D607 - Assignment 1 - Urbanization Index

Marco Castro

2024-09-07

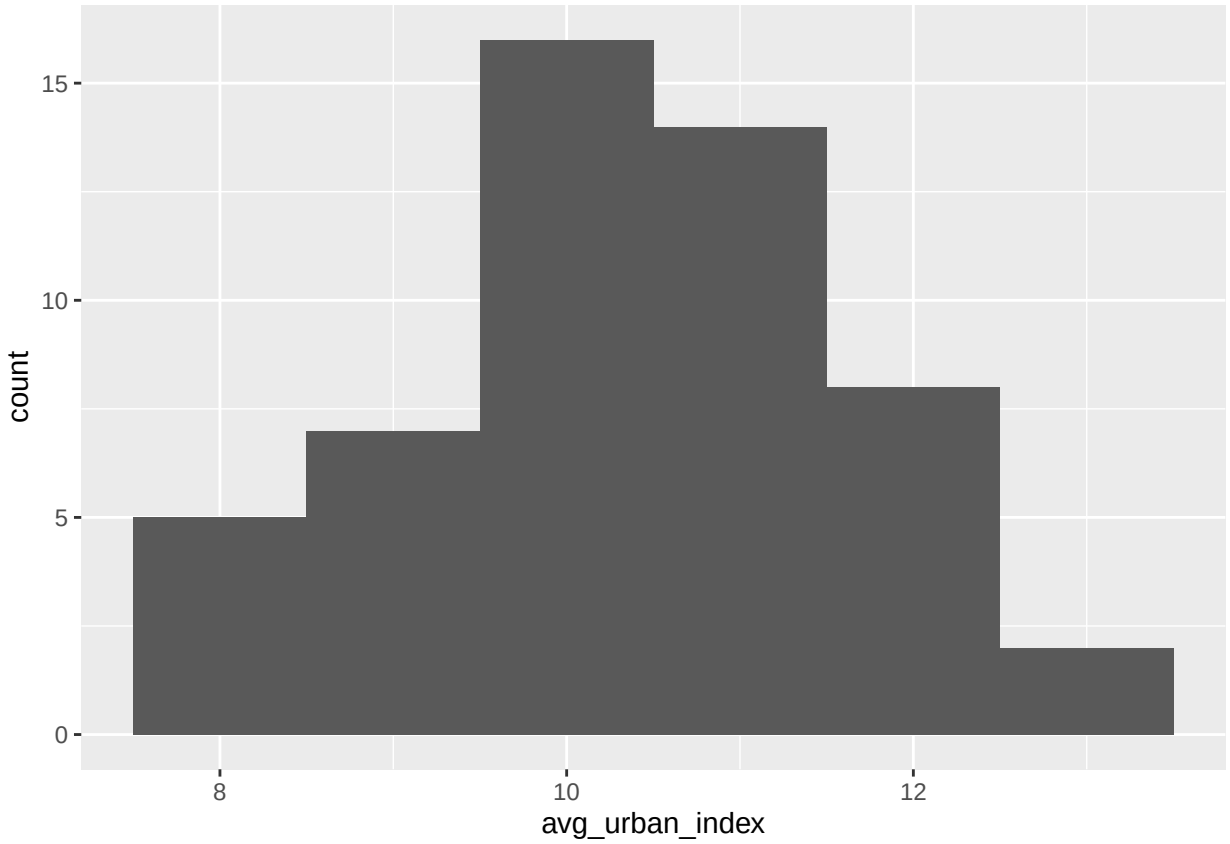
Overview

For the first assignment, I decided to explore FiveThirtyEight's Urbanization Index which was used as one of backbones for the article *How Urban Or Rural Is Your State? And What Does That Mean For The 2020 Election?* (https://fivethirtyeight.com/features/how-urban-or-rural-is-your-state-and-what-does-that-mean-for-the-2020-election/). FiveThirtyEight derived the index by applying the natural logarithm for the average population within a five-mile radius of every census tract. The authors then conduct further analysis using additional datapoints not included in this dataset and concluding that more urban areas vote democrat while more less urban areas tend to vote republican, in line with other prediction models.

