# Section 5.5 Summary for Section 5

• In markdown explain how df\_calls\_disp was created and how it can be useful

Df\_calls\_disp has merged from two data sets: df\_calls and df\_disp.

Df\_Calls was originally imported from the calls table in section 4. Several new columns and data were created after the initial data was imported. This includes converting Date\_Received from an object into a date. In addition, Day\_Name, WEEK, YEAR, YEAR\_WEEK and SUB\_YEAR\_WEEK were added as new columns to the data set. In section 5.1, dummy variables for “Day\_Name” , “MONTH” and “Complaint” were created in order to convert the values into numerical values which was later concatenated with the main data set of df\_calls.

Df\_disp was originally imported in section 4 from the Disposition table. A separate data set (df\_disp\_cat) which was imported from a csv file would be later included in df\_disp. A dummy variable for Disposition from df\_disp was created to get numerical values and then later concatenated with the rest of the main data set for df\_disp. Afterwards, df\_calls and df\_disp would be merged into one data set which is now df\_calls\_disp. Missing values in data set were replaced with 0.

df\_calls\_disp is useful for the police department because it consists of 10 years of data that includes, subzones of canton, complaints ( ex: Burglary, Assault, Welfare check etc.) , date received, month 1-12, calls for armed subject, assault, burglar alarm, fights, noise complaints, possible drugs present, stalking, shots fired and special residence patrol. These columns are in numeric values and are formatted to be ran through models such as a regression, cluster analysis etc. This data will be valuable in analyzing which areas are showing higher volumes of crime in order for Canton P.D. to focus their resources on certain subzones.

# 8.3 Summary for Section 8

• How can spatial lag be helpful in a predictive model using df\_calls\_disp\_week

Spatial lag is a variable that averages the neighboring values of a location accounts for autocorrelation in the model with the weights matrix. Specifically, a common predictive model for spatial lag is a spatial lag regression model that considers dependent variables on an area with other areas associated with it. Regression models investigate what variables explain their location. For example: if you have crime locations in a city, you can use spatial regression to understand the factors behind patterns of crime.

# 9.4 Summary for Section 9

• Congratulations, you made it! How can this final dataset df\_pred\_calls be used to make

predictions?

• Explain in a few paragraphs (with bullets/dashes) what you think the final df\_pred\_calls

can be used for in helping the police chief make better decisions

Based on the final dataset, we have details such as the subzone, sub year week, call all, call burglary, month 1-12, calls for suspicious activity, calls for trespassing, calls for vandalism, calls for welfare check, calls for shots fired, calls for possible drugs present etc. With these details, we can pinpoint complaints with dates when they were made. We are able to determine which subzones have the highest volume of crime in addition to what month of the year the crimes happen the most.

For example, we can determine which subzones have the highest volume of calls involving shots fired with specifics of dates, days, week of a month, month of the year when they happen the most. With this information, Canton police department can optimize its resources to respond to violent crime in certain subzones depending on the month of the year.

The described outcome one can achieve by performing a series of linear regressions between the feature variables (e.g., specific days, week of month, month of year) and a selected target variable (e.g., number of calls for suspicious activity, trespassing, vandalism, shots fired). One would look for high coefficient of correlation between the feature variables and the selected target variable to understand which incidents are more frequent on a particular period in the feature variables; the coefficients of correlation for each linear regression would be available in the correlation matrix generated by the linear regression model.

To determine the correlation coefficients are valid, one would have to review the linear regression model outputs and validate the assumptions the data has a normal distribution (using histogram charts, looking for a close bell curve shape in the histogram) and that there are no multi-collinearity issues by reviewing the variance inflation factors (VIF). The presence of high VIF values (over 10) would indicate that multi-collinearity issues are present, and one could be suspicious that the high correlation coefficients are valid.

Moreover, one could perform a clustering analysis of the target variables (e.g., number of calls for suspicious activity, trespassing, vandalism, shots fired) by subzones, to determine which zones have high clustering of criminal activity, which would enable Canton P.D. to allocate surveillance resources to the subzones that have high concentrated clustering of the target variables.