industries). Union wage gaps are higher in the private than in the public sector, higher for men than women, roughly similar for black and white men, and much higher for Hispanic men than for Hispanic women. The database is updated annually.

See their open-access article "Five decades of CPS wages, methods, and union-nonunion wage gaps at Unionstats.com" for details about their methods and additional visualizations.

- The first Monday in September was officially recognized as Labor Day by the state of Oregon in 1887 and became an official U.S. federal holiday in 1894, 10 years before May first was adopted as International Workers' Day by the International Socialist Congress. May 1 was chosen in part to commemorate the Haymarket affair, a strike and incident of police violence that took place in Chicago in 1886. There's no reason everyone can't recognize both days!

## The Data

```
# Option 1: tidytuesdayR package
## install.packages("tidytuesdayR")

tuesdata <- tidytuesdayR::tt_load('2023-09-05')
## OR
tuesdata <- tidytuesdayR::tt_load(2023, week = 36)

demographics <- tuesdata$demographics
wages <- tuesdata$wages
states <- tuesdata$states

# Option 2: Read directly from GitHub

demographics <-
readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytues
09-05/demographics.csv')
wages <-
readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytues
09-05/wages.csv')
states <-
readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytues
09-05/states.csv')
```

## How to Participate

- Explore the data, watching out for interesting relationships. We would like to emphasize that you should not draw conclusions about **causation** in the data. There are various moderating variables that affect all data, many of which might not have been captured in these datasets. As such, our suggestion is to use the data provided to practice your data tidying and plotting techniques, and to consider for yourself what nuances might underlie these relationships.
- Create a visualization, a model, a shiny app, or some other piece of data-science-related output, using R or another programming language.
- Share your output and the code used to generate it on social media with the #TidyTuesday hashtag.

### Data Dictionary

## `demographics.csv`

Data sources:

- 1973-1981: May Current Population Survey (CPS)
- 1982: No union questions available
- 1983-2022: CPS Outgoing Rotation Group (ORG) Earnings Files

The definition of union membership was expanded in 1977 to include "employee associations similar to a union".

| variable | class | description |
|---|---|---|
| year | double | When the data was collected. |
| sample_size | double | The number of wage and salary workers ages 16 and over who were surveyed. |
| employment | double | Wage and salary employment in thousands. |
| members | double | Employed workers who are union members in thousands. |
| covered | double | Workers covered by a collective bargaining agreement in thousands. |
| p_members | double | Percent of employed workers who are union members. |
| p_covered | double | Percent of employed workers who are covered by a collective bargaining agreement. |
| facet | character | The sector or demographic group contained in this row of data. |

## wages.csv

Data sources:

- 1973-1981: May Current Population Survey (CPS)
- 1982: No union questions available
- 1983-2022: CPS Outgoing Rotation Group (ORG) Earnings Files

The definition of union membership was expanded in 1977 to include "employee associations similar to a union".

| variable | class | description |
|---|---|---|
| year | double | When the data was collected. |

| variable | class | description |
|---|---|---|
| sample_size | double | The number of wage and salary workers ages 16 and over who were surveyed and provided earnings and hours worked information. |
| wage | double | Mean hourly earnings in nominal dollars. |
| at_cap | double | Percent of workers with weekly earnings at the top code of $999 through 1988, $1923 in 1989-97, and $2885 beginning in 1998, with individuals assigned mean earnings above the cap based on annual estimates of the gender-specific Pareto distribution. |
| union_wage | double | Mean wage among union members. |
| nonunion_wage | double | Mean wage among nonunion workers. |
| union_wage_premium_raw | double | The percentage difference between the union and nonunion wage. |
| union_wage_premium_adjusted | double | Estimated as exp(b)-1 where b is the regression coefficient on a union membership variable (equal to 1 if union and 0 otherwise) from a semi-logarithmic wage equation, with controls included for worker/job characteristics. Included in the all-worker wage equation are the control variables: years of schooling, potential years of experience [proxied by age minus years of schooling minus 6] and its square [both interacted with gender], and categorical variables for marital status, race and ethnicity, gender, part-time, large metropolitan area, state, public |

