Exercise Answers

Stuart Barnum

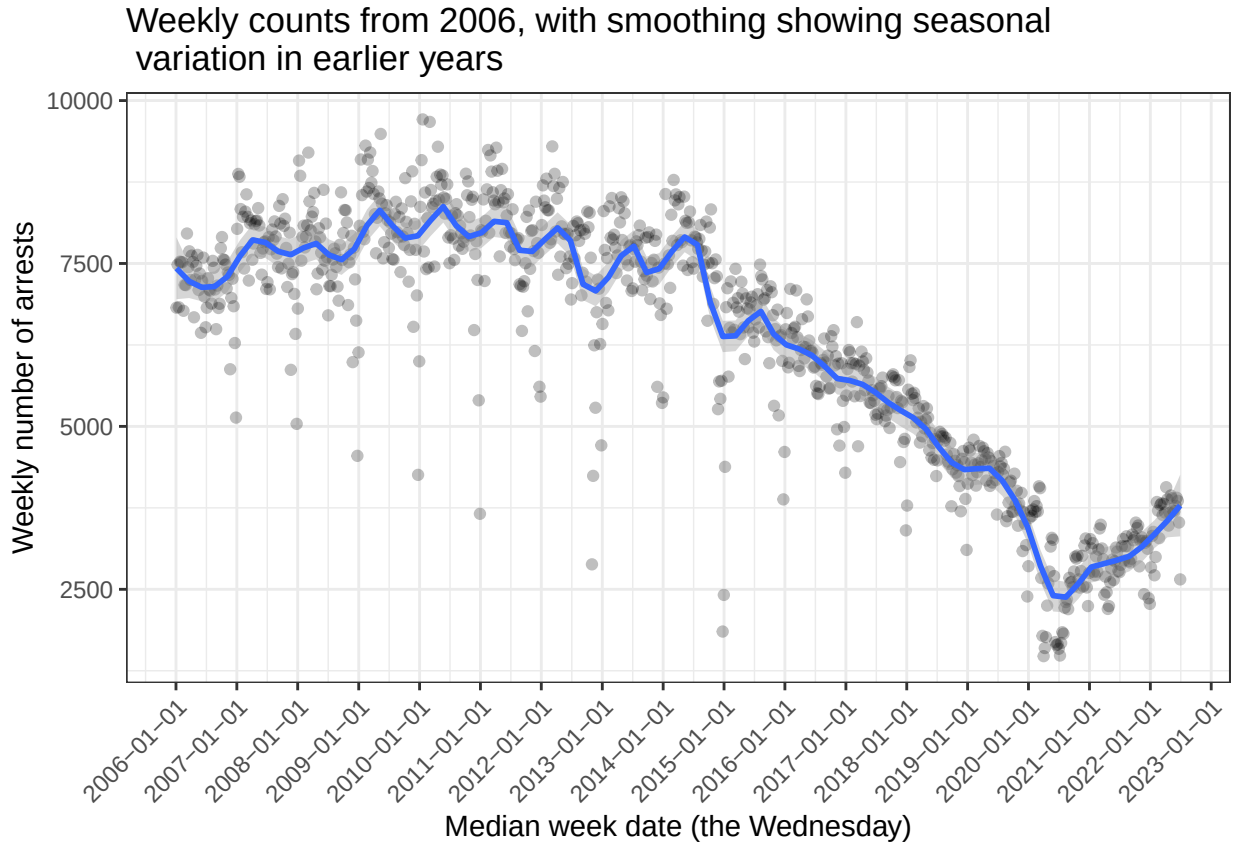
9/7/2022

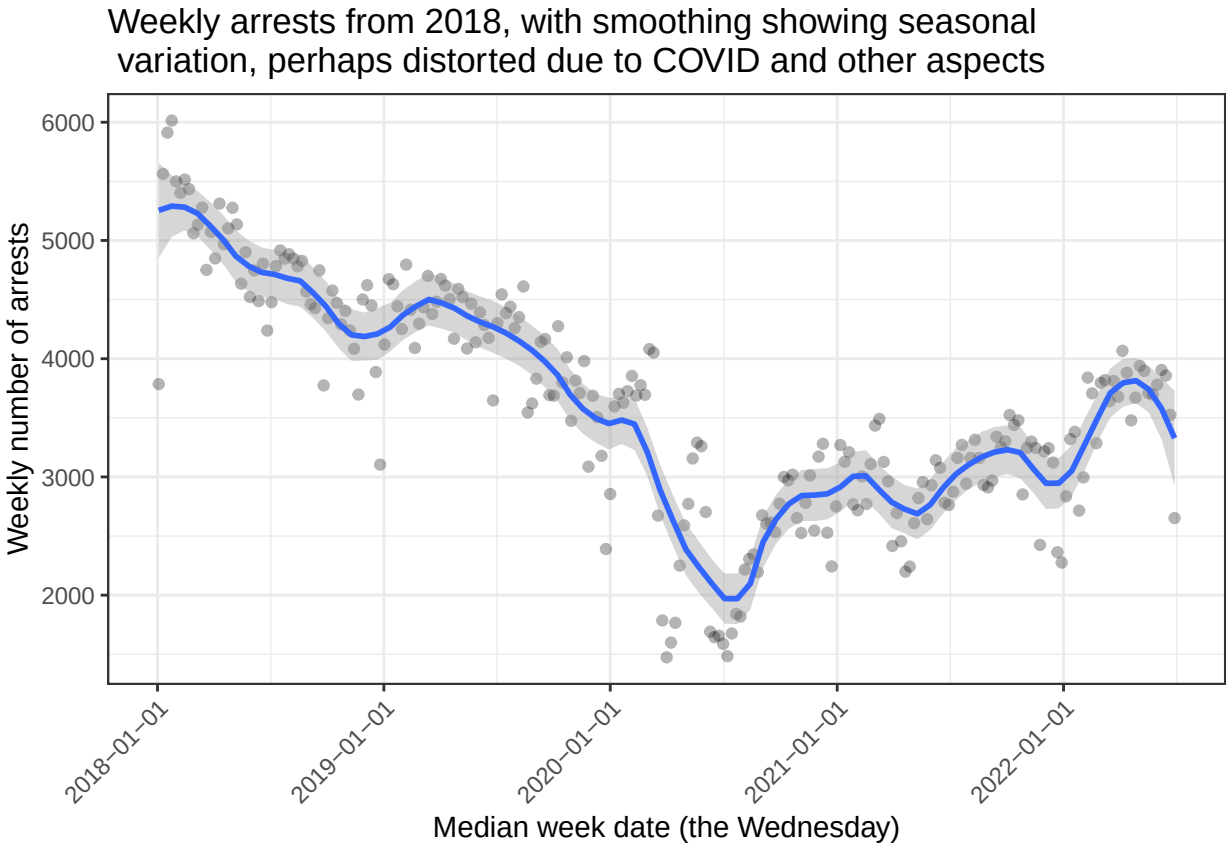
Question 1

Has the arrest rate been decreasing from 2018-2022? Describe the trend and defend any statistical tests used to support this conclusion.

The arrest record has not decreased from 2018. It decreased from 2018 through mid 2020 and then increased thereafter.

