Graph Challenge 2

Aryah Rao

2024-09-20

You can use data to tell many different, sometimes conflicting, stories. Below we have some code written by an analyst who claims there was fraud in the 2012 US Presidential Election between then Democratic incumbent Barack Obama and his Republican challenger Mitt Romney. After reading in the data, this analyst made two data visualizations that he believes show something is amiss about the election. The data records for all counties in the lower 48 US states vote totals and county population in the 2008 and 2012 US Presidential Elections. With this data, the analyst made a bar plot that summarizes the sum total of counties where Obama and Romney won the popular vote. He then made a scatter plot with a smoothed linear regression line of best fit showing the sum of ballots cast for Romney across counties by county population. If you run the code, you’ll see that Romney won a strong majority of counties, and that larger counties cast more ballots for Romney than smaller counties did. *How can it be that Obama won in 2012?!*

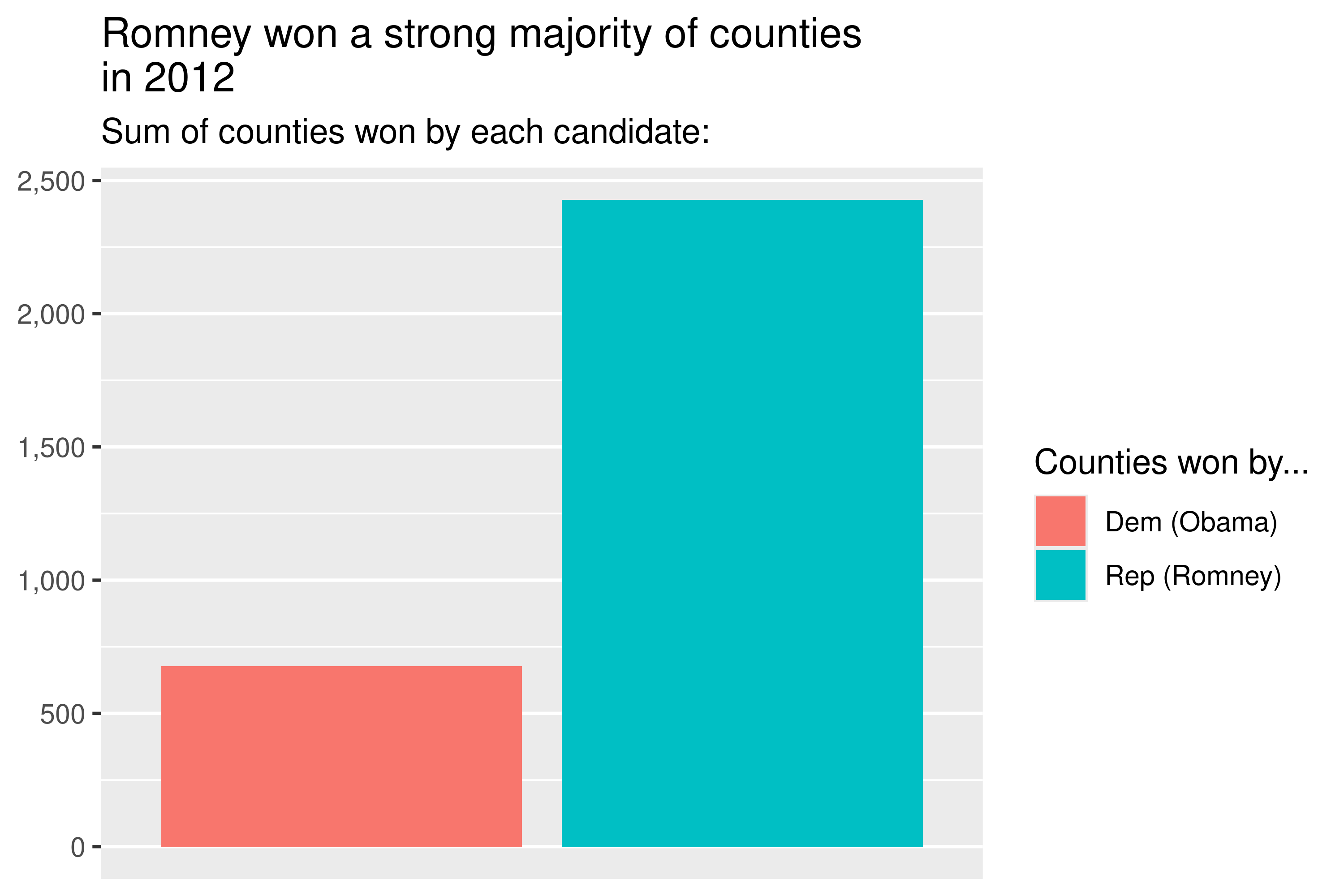
For this graph challenge, you mission is to update the below figures in ways that help our misguided analyst see that he’s missed some important details about the 2012 election. Find better ways to show the data so that these apparent systematic oddities look less odd. When you’re done, render your work and submit to Canvas.