| variable | class | description |
|---|---|---|
| | | sector, broad industry, and broad occupation. Controls are omitted, as appropriate, for estimates within sectors or by demographic group [i.e., by class, gender, race, or industry sector]. Workers who do not report earnings but instead have them imputed [i.e., assigned] by the Census are removed from the estimation samples in all years, except 1994 and 1995 when imputed earners cannot be identified. Inclusion of imputed earners causes union wages to be understated, nonunion wages overstated, and union-nonunion wage differences understated. For 1994-95, the sample includes imputed earners and estimates in those years have been adjusted to remove the bias from imputation. |
| facet | character | The sector or demographic group contained in this row of data. |

## `states.csv`

Data source: Current Population Survey (CPS) Outgoing Rotation Group (ORG) Earnings Files

| variable | class | description |
|---|---|---|
| state_census_code | double | Census state code used in CPS |
| state | character | State name. |
| sector | character | Employment sector. |
| observations | double | CPS sample size. |
| employment | double | Wage and salary employment in thousands. |
| members | double | Employed workers who are union members in |

| variable | class | description |
|---|---|---|
| | | thousands. |
| covered | double | Workers covered by a collective bargaining agreement in thousands. |
| p_members | double | Percent of employed workers who are union members. |
| p_covered | double | Percent of employed workers who are covered by a collective bargaining agreement. |
| state_abbreviation | character | State abbreviation. |
| year | double | Year of the survey. |

