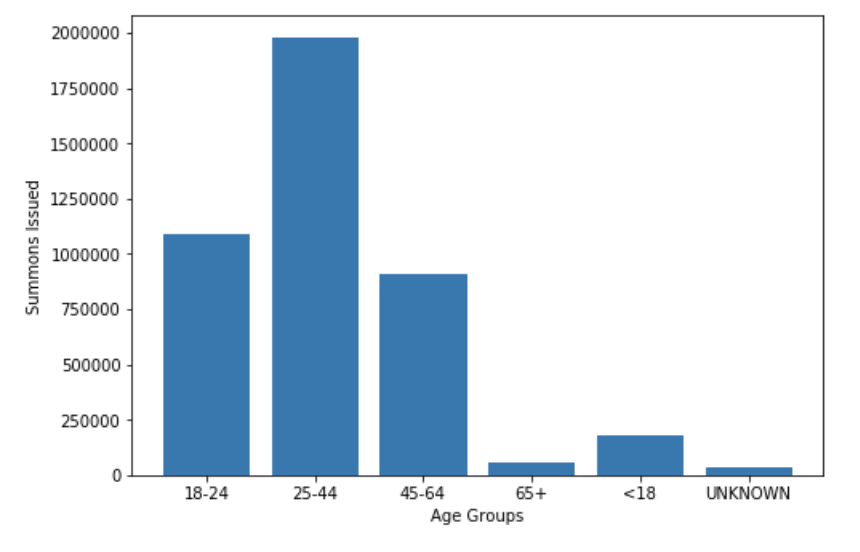
I settled on the NYPD historic criminal summons data, [found here](https://data.cityofnewyork.us/Public-Safety/NYPD-Criminal-Court-Summons-Historic-/sv2w-rv3k). There are several data quality issues, even after cleaning, such as the fact that the native bucketing for the ages is severely skewed (the age categories are <18, 18-24, 25-44, 45-64, and 65+), with 20 years between the 25-44 and 45-64 categories, but only 6 years in the 18-24 category, and then the catchalls of <18 and 65+. I may have to do some kind of normalization, but I’m not sure what that would look like just yet. Additionally, there is an inherent lack of identification of “repeat offenders,” which means that several of these records could be the same people and I wouldn’t be able to tell.

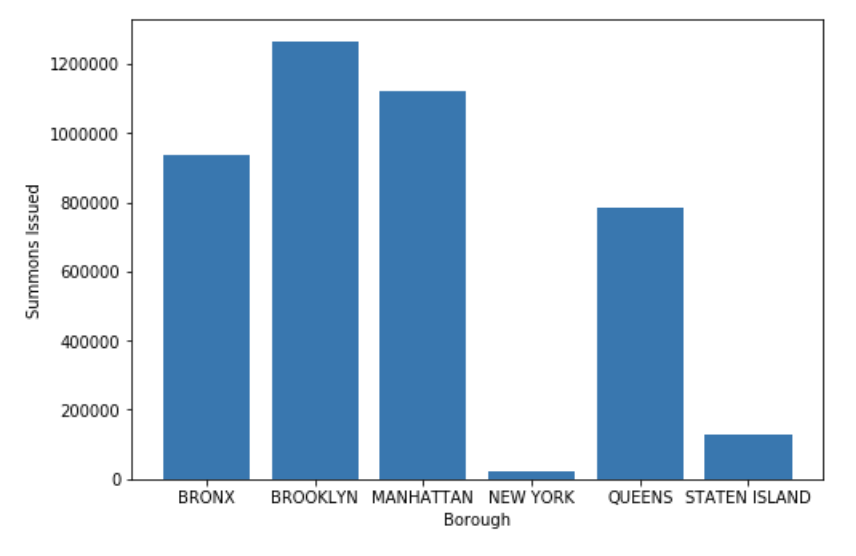
I still think this is the best option for my project because my initial question was really asking what the demographic landscape of our criminal summons system looks like. Though the data is messy (likely due to understaffing) and simultaneously huge (there were originally over 5 million records prior to cleaning; I am planning on either sampling with replacement or creating categorical visualizations that won’t be as difficult to parse with the human eye), I think it’s pertinent to ask who exactly the NYPD is hitting with summons that eat into their time and, depending on the type of charge, money.

The data is in a csv file, with a unique key field and 14 other columns prior to cleaning--I removed the X and Y coordinate columns due to redundancy. I don’t think this dataset inherently engages with Beer’s conception of data rationality--it’s more descriptive than prescriptive, but I can see how someone could take this data and attempt to create a very poor model of “identifying potential criminals,” as has been recorded in the past (see the COMPAS algorithm). In a sense, that could engage with the idea of “making us better people,” but in such a way that it becomes oppressive and dangerous. I can also see a potentially looser view of data rationality here, wherein we turn our gaze from the “offenders” to the police themselves, and begin asking questions about to whom they issue summons and why.

I created a couple of graphs just to play with the data a little bit--because I’m working with largely categorical data, bar charts seemed the most appropriate here. Additionally, I didn’t want to begin sampling just yet, so I’m using the entirety of the cleaned dataset. A quick note before I go into this analysis--the data is entirely from the years 2017-2019 and the summons include several types of law, which all seem to fall under the category of “criminal summons”.



The first graph is the aforementioned age buckets and how many summons were issued. Because of the uneven bucketing, there are way more records in the 25-44 bucket than anywhere else. However, the fact that the 18-24 column has so many records without consistent bucketing is insightful in and of itself--compared to the 10-year buckets, there is a higher concentration of summons issued. On the surface, it seems that young people are more often issued summons than older people, though this could be due to a number of factors (i.e. life expectancy for certain groups).



The second graph is a borough breakdown of the summons data. The most immediate conclusion to be drawn here is that most summons are issued in Brooklyn, with the second-most being issued in Manhattan. More insight could definitely be drawn here with population counts for all of the boroughs, but as it stands, the fact that Brooklyn has a larger surface area than the Bronx or Queens could account for the larger amount of records. Manhattan is very densely populated, which makes the amount of summons issued make sense, because existing in such cramped quarters would naturally give rise to things like noise complaints and “in a public place and uses abusive/obscene language or makes an obscene gesture.” That sort of thing can generally go unremarked upon with fewer witnesses. Another thing about this graph is the fact that it highlights an interesting thing about the people populating the dataset--most of the data entry clerks use “Manhattan,” but some use “New York”. These, of course, are the same borough, but it draws a particularly fascinating distinction: who would call it New York, and who would simply refer to it as Manhattan? What do those particular records look like?