First, open the {tidyverse} then read in the data:

## Open the packages you need  
library(tidyverse)  
  
## Attach the data   
## (you may need to modify the file path)  
read\_csv("/home/rao\_a1/dpr101-aryah/dpr-101-project-files-main/\_data/election.csv") -> dt

Here’s the first data visualization that the analyst created:

dt |>  
 drop\_na() |>  
 ggplot() +  
 aes(x = winner12, fill = winner12) +  
 geom\_bar() +  
 scale\_x\_discrete(  
 breaks = NULL  
 ) +  
 scale\_y\_continuous(  
 labels = scales::comma  
 ) +  
 labs(  
 title = paste0(  
 "Romney won a strong majority of counties\n",  
 "in 2012"  
 ),  
 subtitle = "Sum of counties won by each candidate:",  
 x = NULL,  
 y = NULL,  
 fill = "Counties won by..."  
 )

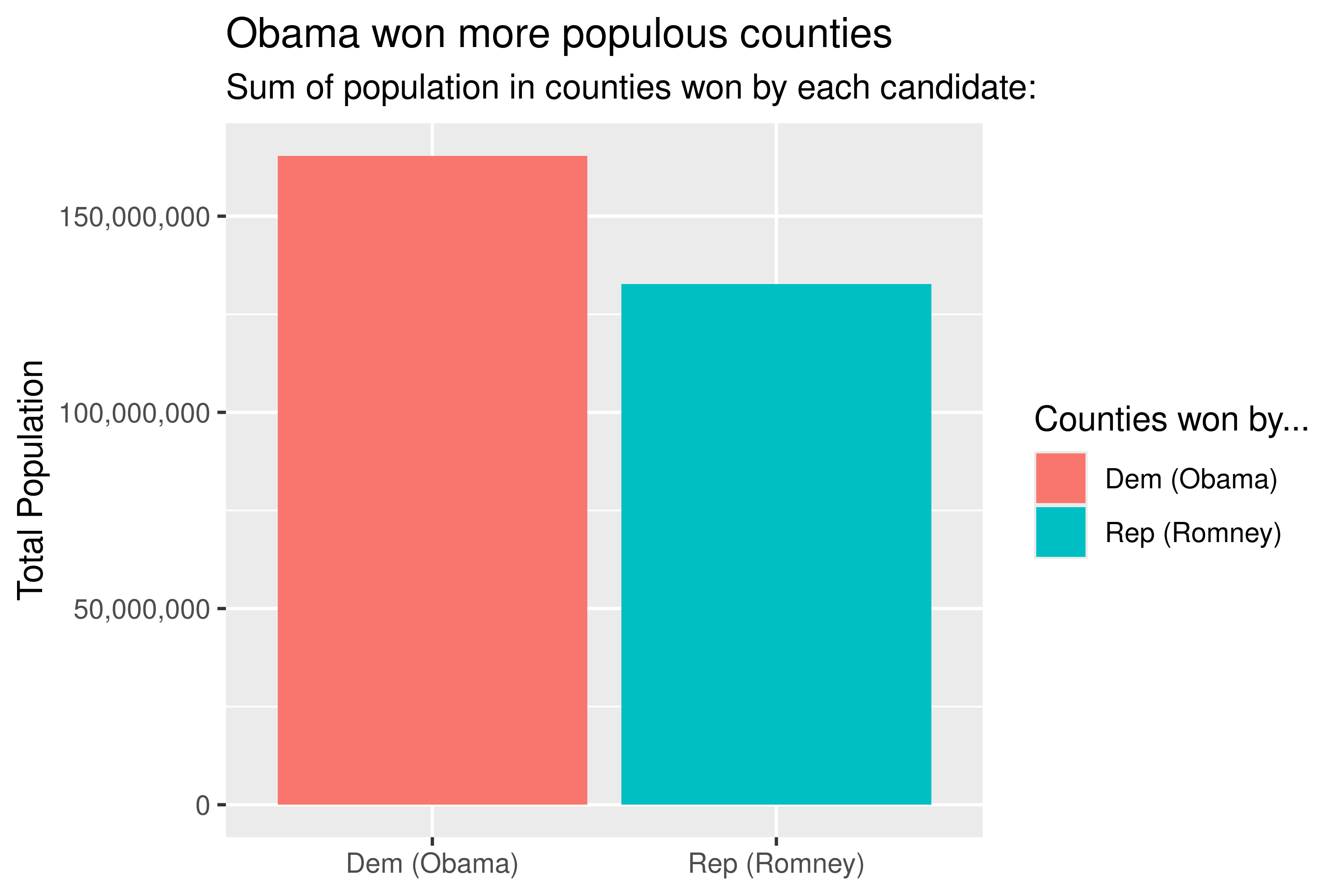