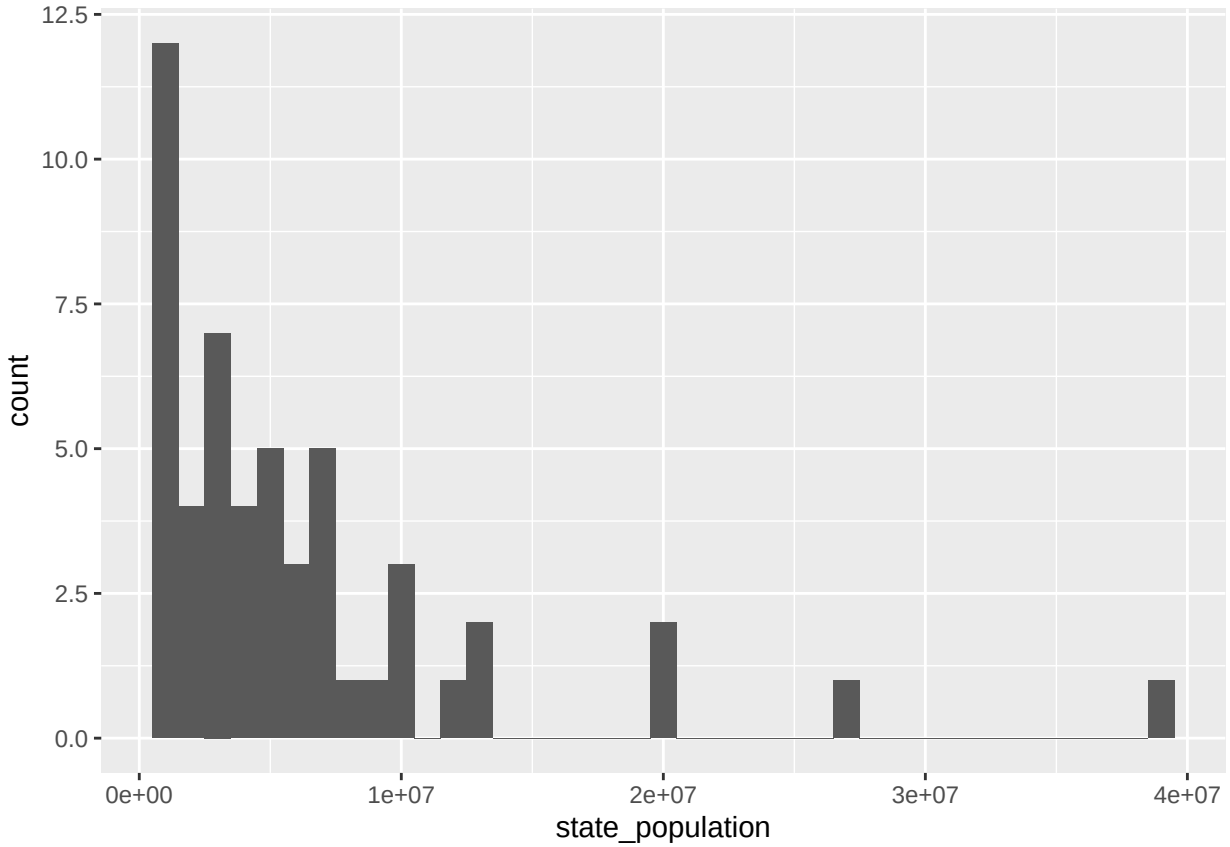
For this assignment, I first imported the data using the read_csv function. Since this dataset is broken down by census tracts, I created a subset of the dataframe stored in variable *urbanization_indexes_by_state* which calculates the sum of the population and average urban index for each state, as well its counts of census tracts. I plotted the distrubution of average urban index to get a sense of where most states fall and where the outliers fall.

```
# load data from 538 github page
urbanization_indexes <- read_csv("https://raw.githubusercontent.com/fivethirtyeight/data/master/urbaniza

# create new df of state populations and their average urban indexes
urbanization_indexes_by_state <- urbanization_indexes |>
    group_by(state) |>
    summarize(
      state_population = sum(population),
      avg_urban_index = mean(urbanindex),
      n = n()
    )

# set min/max population from all states
state_min_pop <- urbanization_indexes_by_state[which.min(urbanization_indexes_by_state$state_population
state_max_pop <- urbanization_indexes_by_state[which.max(urbanization_indexes_by_state$state_population

# set min/max avg indexes from all states
state_min_index <- urbanization_indexes_by_state[which.min(urbanization_indexes_by_state$avg_urban_inde
state_max_index <- urbanization_indexes_by_state[which.max(urbanization_indexes_by_state$avg_urban_inde

# what's the mean of the  average urbanization index
mean_urban_index <- mean(urbanization_indexes_by_state$avg_urban_index)

# show distribution of the average urbanization index
ggplot(
  urbanization_indexes_by_state,
```

```
  aes(x = avg_urban_index)
) +
geom_histogram(binwidth = 1)
```




```
# show distribution of state's population
ggplot(
  urbanization_indexes_by_state,
  aes(x = state_population)
) +
geom_histogram(binwidth = 1000000)
```



##Is a state's population an indicator of its urbanization index?

