Comparison of Campus Crime with Crime in the Surrounding City

Group Number: 41

Group Members: Yizhi Fang, Matt Welch, Jiamin Zhong, Jonathan Pickett

Introduction: For our project we wanted to look at how crime in the surrounding city relates to higher levels of university student drug, weapon, and alcohol violations. We also explored how crime rates, adjusted on a population level, relate to the median income levels of a state. To do this analysis we merged the crime data sets by State and City. We printed the least safe universities and summed the data on a state level to merge with the median income data by State. The university data for on campus violations is located at https://ope.ed.gov/campussafety/#/datafile/list From here one can select the year 2012 and download the zip file. This gives data for the three previous years. The downloaded folder contains 23 XLS files for which we choose the on-campus arrest data set. It contains the past three years 2009-2011, and we chose the year 2010 since this is the year that matches the FBI data described below. The main variables in the on-campus data set include the University's Institution Name, City, State, total university enrollment, and three separate variables for Drug, Weapon, and Liquor violations. The raw data has 10769 observations with 24 variables so is our "large" data set.

Our second FBI data set can be found at https://ucr.fbi.gov/crime-in-the-u.s/2010/crime-in-the-u.s.-2010/tables/10tbl08.xls/view This data set has crime statistics in the surrounding city, and close to 10,000 cities in all. Among the important variables are State and City like before as well as Violent Crime and Property Crime (which we name non-violent crime below). Violent crime can be broken down into Manslaughter, Rape, Robbery, and Aggravated assault. Property crime is the sum of the variables burglary, larceny-theft, and motor vehicle theft. We exclude the the Arson variable due to too many missing values. Finally, our third data set can be found at a census.gov website: https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_05_EST_GCT2001.US13PR&prodType=table It contains two columns: State Name and Median Income. All 50 states and Washington D.C. are included in this data.

Methods: As mentioned above, we are looking to eventually merge three data sets together. However, we do not use all variables in the analysis. The most important variables, and the variables included in the final data set, are State (a character variable of length 25), City (a character variable of length 50), and numeric variables. These numeric variables include Total (which represents total university enrollment); Total_Crimes which is a variable we create as the sum of violent crime and property crime/ non-violent crime in the surrounding city; Total_Campus_Crime which represent total weapon, liquor, and drug violations; and Median_Income which as described above is given on a state level. We first read in the FBI crime dataset with the appropriate firstobs =5, dsd option, formats, and lengths. The max length of the character variables were found in Excel and observations 9311 to 9317 were really just footnotes at the end of the file so we excluded them. The FBI crime file had a state name listed for the first city in each state and then missing values until the next state name. To correct this, we used the retain function to fill in missing state names. Next, we used the compress(variable, , 'd') option in SAS to remove numbers from city and state names. These numbers were footnotes that became attached to the end of states and cities when the data was read in.

Example of a city name with a number
This number must be removed

Obs	State	City
8	ALABAMA	Andalusia2

Number of missing, min, and max for numeric variables in FBI crime data set

The MEANS Procedure

Variable	Label	N	N Miss	Minimum	Maximum
Population		9306	4	17.0000000	8336002.00
Violent_crime	Violent Crime	9081	229	0	48489.00
Rape		9105	205	0	1036.00
Prop_crime	Property crime	9280	30	0	139615.00
Arson1		8932	378	0	1341.00

Then, for validation of the FBI data set we looked at the minimum, mean, and maximums to make sure the crime variables were in reasonable ranges. Since cities as large as New York City were in our data, property crimes as high as 100,000 do make sense. Likewise, values of zero make sense for small towns. In particular, we looked at missing values of populations and extremely small populations. Using proc univariate, we found that Teterboro, New Jersey had the smallest population with 17 and Lakeside had 19 people respectively. Searching these towns, along with other small populations in the 2010 census data, confirmed these small values.