The key issue here is that the original bar chart only showed the number of counties won by each candidate, which made it look like Romney had the upper hand because he won more counties than Obama.

Winning more counties doesn’t automatically mean winning more total votes. If Obama won counties with larger populations, his total vote count would be higher, even though he won fewer counties.

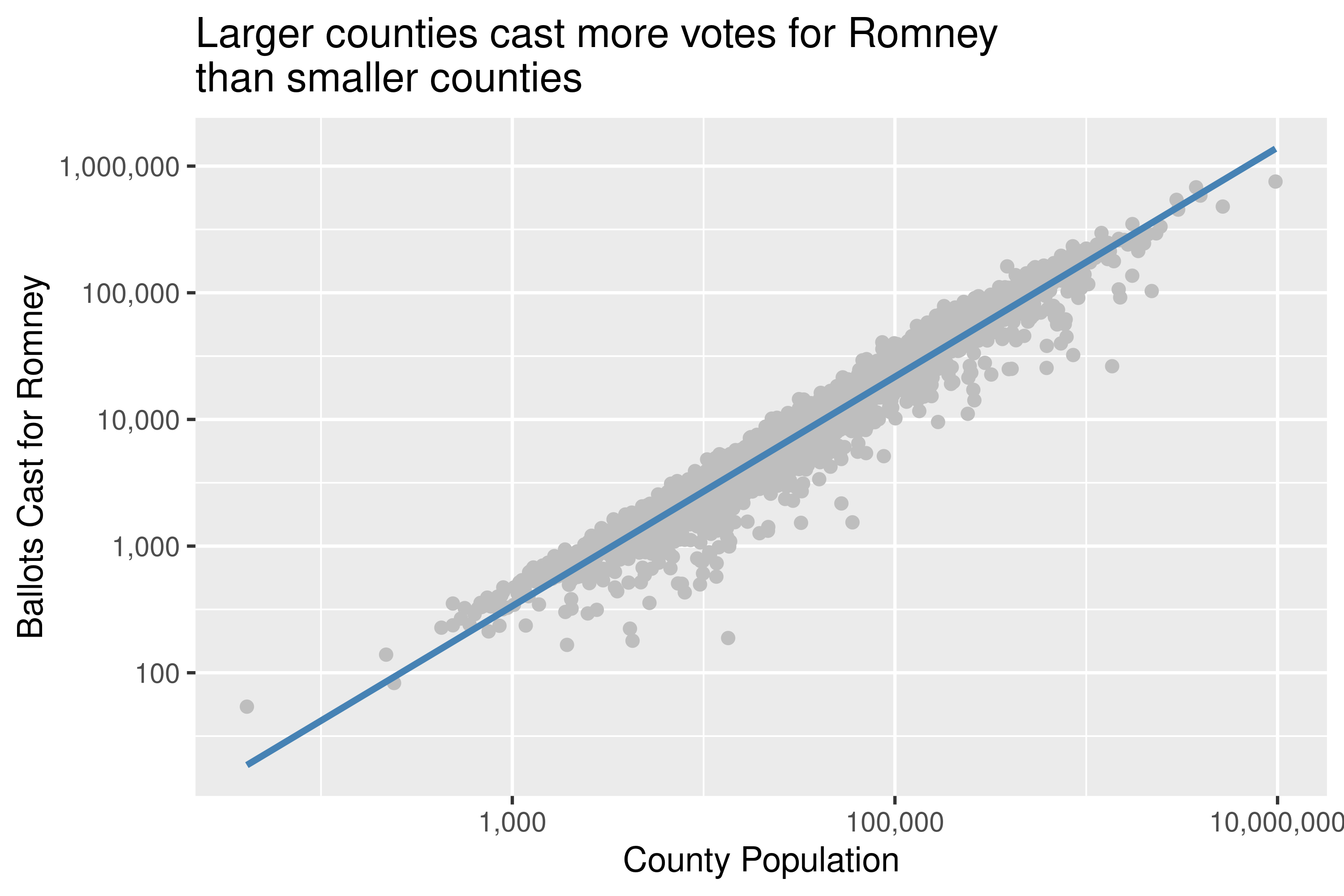
To address this misunderstanding, the improved bar plot not only shows the number of counties won but also weights the counties by their total population. This allows you to see that Obama’s counties, though fewer in number, had more people and therefore more total votes, explaining why Obama won the popular vote.