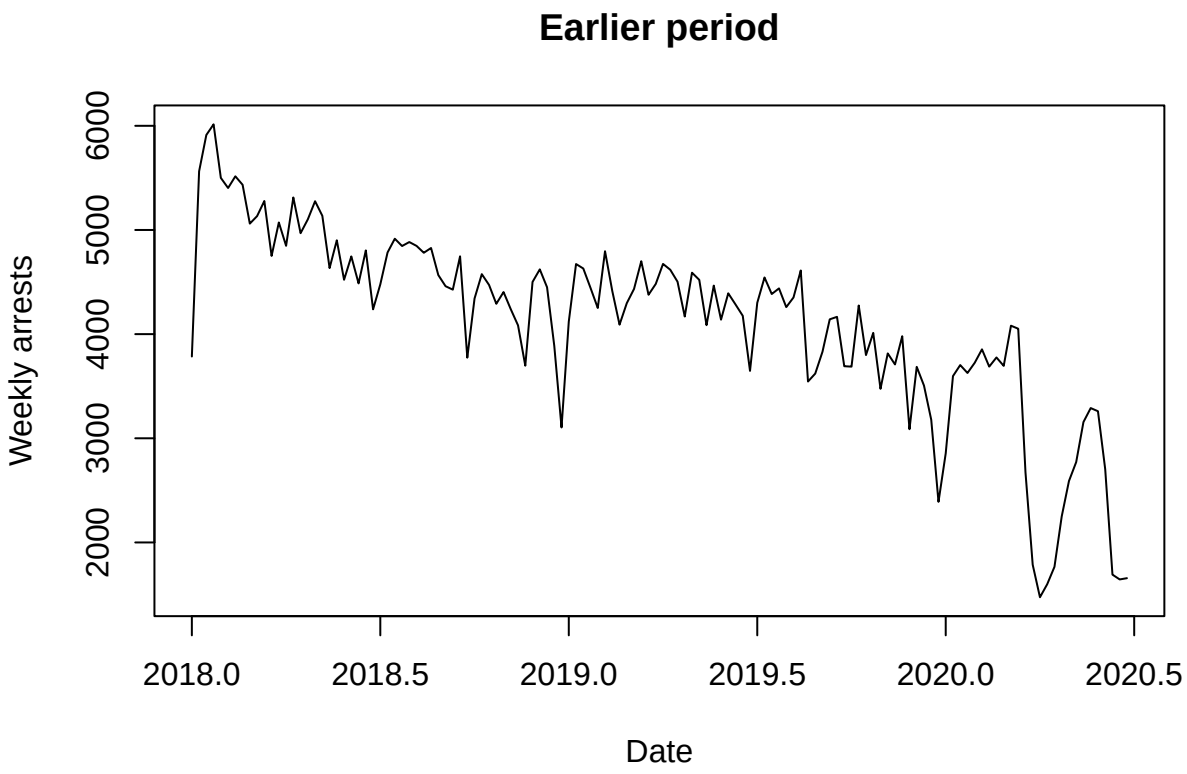
We might expect to see differences between week days, with e.g. more arrests of certain types on weekends. I will bin the weeks, obtaining the count for each week and assigning these counts to the median day of the week (Wednesday). With this, we might better visualize trends over time periods greater than a few weeks. Note in the graphics the cyclical variation before 2015, which disipates after that.