Two smallest towns (populations confirmed to be 17 and 19)

The UNIVARIATE Procedure
Variable: Population

Extreme Observations					
Lowest			Highest		
Value	City	Obs	Value	City	Obs
17	Teterboro	5050	3841707	Los Angeles	708
19	Lakeside	1005	8336002	New York	5392

Next, we examined missing values. Property crime variables had around 30 missing values; quite small considering the data set had 10,000 observations. On the other hand, the violent crime variable had 229 missing values, which was mostly driven by 205 missing values of rape. Therefore, we created a new variable violent_crime1 to represent the sum of the violent crime variables robbery, assault, and murder (excluding rape). This variable could be used to confirm that the missing values do not impact the analysis.

We then used proc SQL to confirm that there were 51 distinct states (including Washington DC). We also deleted cities with extremely large violent crimes/ property crimes (more than 25 percent of population). For example, Vernon, CA had 89 residents and 319 property crimes. We removed 15 observations where the crime likely came from (or was included with) the surrounding county.

Number of Distinct States
51

Examples of cities with Violent Crime or Property Crime more than 25 percent of population

Obs	State	City	Population	Violent Crime	Property crime
662	CALIFORNIA	Industry	930	77	1,183
826	CALIFORNIA	Sand City	372	3	125
905	CALIFORNIA	Vernon	89	26	316

Number of missing values for on-campus data set

The MEANS Procedure

Variable	Label	N Miss
Total	Total enrollment	31
WEAPON10	Weapons Violations: carrying, possessing,	793
DRUG10	etc. in 2010	793
LIQUOR10	Drug Violations in 2010	793
	Liquor Violations in 2010	

The second phase of our analysis involved cleaning the on-campus data. We read in the data, added formats/labels, proper case, and then examined the data. There were 31 missing values of total, 793 missing values of drug, liquor, or weapon violations (in 2010), and 149 missing values of state. We excluded these observations from the analysis. This was in all around 8% of the data. Note that the missing values of state also came from campuses located in foreign countries so should be excluded as well. Then, using proc sql we found 59 distinct regions. We used the SAS stname( ) function to convert all the non-missing 2-digit abbreviated state names to their full written names and excluded universities in 8 territories not in the US. (For example, Guam and Puerto Rico). Using the substr( ) function we removed 4 digit extensions from zip codes. Finally, we examined small campuses and found some enrollments of less than 50. These were not a concern since the small enrollments could be possible for hairdressing school, Beauty College, etc.

For the final analysis we then read in an income data set, including the upcase function, formats, and so on. This gave us median incomes for all 50 states and Washington, DC. The three data sets were then merged together appropriately, as explained next.

Example of Foreign University

Obs	UNITID_P	INSTNM	City	State	ZIP	Total	WEAPON10	DRUG10	LIQUOR10	FILTER10
305	107044004	Harding University	Markopoulo, Attiki		19003	7056	0	0	0	1

Number of Distinct States/Territories
59

total number of schools
455

Results: With the income, FBI city, and on-campus data sets cleaned. We next used arrays to adjust the crime variables by population and then multiplied this by 100,000. For the on-campus data set on the other hand it was more difficult since some universities had multiple campuses located in different towns, but only one university enrollment was given. For our analysis, we made the assumption that the amount in each separate campus was proportional to the size of the city it was in. While this condition may not hold, the focus of our project was not a rigorous statistical analysis so we continued with this assumption. We then merged the FBI crime data set with the on-campus data set by state and city (including only common observations). This gave university data along with crime statistics from the surrounding city.

Top 4 unsafe campuses with over 10,000 in enrollment
Rates displayed per 100,000

State	City	Population	Institution Name	Total enrollment	Violent crimes in surrounding city	Non-violent crimes in surrounding city	Total campus violations in 2010
WISCONSIN	La Crosse	51,184	University of Wisconsin-La Crosse	10,284	302.83	3,366.29	4,959.16
WISCONSIN	Oshkosh	64,428	University of Wisconsin-Oshkosh	13,513	319.74	2,661.89	4,262.56
TEXAS	San Marcos	55,100	Texas State University-San Marcos	34,087	275.86	3,132.49	3,459.91
COLORADO	Boulder	99,255	University of Colorado Boulder	32,558	211.58	2,742.43	3,258.80
.....	.....	.....	.....	.....	.....	.....	.....

