Lab2_Huibin

Huibin Chang

2024-12-06

```
# Data pipeline

# US Census API library
library(tidycensus)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
#library(geojsonio)
# Library for shape files
#library(sf)
#library(scales)
#library(htmltools)
#library(htmlwidgets)
# Library to read Census shape files
library(tigris)
```

```
## To enable caching of data, set `options(tigris_use_cache = TRUE)`
## in your R script or .Rprofile.
```

```
#library(leaflet)
library(knitr)
# census_api_key('a22e52a2d1b1c6c403b1508183a23ce97b59172d', install=TRUE)


# Load ACS data for population estimates
# Population
population <- get_acs(
  geography = "county",
  variables = "B01003_001",  # Total population
  year = 2021,
  survey = "acs5") %>%
  rename(population = estimate) %>%
  select(-variable, -moe)
```

```
## Getting data from the 2017-2021 5-year ACS
```

```
# Number of workers
worker_population <- get_acs(
  geography = "county",
  variables = "B23025_005",
  year = 2021,
  survey = "acs5") %>%
  rename(workers = estimate) %>%
  select(-NAME, -variable, -moe)
```

```
## Getting data from the 2017-2021 5-year ACS
```

```
# Median wage income
wages <- get_acs(
  geography = "county",
  variables = "B20002_001",
  year = 2021,
  survey = "acs5") %>%
  rename(median_wage = estimate) %>%
  select(-NAME, -variable, -moe)
```

```
## Getting data from the 2017-2021 5-year ACS
```

```
# Household size
avg_household_size <- get_acs(
  geography = "county",
  variables = "B25010_001",
  year = 2021,
  survey = "acs5") %>%
  rename(household_size = estimate) %>%
  select(-NAME, -variable, -moe)
```

```
## Getting data from the 2017-2021 5-year ACS
```

```
# Household income
median_household_income <- get_acs(
  geography = "county",
  variables = "B19013_001",
  year = 2021,
  survey = "acs5") %>%
  rename(median_household_income = estimate) %>%
  select(-NAME, -variable, -moe)
```

```
## Getting data from the 2017-2021 5-year ACS
```

```
# Vehicle ownership
vehicle <- get_acs(
  geography = "county",
  variables = c(
    total_households = "B25044_001",
    no_vehicle = "B25044_003",
    one_vehicle = "B25044_004",
    two_vehicles = "B25044_005",
    three_vehicles = "B25044_006",
    four_or_more_vehicles = "B25044_007"),
  year = 2021,
  survey = "acs5") %>%
```

```
  select(GEOID, variable, estimate) %>%
  pivot_wider(names_from = variable,
              values_from = estimate) %>%
  mutate(vehicle_per_hh =
           (no_vehicle * 0 +
              one_vehicle * 1 +
              two_vehicles * 2 +
              three_vehicles * 3 +
              four_or_more_vehicles * 4) /
            total_households) %>%
  select(GEOID, vehicle_per_hh)
```

```
## Getting data from the 2017-2021 5-year ACS
```

```
# Education
education<- get_acs(
  geography = "county",
  variables = c(
    total_population_25_over = "B15003_001",
    bachelor = "B15003_022",
    master = "B15003_023",
    professional = "B15003_024",
    doctoral = "B15003_025"),
  year = 2021,
  survey = "acs5") %>%
  select(GEOID, variable, estimate) %>%
  pivot_wider(names_from = variable,
              values_from = estimate) %>%
  mutate(
    college_or_higher = bachelor +
      master +
      professional +
      doctoral,
    proportion_college_or_higher = college_or_higher / total_population_25_over) %>%
  select(GEOID,
         college_or_higher,
         proportion_college_or_higher)
```

```
## Getting data from the 2017-2021 5-year ACS
```

```
# Housing price
housing_values <- get_acs(
  geography = "county",
  variables = c(median_housing_value = "B25077_001"),
  year = 2021,
  survey = "acs5") %>%
  select(GEOID, estimate) %>%
  rename(median_housing_price = estimate)
```

```
## Getting data from the 2017-2021 5-year ACS
```

```
# Land area of all counties
options(tigris_use_cache = TRUE)  # Cache shapefiles for reuse

# Download county shapefiles
counties <- counties(year = 2021)
```

```r
# Calculate land area in square kilometers
counties <- counties %>%
  mutate(land_area_sqkm = ALAND / 1e6)  # m² to km²

county_land_area <- counties %>%
  select(GEOID, land_area_sqkm)



##
# Putting everything together
##
census_county <- population %>%
  left_join(worker_population, by = 'GEOID') %>%
  left_join(wages, by = 'GEOID') %>%
  left_join(avg_household_size, by='GEOID') %>%
  left_join(median_household_income, by = 'GEOID') %>%
  mutate(per_person_income = median_household_income /
           household_size) %>%
  left_join(vehicle, by = 'GEOID') %>%
  left_join(education, by = 'GEOID') %>%
  left_join(housing_values, by = 'GEOID') %>%
