PS6$_S chmidt$

akeaneschmidt1998

March 2024

1 Data Collection

I chose to move forward with the non-API data I collected in the previous project, as it was already relatively clean and interested me personally. Although I found the API-collected data fun, it wouldn't need cleaning and I didn't find it interesting, which I felt would deny me the essential purposes of the project.

Data was collected from Oklahomawatch using the rvest package. First, I directed the URL of the web site to the page variable by the "read html" function. Next, I used the CSS selector bookmark to extract the CSS titles needed to scrape the data, and fed it into the code, naming each variable as it appeared on the website. I then printed the names of the vectors (optional) because I wanted to debug an issue I kept having in this section. I wrote the lengths of the vectors using the length(variable) command, and used an If function to check that the lengths were all the same. If they were not (which was the case initially), then the function returned a specific error for this issue telling me to check the CSS selectors. If they were, the data was converted into a .CSV data frame. Empty cells were filled with "NA", and moved on to cleaning.

2 Data Cleaning

The data was cleaned by removing towns listwise. I decided on using listwise deletion because I didn't think there was any reason to believe the missing data was MCAR. Rather, it seemed plausible to me that local ordinances might have caused data to be reported or aggregated differently.

I was also careful to keep data which featured a "0," as many districts have 0 population (for instance, university precincts with no permanent residents), 0 crime, etc. To do this, I converted any cell which was filled with nothing into reading "NA," and then used an na.omit function to exclude the entire line when writing the data frame (lines 38-42).

The character "Â" kept appearing in the data, seemingly a translation from when the publisher tried to use a colored dot to color-code data. Since the color coding was only to visually identify crime per residents into brackets, it was not

needed and was discarded using the lapply function and the Unicode for that artifact.

3 Data Visualization

Three charts were created, relating the rate of crime overall, the rate of violent crime, and the rate of property crime to the total population of the district. The crime rate was used instead of the crime, so as to control for the obvious effect that more population should result in more opportunities for crime.

Some modifications were made to make the graphs clearer. First, the scale of the X-axis was altered to intervals of 10,000, as the graph was too hard to read with the population of each specific town displayed. Next, blue bars connected the possible populations of each crime rate, which were removed.

An additional set of graphs was created which eliminated outliers based on the interquartile range.

I expected the data to show an upward trend between population and crime rates, as I assumed a denser city would be harder to police. Instead, the graph shows minimal evidence that this is the case. Instead, population appears to have a moderating effect on all three forms of crime, with (1-3) and without (4-6) outliers included.

