Lab Report 3: Manipulating Data

State Crime Rates

DUE: October 28 at 5 PM

You work for a nonprofit organization interested in reducing violent crime in the United States. Your boss is interested in the answers to two questions:

1. Whether at the state level rates of violent crime (the occurrence of violent crimes—murder, rape, robbery, and aggravated assault—per 100,000 people in a state as collected by the Bureau of Justice Statistics Uniform Crime Reports) increased, decreased, or stayed the same between 2000 and 2010.
2. Whether states with more Republicans, who may favor more "tough on crime policies," have higher, lower, or the same levels of violent crime than states with fewer Republicans.

For this assignment, work through this sheet and answer the questions along the way before writing up your brief memo (described below). You can write your memo at the end of this sheet so that all previous answers are included as well.

Your boss gives you two datasets to work with: one with information on violent crime in the US ('crime_long.csv') and other state-level data from the Correlates of State Policy dataset ('correlates_2000-2010.csv'). These datasets are available on CANVAS or you can import them from online using the following code:

```
Crime <- read.csv("https://raw.githubusercontent.com/ilaydaonder/POLS209/
Lab-Report-3/crime_long.csv")
```

```
Correlates <- read.csv("https://raw.githubusercontent.com/ilaydaonder/POLS209/
Lab-Report-3/correlates_2000-2010.csv")
```

To conduct this exercise, we will need to learn how to merge data, review subsetting data, and create new variables based on values of existing variables.

Notice that we have two datasets, one named "crime" and one named "correlates." Using the 'dim()' function, determine how many rows and columns are in each spreadsheet.

- Rows in correlates data:
- Columns in correlates data:
- Rows in crime data:
- Columns in crime data:

- There are 50 states in each dataset plus the District of Columbia. Why are there more than 51 rows/observations?
- What is the unit of analysis in these data (e.g. what defines each unique row?)

Answering your boss's first question will only require that you use the crime dataset because the crime rate statistics for both 2000 and 2010 are available in that dataset. However, answering your boss's second question will require you to merge the crime dataset and the correlates dataset because the variables about Republicans/Democrats at the state level are in the correlates dataset.

The code to merge datasets in R is not difficult, but you MUST BE CAREFUL that R keeps all the observations you want. Merging datasets requires that each dataset has some indicator (or set of indicators) that uniquely identify the observations. In other words, because you want the data for Alaska from both datasets to align when merging, you need to tell R the name of the variable that uniquely identifies that unit. In these datasets, note that both have a variable called "stateno." This variable gives a number for each state that distinguishes it from all others. For example, Alabama is 1, Alaska is 2, and so on. Since both datasets have this same variable and the name of that variable is the same in both, we can tell R to merge the datasets on that variable to properly align the datasets.

The code to merge data takes the following general structure:

```
Merged1 <- merge(x = dataset1, y = dataset2, by = "ID", all = TRUE)
```

Where 'merged1' can be any name for your new merged object, 'x' is the first dataset, 'y' is the second dataset, and 'by' specifies the name of the variable you want to merge on. As you may have guessed, however, the state number itself does not uniquely identify each row because there are two observations for each state (e.g., Alabama 2000, Alabama 2010). Remember the unit of analysis you specified above?

If observations are uniquely identified by a combination of two indicators (e.g., 'stateno', 'year'), we can give the 'by' argument a vector instead of one variable name: 'by = c("stateno", "year")'.

Try merging the datasets together using the above code but replace "dataset1," "dataset2," and "ID" with the object and variable names you've specified in your code. Be sure to include 'all = TRUE'.

- What is the unit of analysis in these data (e.g. what defines each unique row?)
- Now create a new merged dataset (call it 'merged2') where you replace 'all = TRUE' with 'all = FALSE.' What changed?

You now have a merged dataset! Use the first merged dataset for the rest of the assignment.

Your boss is interested in the difference between crime rates in 2000 and 2010. Using your merged dataset, create two separate datasets where one represents the data from 2000 and

2010. Call them "data2000" and "data2010." To do this, you'll need to subset the data. We have done this before, but as a reminder, the code (in general) looks like this:

```
Data2000 <- subset(dataset, variable == value)
```

Create your two subsetted datasets based on whether the year is 2000 or 2010. Then calculate the following (using 'summary()' may be the most efficient):

- Mean of crime_rate in 2000:
- Mean of crime_rate in 2010:
- Range (max – min) of crime_rate in 2000:
- Range (max – min) of crime_rate in 2010:
- Standard deviation of crime_rate in 2000:
- Standard deviation of crime_rate in 2010:
- Mean of Republicans in state houses in 2000 (hs_rep_in_sess):
- Mean of Republicans in state houses in 2010 (hs_rep_in_sess):
- Mean of Republicans in state senates in 2000 (sen_rep_in_sess):
- Mean of Republicans in state senates in 2010 (sen_rep_in_sess):

The last thing we will want to do is create a new variable representing whether Republicans are in the majority in state legislatures. We have data on the number of Republicans and Democrats in each state legislative chamber, but we may be interested in creating a new binary variable that takes a value of 1 if the Republicans are in the majority and 0 if not.

There are many ways to create new variables from existing data. Let's use the 'ifelse()' function. This function works by defining a new variable (or object) based on the values of some other variable that you specify. It is generally structured as follows:

```
Dataset$new_variable_name <- ifelse(Dataset$variable == X, value1, value_for_
all_other_observations)
```

We can read the above code as the following: "create a new variable called 'new_variable_name' in my Dataset object. If the value of the old variable is X, then give my new variable a value of 'value1' and 'value_for_all_other_observations' for everything else." Note that we don't need to use the '==' operator; we can use greater than ('>') or less than ('<') too.

Since a legislative chamber that has more Republicans than Democrats means that that chamber has a majority (technically a plurality) of seats, let's create a new variable in our merged dataset defined as a 1 if the number of Republican seats is greater than the number of Democrat seats and 0 otherwise. Your object names may differ from below, so alter this code accordingly:

```
Merged1$house_r_majority <- ifelse(Merged1$hs_rep_in_sess > Merged1$hs_dem_in_
sess, 1, 0)
```

You should have a new variable called 'house_r_majority' defined as 1 if there are more Republicans than Democrats in the house of representatives and 0 otherwise.

Now do the same but for the senate chambers. Call this variable 'senate_r_majority.' Use 'sen_rep_in_sess' as the data for senate Republican seats and 'sen_dem_in_sess' for senate Democratic seats.

You should now have two new variables! You can view the basic summary statistics using 'summary()' but remember these are now binary (nominal) variables.

Now that you have manipulated the datasets to thoroughly answer your boss's two questions, write a short memo (about 1 page including figures) below that includes answers to the following, some of which you have calculated above:

- What is the average crime rate in 2000 and 2010? Did this number increase or decrease? Is this difference statistically significant? (HINT: use 't.test()' to conduct a difference of means test and report the p-value and what that means)
- Using your two subsetted datasets, create a scatterplot with the number of Republicans in the house of representatives ('hs_rep_in_sess') on the x-axis and crime_rate on the y-axis. (note: you should have two scatterplots, one for 2000 and one for 2010). Does there seem to be a relationship? If so, is that relationship positive or negative?
- Are states with Republican majorities in their house of representatives associated with more/less crime? Report the mean crime rates in majority Republican houses and majority non-Republican houses and conduct a difference of means t-test.