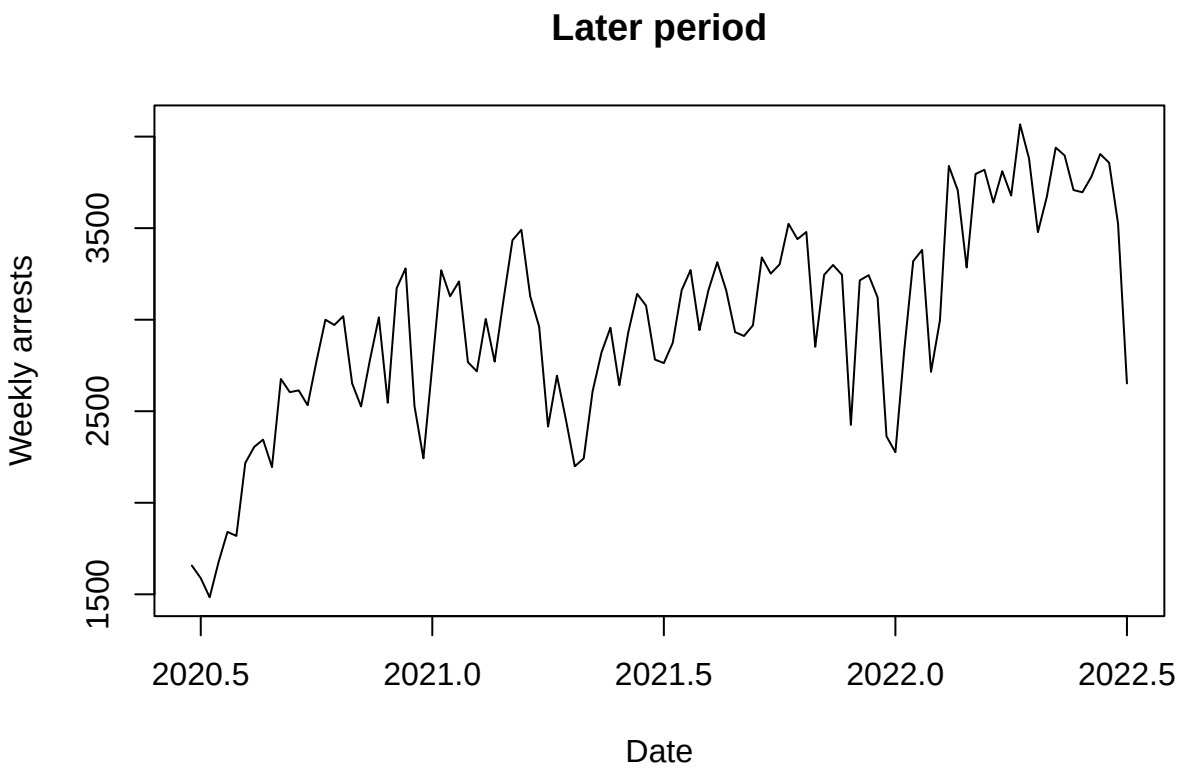




It appears that arrests decreased until the middle of 2020 and then started to increase. I will divide the data into these two time periods and apply a Mann-Kendall test with sieve bootstrap to the two time-series resulting from separation of weeks before middle 2020 from weeks after middle 2020. This test allows that data be autocorrelated and that there be periodicity, which seems to exist, even if it is rather irregular. I will use this statistical test repeatedly, and then use a simple bootstrap approach at the end.