## Cleaning Script

```
library(haven)
library(rvest)
library(tidyverse)
library(here)

working_dir <- here::here("data", "2023", "2023-09-05")

base_url <- "https://www.unionstats.com/"

members_url <- "https://www.unionstats.com/members/members_index.html"
members_dta_urls <- rvest::read_html(members_url) |>
  rvest::html_nodes("td:nth-child(4) a") |>
  rvest::html_attr("href") |>
  rvest::url_absolute(members_url)

# Fix typos
members_dta_urls <- stringr::str_replace_all(members_dta_urls, "wages", "me
double_public <- which(
  members_dta_urls == "https://www.unionstats.com/members/dta/members_publi
)
members_dta_urls[[double_public[[1]]]] <- "https://www.unionstats.com/membe

members_data <- purrr::map(
  members_dta_urls,
  \(url) {
    haven::read_dta(url) |>
      dplyr::rename(
        "sample_size" = "nobs",
        "employment" = "empl",
        "members" = "member",
        "p_members" = "pctmem",
        "p_covered" = "pctcov"
      ) |>
      dplyr::mutate(
        dplyr::across(
          c(year, sample_size),
          as.integer
        )
      ) |>
      dplyr::mutate(
        facet = stringr::str_extract(url, "members_([^.]+)\\.dta", group =
      )
  }
) |>
  purrr::list_rbind() |>
  dplyr::mutate(
    facet = dplyr::case_match(
      facet,
      "all" ~ "all wage and salary workers",
      "constr" ~ "construction",
      "manuf" ~ "manufacturing",
      "whole_ret" ~ "wholesale/retail",
      "trans_comm_util" ~ "transportation, communication, and utilities",
      "fire" ~ "finance, insurance, and real estate",
      "serv" ~ "services",
      "public_admin" ~ "public administration",
      "priv" ~ "private sector: all",
      "priv_nonag" ~ "private sector: nonagricultural",
      "priv_const" ~ "private sector: construction",
      "priv_manuf" ~ "private sector: manufacturing",
      "public" ~ "public sector: all",
      "fed" ~ "public sector: federal",
      "stateg" ~ "public sector: state government",
      "localg" ~ "public sector: local government",
      "postal" ~ "public sector: postal service",
      "police" ~ "public sector: police",
      "less_th_coll" ~ "demographics: less than college",
      "coll_plus" ~ "demographics: college or more",
      "male" ~ "demographics: male",
      "fem" ~ "demographics: female",
      "wh_male" ~ "demographics: white male",
      "wh_fem" ~ "demographics: white female",
      "bl_male" ~ "demographics: black male",
      "bl_fem" ~ "demographics: black female",
      "hisp_male" ~ "demographics: hispanic male",
      "hisp_fem" ~ "demographics: hispanic female",
    )
  )

state_url <- "https://www.unionstats.com/state/dta/state.dta"
state_data <- haven::read_dta(state_url) |>
  dplyr::select(
    "state_census_code" = "state_cens",
    "state",
    "sector",
    "observations" = "nobs",
    "employment" = "empl",
    "members" = "member",
    "covered",
    "p_members" = "pctmem",
    "p_covered" = "pctcov",
    "state_abbreviation" = "state2",
    "year"
  ) |>
  dplyr::mutate(
    dplyr::across(
      c(state_census_code, observations, year),
      as.integer
    )
  )

wages_url <- "https://www.unionstats.com/wages/wages_index.html"
wages_dta_urls <- rvest::read_html(wages_url) |>
  rvest::html_nodes("td:nth-child(4) a") |>
  rvest::html_attr("href") |>
  rvest::url_absolute(wages_url)

# One of the URLs has a typo on the site.
wages_dta_urls <- stringr::str_replace_all(wages_dta_urls, "members", "wage

wages_data <- purrr::map(
  wages_dta_urls,
  \(url) {
    haven::read_dta(url) |>
      dplyr::select(
        "year",
        "sample_size" = "nobs_wage",
        "wage",
        "at_cap" = "atcap",
        "union_wage" = "memwage",
        "nonunion_wage" = "nonwage",
        "union_wage_premium_raw" = "un_wage_prem_raw",
        "union_wage_premium_adjusted" = "un_wage_prem_adj"
      ) |>
      dplyr::mutate(
        dplyr::across(
          c(year, sample_size),
          as.integer
        )
      ) |>
      dplyr::mutate(
        facet = stringr::str_extract(url, "wages_([^.]+)\\.dta", group = 1)
      )
  }
) |>
  purrr::list_rbind() |>
  dplyr::mutate(
    facet = dplyr::case_match(
```

```r
      facet,
      "all" ~ "all wage and salary workers",
      "constr" ~ "construction",
      "manuf" ~ "manufacturing",
      "whole_ret" ~ "wholesale/retail",
      "trans_comm_util" ~ "transportation, communication, and utilities",
      "fire" ~ "finance, insurance, and real estate",
      "serv" ~ "services",
      "public_admin" ~ "public administration",
      "priv" ~ "private sector: all",
      "priv_nonag" ~ "private sector: nonagricultural",
      "priv_const" ~ "private sector: construction",
      "priv_manuf" ~ "private sector: manufacturing",
      "public" ~ "public sector: all",
      "fed" ~ "public sector: federal",
      "stateg" ~ "public sector: state government",
      "localg" ~ "public sector: local government",
      "postal" ~ "public sector: postal service",
      "police" ~ "public sector: police",
      "less_th_coll" ~ "demographics: less than college",
      "coll_plus" ~ "demographics: college or more",
      "male" ~ "demographics: male",
      "fem" ~ "demographics: female",
      "wh_male" ~ "demographics: white male",
      "wh_fem" ~ "demographics: white female",
      "bl_male" ~ "demographics: black male",
      "bl_fem" ~ "demographics: black female",
      "hisp_male" ~ "demographics: hispanic male",
      "hisp_fem" ~ "demographics: hispanic female",
    )
  )

readr::write_csv(
  members_data,
  here::here(working_dir, "demographics.csv")
)
readr::write_csv(
  wages_data,
  here::here(working_dir, "wages.csv")
)
readr::write_csv(
  state_data,
  here::here(working_dir, "states.csv")
)
```