Class Exercises

David Kane

2020-09-10

Today's class will be based on the first chapter of the textbook, which covers how to make basic and advanced plots using the **tidyverse** package. We will be using the "kenya" data set from the **PPBDS.data** package. To learn more about this package, read the description here. Try `?kenya`. Data is from "Electoral Administration in Fledgling Democracies: Experimental Evidence from Kenya" (pdf).

Scene 1

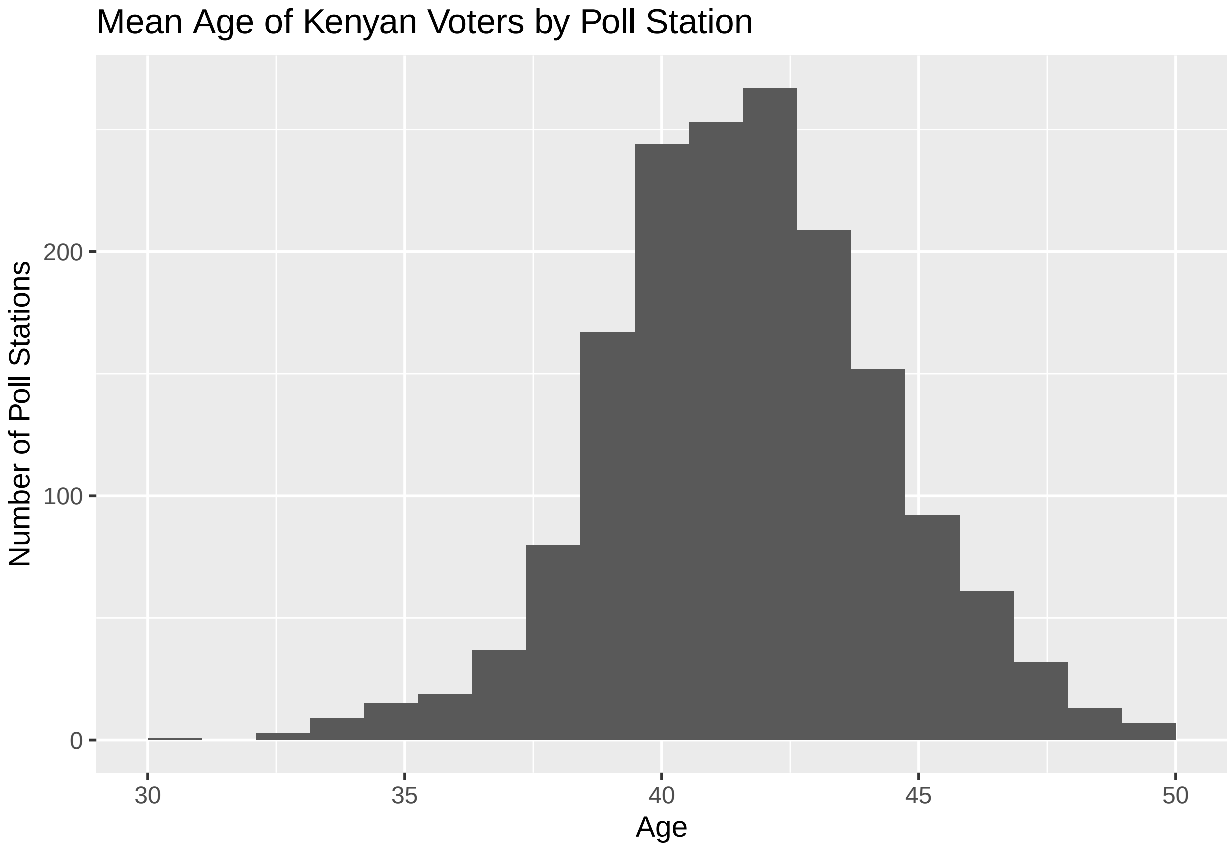
Prompt: Let's first explore the data by creating a basic histogram using `mean_age`. Use the `bins` argument to set an appropriate number of bins in your histogram, and set your range of values to ignore outliers. As always, label your plot appropriately with axis labels and a title. What does your plot tell you about the general age of Kenya's voting population? Along the way, you and your posse should do the usual data exploration: `glimpse()`, `summary()`, `sample_n()`, `view()` and so on. `xlim()` may be a useful function.

```r
library(ggplot2)
glimpse(kenya)
```

```
## Rows: 1,672
## Columns: 9
## $ block        <chr> "KWALE/42", "KWALE/35", "KWALE/40", "KWALE/18", "KWALE...
## $ poll_station <chr> "007/001", "007/004", "007/009", "007/011", "007/017",...
## $ treatment    <fct> control, local + canvass, local, local + SMS, local + ...
## $ poverty      <dbl> 0.2467021, 0.3293527, 0.2630458, 0.4294269, 0.3411467,...
## $ distance     <dbl> 22.022943, 25.142790, 27.822859, 27.220764, 19.301405,...
## $ pop_density  <dbl> 2.957305e-03, 8.879929e-04, 1.840065e-03, 2.695952e-04...
## $ mean_age     <dbl> 39.57318, 43.81937, 34.69372, 44.55333, 38.95662, 37.1...
## $ reg_byrv13   <dbl> 0.003584229, 0.074175824, 0.006908463, 0.260000000, 0....
## $ rv13         <dbl> 1116, 364, 4632, 150, 833, 1646, 616, 775, 696, 722, 2...
```

```r
ggplot(kenya,aes(mean_age)) +
  geom_histogram(bins=20) +
  xlim(30, 50) +
  labs(title="Mean Age of Kenyan Voters by Poll Station",
       x="Age",
       y="Number of Poll Stations")
```

```
## Warning: Removed 11 rows containing non-finite values (stat_bin).
```



Scene 2

Prompt: Next, let's explore the `reg_inc` variable in more detail by creating another histogram. Again, be sure to use an appropriate number of bins using the `bins` argument. You will need to re-scale the x-axis. Explore the difference between using a log scale versus a square root scale. Which one is better for this particular dataset, and why? `reg_inc` is the increase in voter registration in a community. Do you believe those outlier values? Why or why not?

Scene 3

Prompt: Let's look at the effectiveness of the "local + SMS" treatment relative to "control" using a box plot.

Scene 4

Prompt: We have seen that at least one treatment method was clearly effective in increasing voter registration. Now we want to compare the changes in voter registration across all treatments. Calculate the mean increase in registration for each treatment, and plot your results onto a bar graph.

Scene 5

Prompt: Now we know which treatment methods were most effective by looking at the mean increases. However, the previous graph doesn't tell us much about the distribution of `reg_inc` values across treatments.

Create a violin plot that shows the density of registration increases by treatment, and layer a box plot on top. Remember to use an appropriate scaling method for your x-axis.