The following two plots show the results of the separation of the two time periods


Earlier period
Weekly arrests
2000
3000
4000
5000
6000
2018.0
2018.5
2019.0
2019.5
2020.0
2020.5
Date



With my statistical tests using the null hypothesis of no trend, both of the trends—the downward trend before middle 2020 and the upward trend after middle 2020 are significant ($P < .01$).

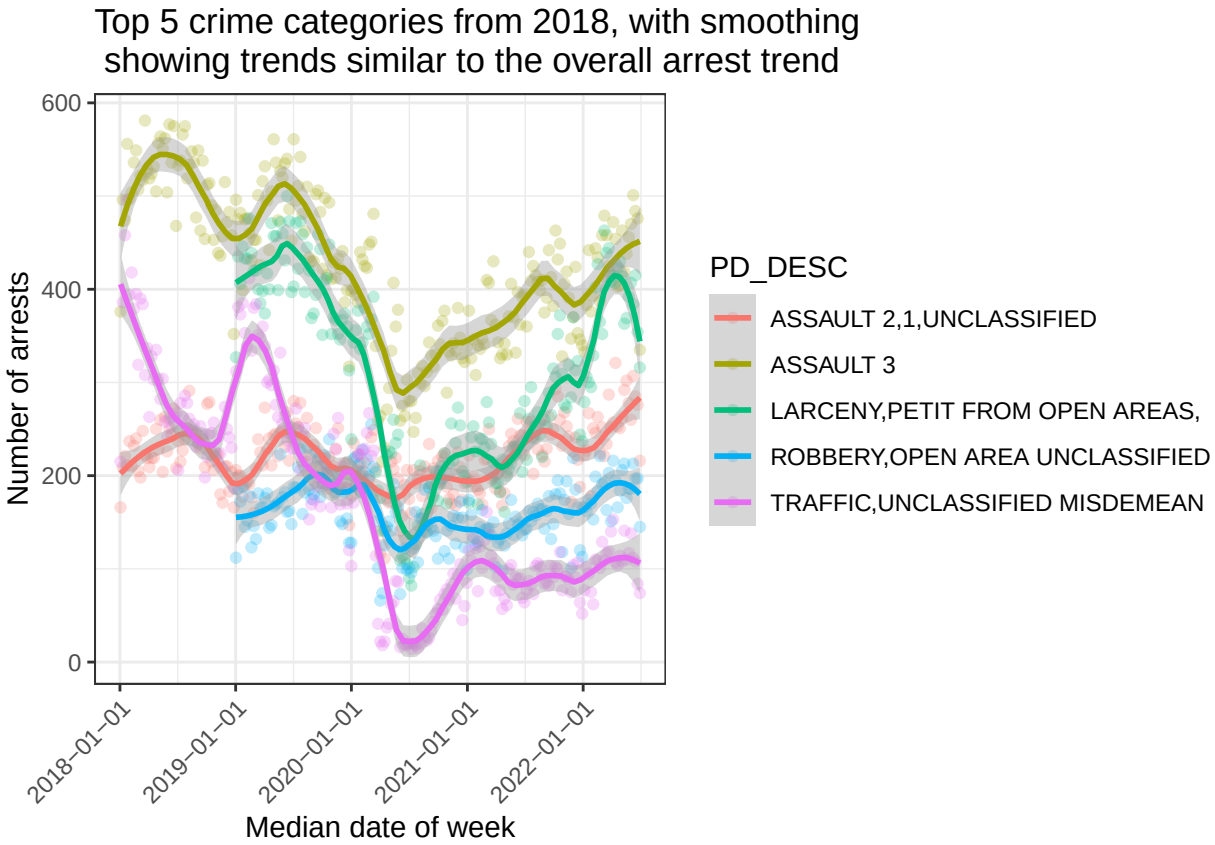
Question 2

What are the top 5 most frequent arrests as described in the column 'pd_desc'in 2018-2022? Compare & describe the overall trends of these arrests across time.

