EDS241: Assignment 1

Wylie Hampson

01/14/2022

In this assignment, we use data from CalEnviroScreen 4.0, a mapping data tool produced by the California Office of Environmental Health Hazards Assessment (OEHHA). The data are compiled and constructed from a variety of sources and cover all 8,035 census tracts in California. Source: https://oehha.ca.gov/calenviroscreen/report/calenviroscreen-40

1 Clean and plot data

The following code loads and cleans the data, then selects the columns that we are interested in.

```
# Load data

data <- read_excel(here("data", "CES4.xlsx"), na = "NA")
data
```

```
## # A tibble: 8,035 x 58
##    `Census Tract` `Total Population` `California County`   ZIP `Approximate Loc~
##             <dbl>              <dbl> <chr>                <dbl> <chr>
##  1     6019001100               2780 Fresno               93706 Fresno
##  2     6077000700               4680 San Joaquin          95206 Stockton
##  3     6037204920               2751 Los Angeles          90023 Los Angeles
##  4     6019000700               3664 Fresno               93706 Fresno
##  5     6019000200               2689 Fresno               93706 Fresno
##  6     6037542402               3306 Los Angeles          90221 Compton
##  7     6019001000               4255 Fresno               93706 Fresno
##  8     6037543202               5124 Los Angeles          90220 Compton
##  9     6019001202               4561 Fresno               93725 Unincorporated F~
## 10     6077000100               3688 San Joaquin          95202 Stockton
## # ... with 8,025 more rows, and 53 more variables: Longitude <dbl>,
## #   Latitude <dbl>, CES 4.0 Score <dbl>, CES 4.0 Percentile <dbl>,
## #   CES 4.0 Percentile Range <chr>, Ozone <dbl>, Ozone Pctl <dbl>, PM2.5 <dbl>,
## #   PM2.5 Pctl <dbl>, Diesel PM <dbl>, Diesel PM Pctl <dbl>,
## #   Drinking Water <dbl>, Drinking Water Pctl <dbl>, Lead <dbl>,
## #   Lead Pctl <dbl>, Pesticides <dbl>, Pesticides Pctl <dbl>,
## #   Tox. Release <dbl>, Tox. Release Pctl <dbl>, Traffic <dbl>, ...
```

```
# Clean data

data <- data %>% clean_names()
```

```r
# Select the columns that we are interested in.

data <- data %>%
  select(
    census_tract,
    total_population,
    california_county,
    low_birth_weight,
    pm2_5,
    poverty
  )
```

Question a: What is the average concentration of PM2.5 across all census tracts in California?

```r
mean_pm <- round(mean(data$pm2_5), 3)
mean_pm
```

```
## [1] 10.153
```

The mean PM2.5 concentration across all census tracts in California is **10.153** $\mu g/m^3$.

Question b: What county has the highest level of poverty in California?

```r
poverty_by_county <- data %>%
  group_by(california_county) %>%
  summarize(county_poverty = mean(poverty, na.rm = TRUE)) %>%
  arrange(desc(county_poverty)) %>%
  head()

highest_poverty_rate <- round(max(poverty_by_county$county_poverty), 1)
highest_poverty_county <- poverty_by_county$california_county[1]

poverty_by_county
```

```
## # A tibble: 6 x 2
##   california_county county_poverty
##   <chr>                      <dbl>
## 1 Tulare                      51.8
## 2 Del Norte                   48.4
## 3 Imperial                    47.8
## 4 Merced                      47.3
## 5 Kern                        47.2
## 6 Fresno                      45.8
```