4 Appendix

Figure 1: Offense Rate vs. Population

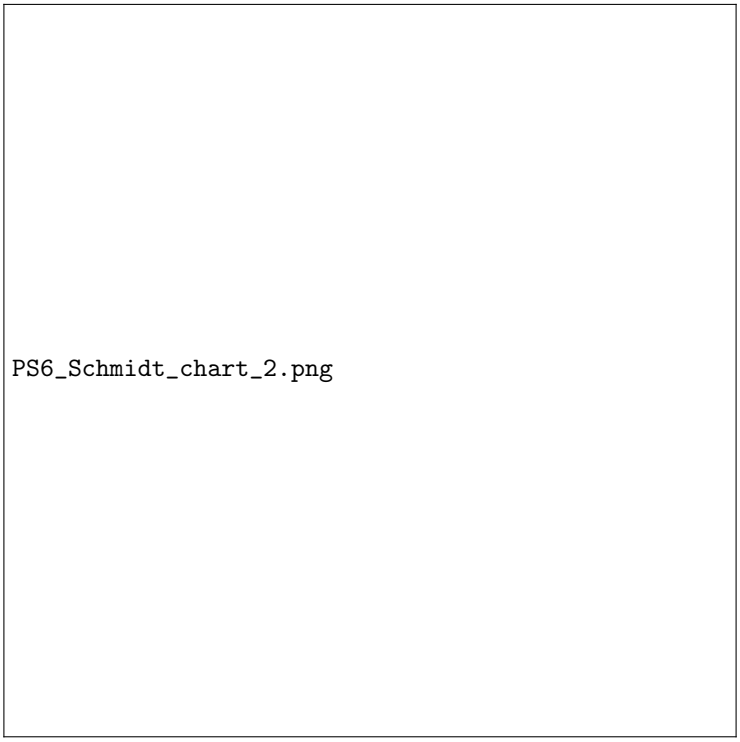

Figure 2: Violent Crime Rate vs. Population

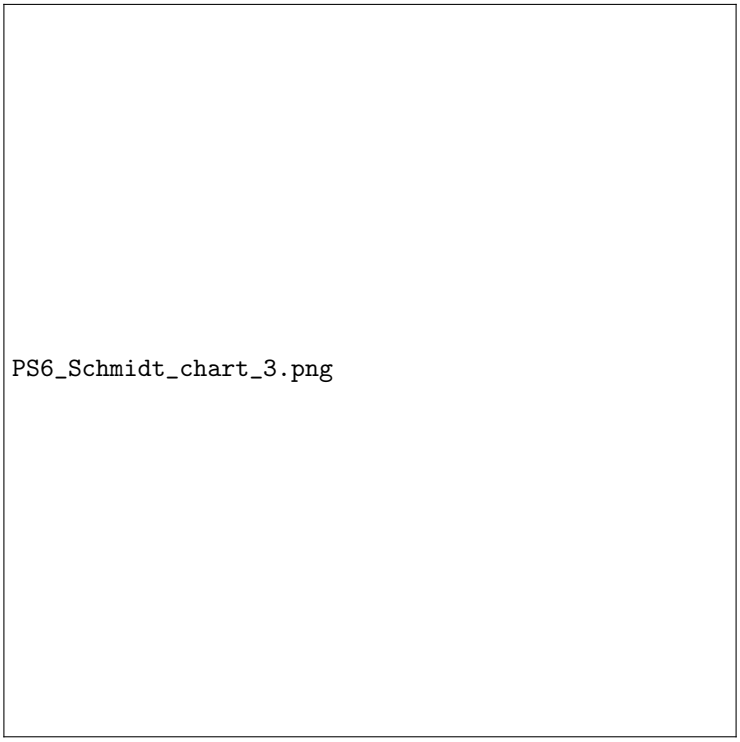

Figure 3: Property Crime Rate vs. Population

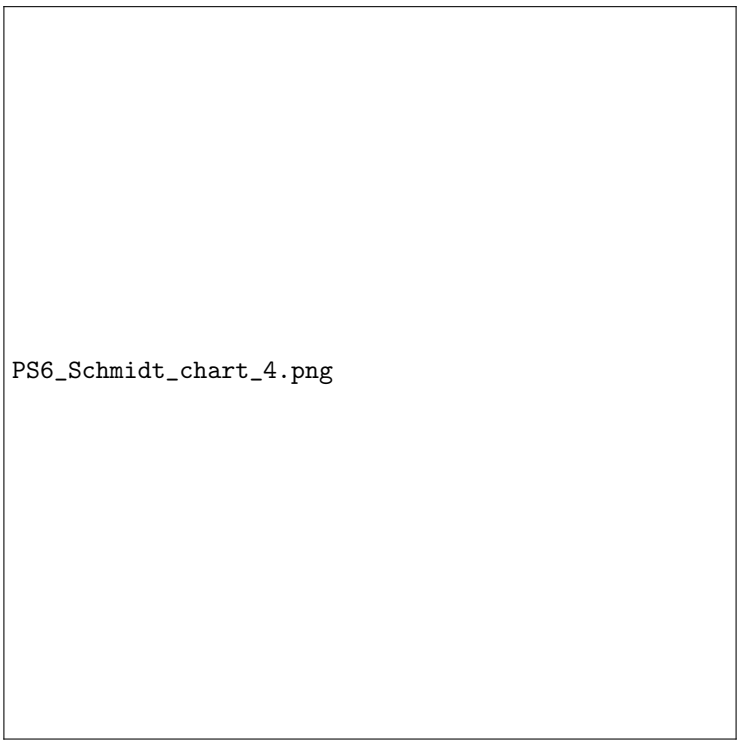

Figure 4: Offense Rate vs. Population (Excluding Outliers)