The crime categories with the greatest numbers of arrests since 2018:

PD_DESC	n
ASSAULT 3	99757
LARCENY,PETIT FROM OPEN AREAS,	56001
ASSAULT 2,1,UNCLASSIFIED	51694
TRAFFIC,UNCLASSIFIED MISDEMEAN	40527
ROBBERY,OPEN AREA UNCLASSIFIED	29773

For the 5 categories with the highest count, we tend to see similar trends as with the overall counts–decreasing until about middle 2020 and then increasing. Except that at least one of the crime categories shows weak trends or perhaps no statistically significant trend at all. We check the statistical significance using the same test as with Question 1. Note that two of the categories show no 2018 data. These may be new or renamed categories (I have not explored this).



Assault 3 and Larceny show strong decreases before middle 2020, while robbery shows little decrease over this period. All five of the crime categories show moderate increases after middle 2020. Statistical tests for all 5 categories over the two time periods (10 tests) support rejection of the null hypothesis of no trend ($p = .05$) in every case except for Robbery before middle 2020.

Question 3

If we think of arrests as a sample of total crime, is there more crime inprecinct 19 (Upper East Side) than precinct 73 (Brownsville)? Describe the trend, variability and justify any statistical tests used to support this conclusion.

This question asks us to compare crime rates, not just trends in crime rates. It can be misleading to say that one area has more crime than another area, without consideration of the population of those areas. I therefore found some population numbers and compared the per capital crime rates.

My sources for the population data:

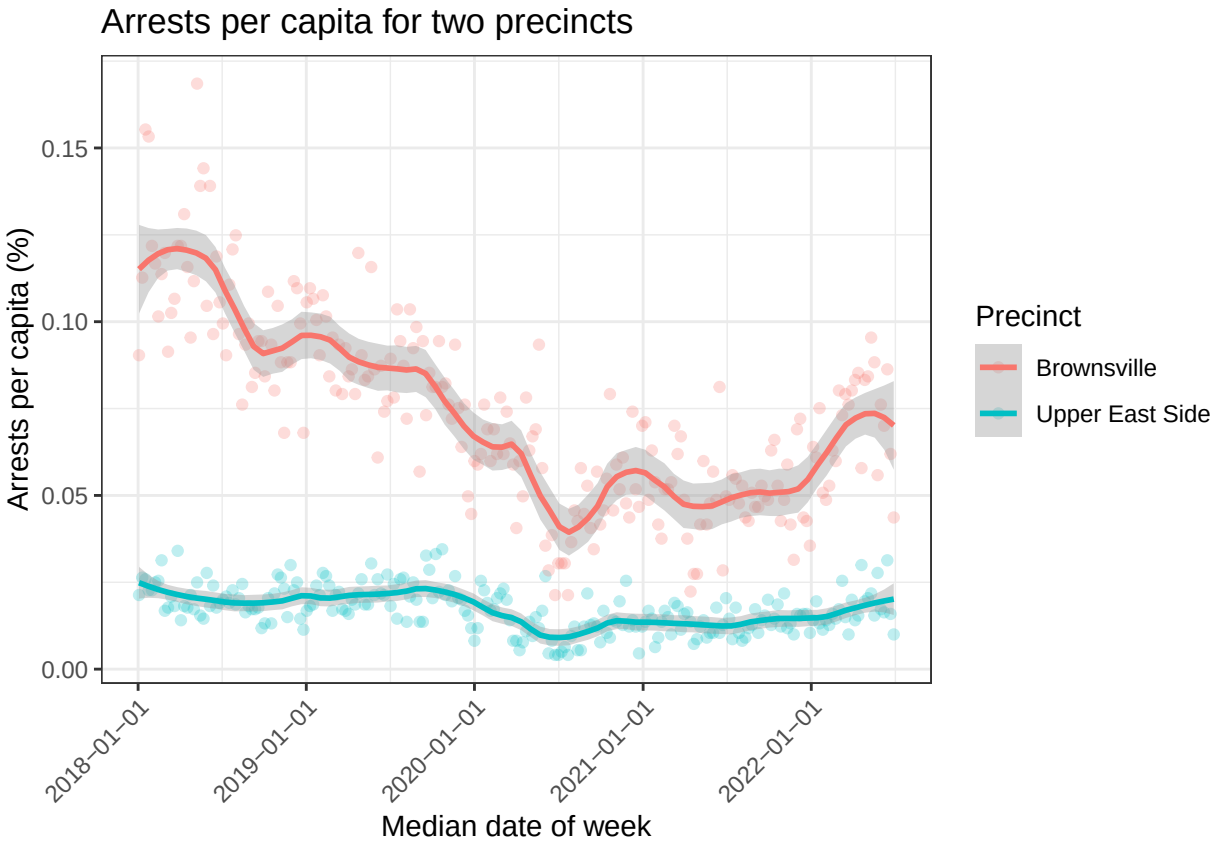
https://johnkeefe.net/nyc-police-precinct-and-census-data