The next part of the analysis we created the variables Total Crime, the sum of property and violent crime, and Total Campus Violations, the sum of alcohol, drug, and weapon violations, and printed the 4 least safe cities as shown (adjusted for population as well). Next, we summed the on campus and FBI crime data sets by State, then adjusted for population, and finally merged them together in the same way as before along with median income data for each state. Finally, we formatted the Total campus, Total crime (for the surrounding city), and median income into low, medium, and high categories. We provided cross tabulation tables where one can see states with high on-campus crime also tended to have high crime in the surrounding city, or states with high median incomes tended to have low crime both in surrounding city and on-campus (shown in Appendix A). Appendix B has the full list of Total Crime, Total campus, and median income in every state.

Classifying States by Total Campus Violations and Total Crime in the State (Low, Medium, High)

The FREQ Procedure

Table of Total_Campus_Crime by Total_crime

Total_Campus_Crime / Frequency Percent	Total_crime: Low	Medium	High	Total
Low	5 9.80	5 9.80	2 3.92	12 23.53
Medium	6 11.76	14 27.45	6 11.76	26 50.98
High	1 1.96	7 13.73	5 9.80	13 25.49
Total	12 23.53	26 50.98	13 25.49	51 100.00

APPENDIX A

Classifying States by Total Campus Violations and Median Income for the State (Low, Medium, High)

The FREQ Procedure

Table of Total_Campus_Crime by Median_Income

Total_Campus_Crime Frequency Percent	Median_Income Low	Medium	High	Total
Low	4 7.84	5 9.80	3 5.88	12 23.53
Medium	6 11.76	14 27.45	6 11.76	26 50.98
High	2 3.92	7 13.73	4 7.84	13 25.49
Total	12 23.53	26 50.98	13 25.49	51 100.00

APPENDIX B

Total crime, Total Campus Violations and Median Income (Low, Medium, High) for every state

State	Total_crime	Total_Campus_Crime	Median_Income
ALABAMA	High	Low	Low
ALASKA	Medium	High	High
ARIZONA	Low	Medium	Medium
ARKANSAS	High	Low	Low
CALIFORNIA	Low	Low	High
COLORADO	Low	High	High
CONNECTICUT	Low	Medium	High
DELAWARE	High	High	High
DISTRICT OF COLUMBIA	Low	Medium	Medium
FLORIDA	Medium	Low	Medium
GEORGIA	High	Medium	Medium
HAWAII	Low	Low	High
IDAHO	Medium	Medium	Medium
ILLINOIS	Medium	Low	Medium
INDIANA	Medium	Medium	Medium
IOWA	Medium	Medium	Medium
KANSAS	Medium	Medium	Medium
KENTUCKY	Medium	Low	Low
LOUISIANA	High	Medium	Low
MAINE	High	Medium	Medium
MARYLAND	Medium	Medium	High
MASSACHUSETTS	Low	Medium	High
MICHIGAN	Medium	High	Medium
MINNESOTA	Low	Medium	High
MISSISSIPPI	Medium	Medium	Low
MISSOURI	High	Medium	Medium
MONTANA	High	Medium	Low
NEBRASKA	Low	Medium	Medium

NEVADA	Low	Low	Medium
NEW HAMPSHIRE	High	High	High
NEW JERSEY	Medium	Medium	High
NEW MEXICO	Medium	Medium	Low
NEW YORK	Low	Low	Medium
NORTH CAROLINA	Medium	Medium	Medium
NORTH DAKOTA	Medium	High	Medium
OHIO	Medium	Medium	Medium
OKLAHOMA	Medium	Low	Low
OREGON	Medium	High	Medium
PENNSYLVANIA	Medium	High	Medium
RHODE ISLAND	Low	Low	High
SOUTH CAROLINA	High	Medium	Low
SOUTH DAKOTA	Medium	High	Low
TENNESSEE	Medium	Medium	Low
TEXAS	Medium	Medium	Medium
UTAH	Medium	Low	Medium
VERMONT	High	High	Medium
VIRGINIA	Medium	Medium	High
WASHINGTON	Medium	Medium	Medium
WEST VIRGINIA	High	High	Low
WISCONSIN	Medium	High	Medium
WYOMING	High	High	Medium