Michael Reyes
CIS 320
T 6:10pm – 10:00pm
2 March 2016

CIS 320 Report

The data I decided to use for this project was the Chicago crime reports. The data has been recorded since 2001, and stays up to date to the past seven days. I downloaded the .csv file and saved it as "Crimes_-_2001_to_present", I then uploaded that information into Microsoft Excel. I discovered in Excel that there was over 290,000 entries spanning from late 2014 to early 2016. I created a pivot table, finding the "Count of Arrest", with "Primary Type" and "Description" as rows, and "Year" as columns. I also put a "Block", "Community Area", and "District" as filters to show a general, to more specific area of the crimes. I examined this information more, and realized it was not arrest made, but Crimes reported. The arrest column was counted regardless of a true/false outcome. I then changed the name from "Count of Arrest" to "Count of Crime Reported" to more accurately describe the pivot table, then I put the "Arrest" field into the filter, so we can see what crimes people were actually arrested for. I tried to make a graph here, to show what was the most reported crime, as well as the number of reports per district. My Excel kept crashing while attempting this, So I decided to find them in R.

The R language seemed daunting at first, because I was unfamiliar with the language. After doing some research, it was not too bad. What took the most time was trying to upload the data into R, that worked for both my home, and on campus. I was able to load the data at home relatively quick, but when I tried uploading the data on school I was getting an error message. The message said something along the lines of "cannot open the connection, unsupported URL

scheme". After numerous attempts of trying, I read somewhere that https was not supported, so I decided to try my code, but removing the "s" in "https". The line of code was accepted. In the code to upload the data I stated there was a header, and in order to check that I used the "str()" input. Everything looked in check, and moved on to find the summary of the data, primary type of crime reported, as well as the arrest made. I had to look for a way for R to find the mode of a field, and found "names(sort(-table(x$y)))", where x was the uploaded data, and y was the particular column. I found the mode for primary type of crime, IUCR code, and location description. The IUCR code stand for Illinois Uniform Crime Report code, which are codes for primary types and the crime description. The list can be seen here <https://data.cityofchicago.org/Public-Safety/Chicago-Police-Department-Illinois-Uniform-Crime-R/c7ck-438e>. For example, a 0486 means Domestic Battery, while a 0460 is simple battery. I then tried to find how to graph categorical data, so I can have a visual representation of primary types of crimes reported. I found out I had to create a new table, containing information only from that one column. From there it was using the bargraph() input to create the graph. I also found out how to put it in descending order, so I was able to get a visual representation of the primary type mode as well. I looked at data, and decided that the best way to show high crime areas was to use Community Areas. A map of community areas can be seen here < https://portal.chicagopolice.org/portal/page/portal/ClearPath/Communities/Districts/community_area.pdf>. After creating a new dataset to ignore NA variables, I used the hist() input but was not happy with the results. The graph was not matching the mode I found for community area, and after analyzing the graph, I realized the only problem was that areas 1 and 2, were being combined. I looked up a way to separate the two areas, and the graph now matches the mode found.