## Your updated plot should go here:  
dt |>  
 group\_by(winner12) |> # Group the data by the 'winner12' column  
 drop\_na() |>  
 summarise(total\_population = sum(pop)) |> # Calculate total population for each group  
 ggplot() + # Create a ggplot object  
 aes(x = winner12, y = total\_population, fill = winner12) +  
 geom\_bar(stat = "identity") + # Map 'winner12' & total population  
 scale\_y\_continuous(labels = scales::comma) + # Scale bar plot  
 # Add labels  
 labs(  
 title = "Obama won more populous counties",  
 subtitle = "Sum of population in counties won by each candidate:",  
 x = NULL,  
 y = "Total Population",  
 fill = "Counties won by..."  
 )



Here’s the second plot the analyst made:

ggplot(dt) +  
 aes(x = pop, y = rep2012) +  
 geom\_point(  
 color = "gray"  
 ) +  
 geom\_smooth(  
 method = "lm",  
 se = F,  
 color = "steelblue"  
 ) +  
 scale\_x\_log10(  
 labels = scales::comma  
 ) +  
 scale\_y\_log10(  
 labels = scales::comma  
 ) +  
 labs(  
 x = "County Population",  
 y = "Ballots Cast for Romney",  
 title = paste0(  
 "Larger counties cast more votes for Romney\n",  
 "than smaller counties"  
 )  
 )