https://github.com/jkeefe/census-by-precincts/blob/master/data/nyc/DECENNIALPL2020.P1_metadata_2022-02-06T162722.csv

Another tricky aspect with Question 3 is the seeming autocorrelation (over time) in the weekly crime rates. Most statistical tests that would compare, say, the average weekly crime rates of these areas, require independence of the data points. This condition fails here. I use a simple bootstrap to ameliorate this issue. I am not certain that my solution is ideal.

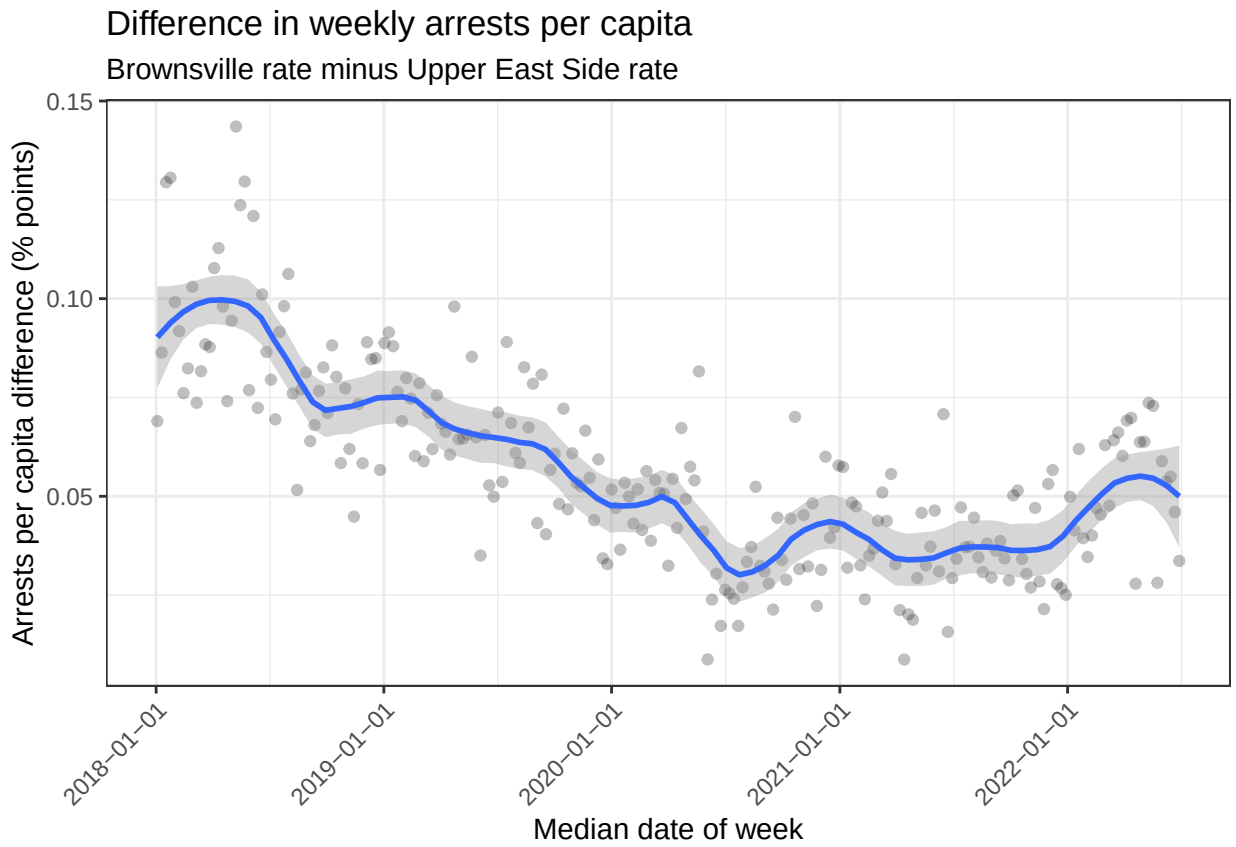
Look at the arrests per capita for the two precincts. From the Keefe data, per the 2020 cencus, Precinct 19 has population 220,261, while Precinct 73 has population 98,506. I will assume that these numbers are