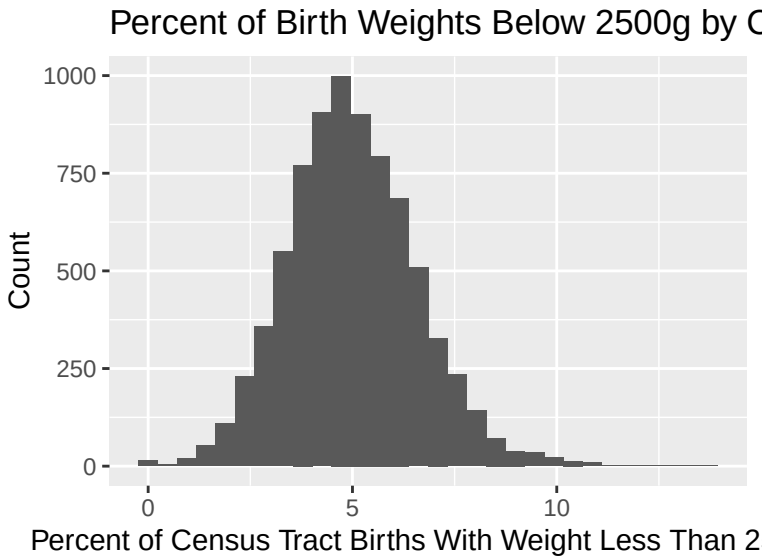
In the above summary table we can see that the California county that has the highest rate of poverty is Tulare with a poverty rate of 51.8%.

Question c: Make a histogram depicting the distribution of percent low birth weight and PM2.5.

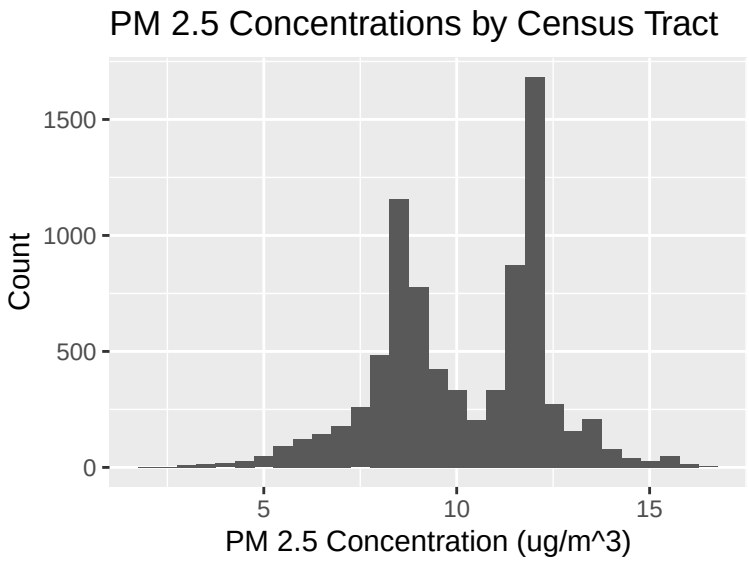
```
birth_hist <- ggplot(data = data, aes(x = low_birth_weight)) +
  geom_histogram() +
  labs(x = "Percent of Census Tract Births With Weight Less Than 2500g",
       y = "Count",
       title = "Percent of Birth Weights Below 2500g by Census Tract")


pm_hist <- ggplot(data = data, aes(x = pm2_5)) +
  geom_histogram() +
  labs(x = "PM 2.5 Concentration (ug/m^3)",
       y = "Count",
       title = "PM 2.5 Concentrations by Census Tract")

birth_hist
```




```
pm_hist
```



Question d: Estimate a OLS regression of low_birth_weight on pm2_5. Report the estimated slope coefficient and its heteroskedasticity-robust standard error. Interpret the estimated slope coefficient. Is the effect of PM25 on LowBirthWeight statistically significant at the 5%?

```
model_1 <- lm_robust(formula = low_birth_weight ~ pm2_5, data = data)
summary(model_1)
```

```
##
## Call:
## lm_robust(formula = low_birth_weight ~ pm2_5, data = data)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value
## (Intercept)   3.8010   0.088583   42.91
## pm2_5         0.1179   0.008402   14.04
##                                                      Pr(>|t|) CI Lower CI Upper
## (Intercept) 0.00000000000000000000000000000000000000000000000   3.6273   3.9746
## pm2_5       0.00000000000000000000000000000000000000003256   0.1015   0.1344
##               DF
## (Intercept) 7806
## pm2_5       7806
##
## Multiple R-squared:  0.02499 ,   Adjusted R-squared:  0.02486
## F-statistic:   197 on 1 and 7806 DF,  p-value: < 0.00000000000000022
```

Here the estimated slope coefficient is 0.1179, meaning that with every unit increase of PM 2.5 concentration, we would expect to see an increase of 0.1179% of babies that are born under 2500 $\mu g/m^3$.