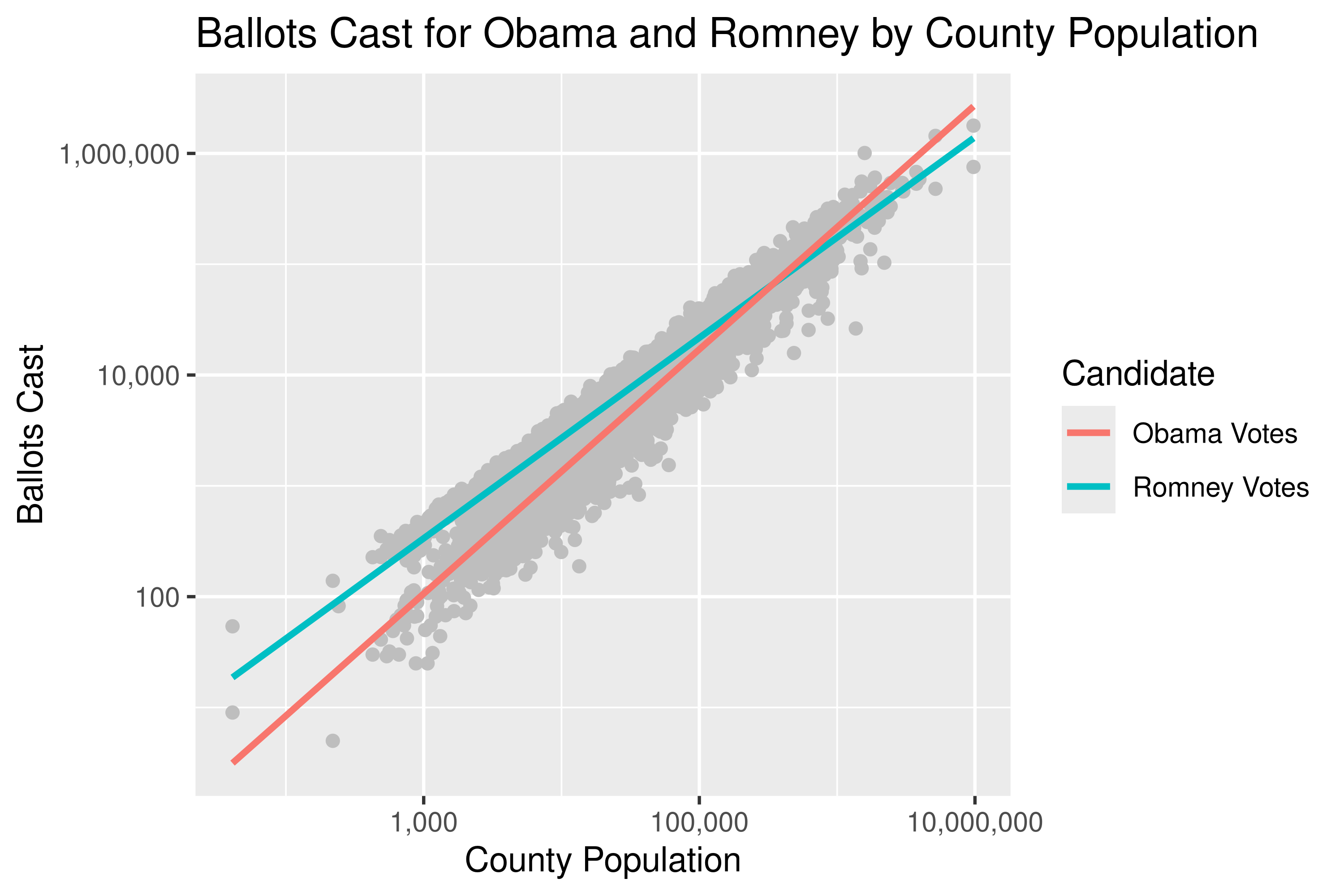


In the original scatter plot, the analyst only plotted Romney’s votes against county population. The regression line suggested that larger counties cast more ballots for Romney, but the plot ignored Obama’s votes, which made it one-sided and led to a potentially misleading interpretation.

In the updated version, I introduced both Obama’s and Romney’s vote totals to show a more balanced comparison.

Obama’s regression line has a steeper slope, this indicates that his vote totals increase significantly in larger counties. For Romney, the line shows that his vote totals increase more steadily, reflecting stronger performance in smaller or less densely populated areas.

## Your updated second plot should go here:  
ggplot(dt) + # Create a ggplot object  
 aes(x = pop) +  
 # Add points for Obama and Romney  
 geom\_point(aes(y = dem2012),   
 color = "gray") +  
 geom\_point(aes(y = rep2012),   
 color = "gray") +  
 # Add regression line for Obama and Romney  
 geom\_smooth(aes(y = rep2012, color = "Romney Votes"),   
 method = "lm",   
 se = F) +  
 geom\_smooth(aes(y = dem2012, color = "Obama Votes"),  
 method = "lm",   
 se = F) +  
 # Use log10 scale  
 scale\_x\_log10(  
 labels = scales::comma) +  
 scale\_y\_log10(  
 labels = scales::comma) +  
 # Add labels  
 labs(  
 x = "County Population",  
 y = "Ballots Cast",  
 title = "Ballots Cast for Obama and Romney by County Population",  
 color = "Candidate"  
 )



Why Obama won despite Romney’s lead in counties: Romney won a majority of smaller, less populated rural counties, but these contribute fewer votes. Obama won fewer but much larger, more densely populated counties. These counties contribute many more votes, even if there are fewer of them.