  left_join(county_land_area, by = 'GEOID') %>%
  mutate(worker_density = workers /
           land_area_sqkm,
         ln_wage = log(median_wage),
         ln_density = log(worker_density),
         ln_housing_price = log(median_housing_price),
         ln_income = log(per_person_income))

census_county<-
  census_county[census_county$ln_density != -Inf
                & !is.na(census_county$ln_wage), ]
census_county
```

```
## # A tibble: 3,198 x 19
##    GEOID NAME                      population workers median_wage household_size
##    <chr> <chr>                          <dbl>   <dbl>       <dbl>          <dbl>
##  1 01001 Autauga County, Alabama        58239     752       35154           2.64
##  2 01003 Baldwin County, Alabama       227131    3994       35999           2.57
##  3 01005 Barbour County, Alabama        25259     808       27623           2.45
##  4 01007 Bibb County, Alabama           22412     884       28108           2.96
##  5 01009 Blount County, Alabama         58884    1554       35567           2.74
##  6 01011 Bullock County, Alabama        10386     118       27256           2.92
##  7 01013 Butler County, Alabama         19181     530       27892           2.89
##  8 01015 Calhoun County, Alabama       116425    3702       30506           2.56
##  9 01017 Chambers County, Alabama       34834     504       30253           2.61
## 10 01019 Cherokee County, Alabama       24975     483       33593           2.55
## # i 3,188 more rows
## # i 13 more variables: median_household_income <dbl>, per_person_income <dbl>,
## #   vehicle_per_hh <dbl>, college_or_higher <dbl>,
## #   proportion_college_or_higher <dbl>, median_housing_price <dbl>,
## #   land_area_sqkm <dbl>, geometry <MULTIPOLYGON [°]>, worker_density <dbl>,
```

```
## #   ln_wage <dbl>, ln_density <dbl>, ln_housing_price <dbl>, ln_income <dbl>
```

Splitting, randomly choose 30% as the exploration data set

```
set.seed(1)
shuffled_census_county <- census_county[sample(nrow(census_county)), ]

train_size = 0.3
train_rows = floor(train_size * nrow(census_county))

train_data = shuffled_census_county[1:train_rows, ]
test_data = shuffled_census_county[(train_rows + 1):nrow(census_county), ]
```

```
mod1 <-
  lm(median_wage ~ worker_density, data = train_data)
summary(mod1)
```

```
##
## Call:
## lm(formula = median_wage ~ worker_density, data = train_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -20725  -3658   -688   3484  45309
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    33701.117    225.302 149.582   <2e-16 ***
## worker_density    18.832      7.617   2.472   0.0136 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6912 on 957 degrees of freedom
## Multiple R-squared:  0.006346,   Adjusted R-squared:  0.005308
## F-statistic: 6.112 on 1 and 957 DF,  p-value: 0.0136
```

```
# mod2:
mod2 <-
  lm(ln_wage ~ ln_density, data = train_data)
summary(mod2)
```

```
##
## Call:
## lm(formula = ln_wage ~ ln_density, data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.95683 -0.09129  0.00153  0.12072  0.81724
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 10.417798   0.007174 1452.068  < 2e-16 ***
## ln_density   0.013982   0.003260    4.289 1.98e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.2065 on 957 degrees of freedom
## Multiple R-squared:  0.01886,    Adjusted R-squared:  0.01783
## F-statistic: 18.39 on 1 and 957 DF,  p-value: 1.978e-05
```

```
# mod3:
mod3 <-
  lm(ln_wage ~ ln_density + proportion_college_or_higher, data = train_data)
summary(mod3)
```

```
##
## Call:
## lm(formula = ln_wage ~ ln_density + proportion_college_or_higher,
##     data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.04318 -0.07360  0.02098  0.11085  0.69690
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  10.183469   0.017790 572.419   <2e-16 ***
## ln_density                   -0.005198   0.003260  -1.594    0.111
## proportion_college_or_higher  0.944467   0.066704  14.159   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1878 on 956 degrees of freedom
## Multiple R-squared:  0.1889, Adjusted R-squared:  0.1872
## F-statistic: 111.4 on 2 and 956 DF,  p-value: < 2.2e-16
```

```
# mod4:
mod4 <-
  lm(ln_wage ~ ln_density + proportion_college_or_higher + ln_housing_price, data = train_data)
summary(mod4)
```

```
##
## Call:
## lm(formula = ln_wage ~ ln_density + proportion_college_or_higher +
##     ln_housing_price, data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.92242 -0.07283  0.02793  0.10610  0.52044
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   7.457233   0.199487  37.382  < 2e-16 ***