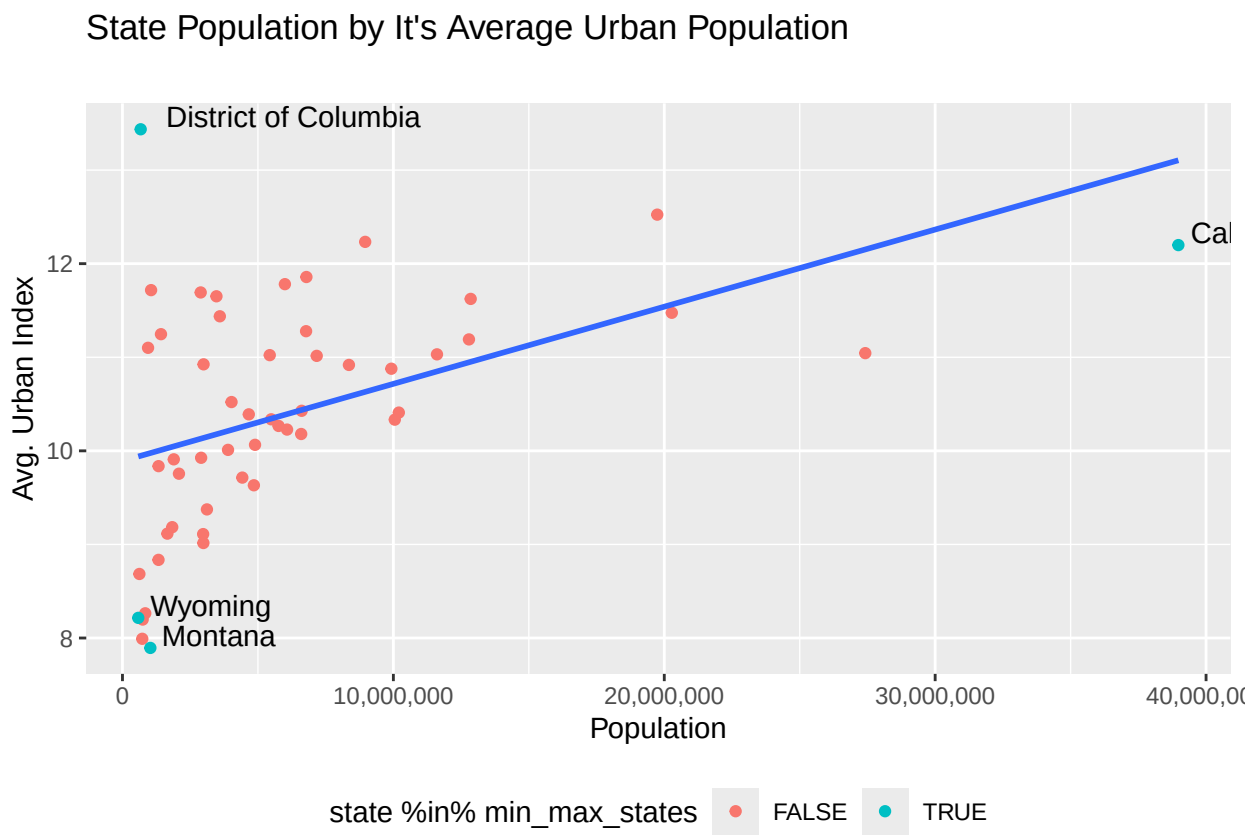
Results were plotted using ggplot with a trend line to see the rough relationship between population size and its urban index.

Using R Base's Min/Max function, I calculated the states with the highest and lowest populations. The state with a smallest population was Wyoming with an average urbanization index of 8.215266, while the state with largest avg. urbanization index was Montana with an average urbanization index of 7.8940607.

On the other end of the spectrum,the state with a largest population was California with an average urbanization index of 12.1985268, while the state with largest avg. urbanization index was District of Columbia with an average urbanization index of 13.4360377.

```
# set states with min/max pop and min/max avg indexes to list
min_max_states <- c(state_min_pop$state, state_max_pop$state, state_min_index$state, state_max_index$sta

# Plot the population vs. average indexes as a scatter plot
ggplot(
  data = urbanization_indexes_by_state,
  mapping = aes(x = state_population, y = avg_urban_index)
) +
  geom_point(aes(color = state %in% min_max_states)) +
  geom_smooth(method = "lm", se=F) +
  labs(
    title = "State Population by It's Average Urban Population",
    subtitle = "",
    x = "Population",
    y = "Avg. Urban Index"
```

```
) +
scale_x_continuous(labels = label_comma()) +
geom_text(aes(label=ifelse(state %in% min_max_states, as.character(state),'')),hjust=-.1,vjust=-.1) +
theme(
  legend.position = "bottom"
)
```



Conclusion

The scatter plot suggests a weak relationship between higher populations and higher urban indexes at the state level, as most data points are clusttered around the mean of the urbanization index (10.4056831) and have populations lower under 13 million from a range of 5.832×10^5-3.8977899×10^7. The distribution plots seem to confirm this clusttering effect. Furthermore, we can observe the presence of outliers. For example, the District of Columbia has the highest urbanization index but one of the lowest populations due to its relatively small size, number of census tracts presents, and other qualities like geographic composition that are not captured in the dataset. A more robust analysis, including implementing methods for handling outliers, are needed to arrive at a more statiscally sound conclusion.