roughly accurate from 2018 to the present.



Altough Brownsville shows much sharper trends than the Upper East Side, downward and then upward, the both the trends for both areas are statistically significant with null hypothesis of no trend (the Mann-Kendall test with sieve bootstrap, again.)

I now look at the trend in the weekly differences between these two areas.



We find statistically significant trends in the *differences*, between the weekly arrest rates, downward and then upward (p«.001).

The analysis of the overall differences, such as whether the differences in weekly arrest rates, between the two precincts, is statistically significant is complicated by the fact that the observations for each week are not independent. Most statistical tests, including non-parametric tests, assume independence. I ameliorate this issue by using a simple bootstrap—resampling the data with replacement. We use, again, population-adjusted arrest rates.

To quantify the overall differences between Brownsville and the Upper East side, I use confidence intervals. With bootstrap, the confidence interval of the difference for the period 2018 through middle 2020, in percentage points (for the numbers of arrests per capita per week), is (0.066, 0.073), while the confidence interval for the later period is (0.038, 0.044)

Question 4

Given the available data, what model would you build to predict crime to better allocate NYPD resources?

I would try a cluster analysis, or a number of cluster analyses, using as independent variables geographic coordinates, (x, y) pairs indicating specific points in the city, together with variables such as time of day, day of the week, the month, type of crime or zoning status (e.g. residential), etc. The type of model could be a simple k-means analysis or (preferably) a mixture model, with statistical distributions assigned to the clusters (the task of the model fit is to estimate the parameters of these distributions). Either type of model (k-means or mixture model) may produce centers where crimes and arrests of certain types tend to be concentrated. These centers in n-dimensional space, e.g. in 3 dimensions with the x and y coordinates and time since the start of the week. It might be useful to create separate models for certain, different, crime categories.

In a mixture model, the value of the dependent variable we are predicting (e.g a y on the left-hand side of the model equation) would be the value of the probability distribution function (for the probability of, say, an arrest). However, in addition to independent variables that have values in the data (e.g. the geographic coordinates), a model such as this will have latent traits or variables such as the cluster centers or the variances of the clusters. Variables or traits of particular interest will include (e.g.) cluster centers and (for mixture models) other parameters describing the clusters (e.g. variance of a normal distribution). In a similar fashion, in k-means clustering, the cluster centers will be of particular interest

A model such this might enhance efficient placement of police resources in certain locations at certain times, based on the properties of the clusters this model might discern. It might also, depending on the data that is used, allow allocation of police officers with certain specialties (with focus on certainly types of crimes).

As for model evaluation, the AIC or BIC criteria could be used to select one model over another. However, these measures will not tell us whether the final result is a good one (only that it is the best among possibly poorly-performing models). Cross validation could also be used for selecting one model over another. Two ways of evaluating a final model would be validation with data that is held out (a test set) or building a simulated dataset using the model. The actual data distributions and the simulated data distributions might then be compared using QQ plots (quantiles of one distribution plotted against quantiles of the other).

Challenges I might face include lack of direct access to what we want to measure. For example, we may know the arrest rate and need to use the arrest rate as proxy for the crime rate. The arrest rates might somewhat reflect greater amounts of policing in certain areas rather than greater amounts of crimes in those areas. Further analysis might be required to disentangle numbers of arrests due to actual crime numbers and numbers or arrests due to greater presence of policing. Another challenge, as always with large amounts of data and sophisticated models, will be limitations on computing resources.