## ln_density                   -0.010964   0.003011  -3.641 0.000286 ***
## proportion_college_or_higher  0.141232   0.084575   1.670 0.095267 .
## ln_housing_price              0.244284   0.017815  13.712  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1718 on 955 degrees of freedom
## Multiple R-squared:  0.3224, Adjusted R-squared:  0.3202
```

```
## F-statistic: 151.4 on 3 and 955 DF,  p-value: < 2.2e-16
```

```
# mod5:
mod5 <-
  lm(ln_wage ~ ln_density + college_or_higher + ln_housing_price, data = train_data)
summary(mod5)
```

```
##
## Call:
## lm(formula = ln_wage ~ ln_density + college_or_higher + ln_housing_price,
##     data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.87845 -0.07602  0.02693  0.10652  0.53704
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        7.418e+00  1.582e-01  46.877  < 2e-16 ***
## ln_density        -1.430e-02  3.105e-03  -4.607 4.64e-06 ***
## college_or_higher  3.186e-07  7.581e-08   4.203 2.89e-05 ***
## ln_housing_price   2.494e-01  1.327e-02  18.793  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1705 on 955 degrees of freedom
## Multiple R-squared:  0.3327, Adjusted R-squared:  0.3306
## F-statistic: 158.7 on 3 and 955 DF,  p-value: < 2.2e-16
```

```
library(stargazer)
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```
# Example of a few regression models with different independent variables

# Create a table with stargazer
stargazer(mod2, mod3, mod4, mod5, type = "latex",
          results = 'asis',
          column.labels = c("Model 2", "Model 3", "Model 4", "Model 5"),
          digits = 3, out = "model_comparison.txt")
```

```
##
## % Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac a
## % Date and time: Fri, Dec 06, 2024 - 00:25:19
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lcccc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & \multicolumn{4}{c}{\textit{Dependent variable:}} \\
## \cline{2-5}
```

```
## \\[-1.8ex] & \multicolumn{4}{c}{ln\_wage} \\
##  & Model 2 & Model 3 & Model 4 & Model 5 \\
## \\[-1.8ex] & (1) & (2) & (3) & (4)\\
## \hline \\[-1.8ex]
##  ln\_density & 0.014$^{***}$ & $-$0.005 & $-$0.011$^{***}$ & $-$0.014$^{***}$ \\
##   & (0.003) & (0.003) & (0.003) & (0.003) \\
##   & & & & \\
##  proportion\_college\_or\_higher &  & 0.944$^{***}$ & 0.141$^{*}$ &  \\
##   &  & (0.067) & (0.085) &  \\
##   & & & & \\
##  college\_or\_higher &  &  &  & 0.00000$^{***}$ \\
##   &  &  &  & (0.00000) \\
##   & & & & \\
##  ln\_housing\_price &  &  & 0.244$^{***}$ & 0.249$^{***}$ \\
##   &  &  & (0.018) & (0.013) \\
##   & & & & \\
##  Constant & 10.418$^{***}$ & 10.183$^{***}$ & 7.457$^{***}$ & 7.418$^{***}$ \\
##   & (0.007) & (0.018) & (0.199) & (0.158) \\
##   & & & & \\
## \hline \\[-1.8ex]
## Observations & 959 & 959 & 959 & 959 \\
## R$^{2}$ & 0.019 & 0.189 & 0.322 & 0.333 \\
## Adjusted R$^{2}$ & 0.018 & 0.187 & 0.320 & 0.331 \\
## Residual Std. Error & 0.206 (df = 957) & 0.188 (df = 956) & 0.172 (df = 955) & 0.170 (df = 955) \\
## F Statistic & 18.395$^{***}$ (df = 1; 957) & 111.353$^{***}$ (df = 2; 956) & 151.429$^{***}$ (df = 3
## \hline
## \hline \\[-1.8ex]
## \textit{Note:}  & \multicolumn{4}{r}{$^{*}$p$<$0.1; $^{**}$p$<$0.05; $^{***}$p$<$0.01} \\
## \end{tabular}
## \end{table}
##
## % Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac a
## % Date and time: Fri, Dec 06, 2024 - 00:25:19
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}} c}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
## asis \\
## \hline \\[-1.8ex]
## \end{tabular}
## \end{table}
```

$ $