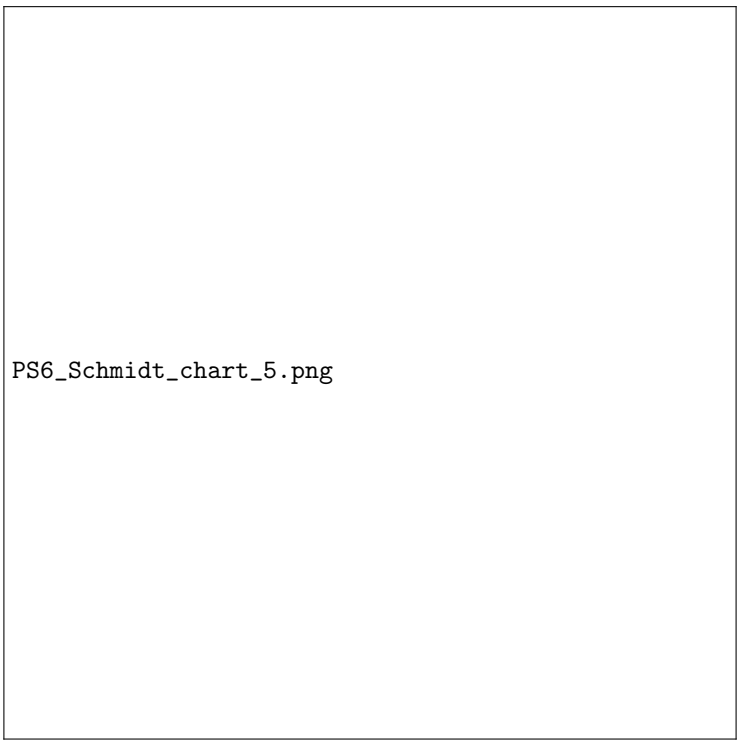

Figure 5: Violent Crime Rate vs. Population (Excluding Outliers)

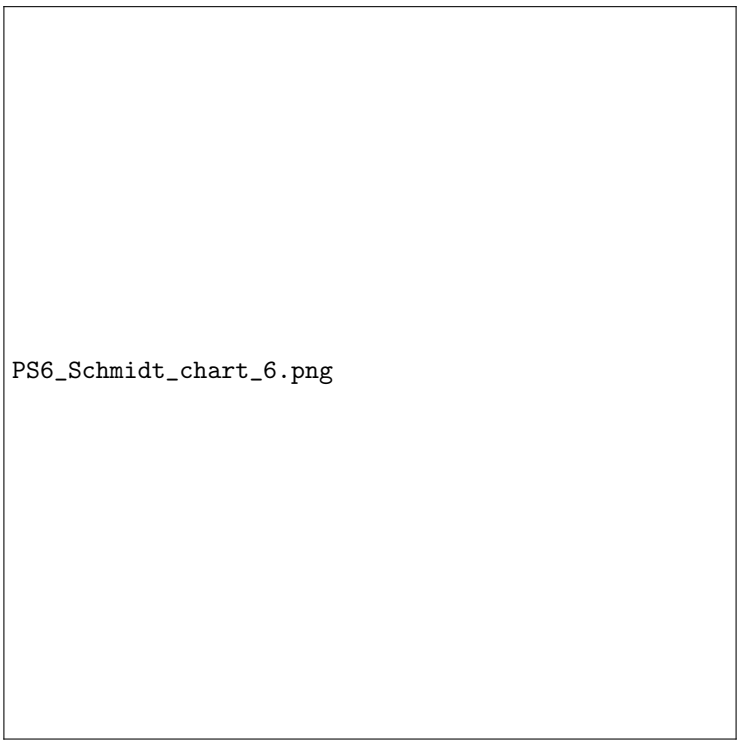

Figure 6: Property Crime Rate vs. Population (Excluding Outliers)