DELIVERABLE 1

- **Establish majority racial and ethnic composition of each 311 request based on block group demographics**

We use two sources of data, one is the census data for Boston in 2020 and the other is the 311 data from 2022. The census data has populations in different races and ages for each precinct and each record in the 311 data also has precinct numbers associated with it. Therefore, it is very natural for us to connect this two dataset by the precinct numbers and using the precinct number to represent a neighborhood.

We first analyze the census data by precinct: We calculate the ratio of a certain population over the total population in that precinct so for each 311 request, we know which precinct it comes from and the racial composition of that precinct.

❖ Conduct Analysis of requests

We first focus on the on-time rate and the case-closed rate. We want to know if a certain racial composition of a neighborhood can affect the efficiency of the 311 service for that area. From the last step we can get the racial composition for each precinct, and we calculate the spearman correlation coefficients of the on-time ratio and the case-closed ratio of 311 service of all the precinct and the racial composition, and we have some findings. In terms of on-time, We found that the White and Hispanic people ratio in a community both have positive correlation coefficient with the on-time rate; while the Asian and Black people ratio both have negative correlation coefficient with the on-time rate. However, in terms of case-closed rate, only the Hispanic ratio has a negative correlation coefficient with the case-closed rate.

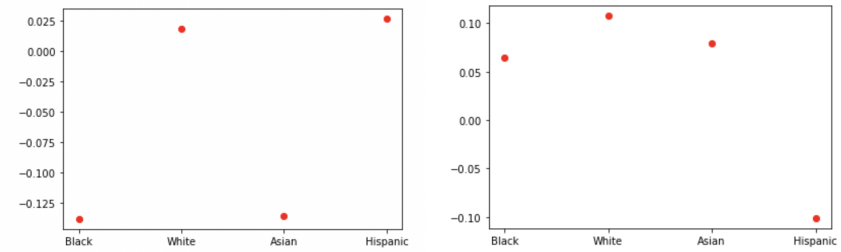

Figure 1. The left plot is the correlation coefficient of on-time rates versus respective racial proportion of communities; The right plot is the correlation coefficient of case-closed rates versus respective racial proportion of communities.

Then we try to figure out what kinds of cases a certain race of people tends to report to 311. The way we choose to analyze the problem is we focus on one race at a time: first we classify the precincts by population of a certain race into 3 clusters using k-means, each of which represents low, mid and high proportion of that race. Then we plot the number of different cases for the 3 clusters.

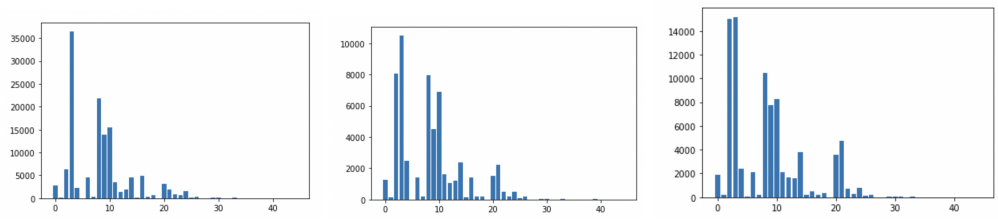

Figure 2. The left plot is the number of different cases for low white people population communities; The middle plot is the number of different cases for middle white people population communities; The right plot is the number of different cases for high white people population communities.

We first focus on the White population(as in Figure 2), the shape of the number of different types of cases seems to be similar, but we notice that the third bar in the left plot is lower than the ones in others. The third type corresponds to the 'Sanitation'. Therefore, it seems like that for communities with lower White population, cases about Sanitation is reported less. The reason could be there are less phenomena about sanitation in those communities or people of this race report more often, or something that is more complicated.

We perform the same procedure with respect to Black, Asian and Hispanic race. For Black race, we notice a similar pattern, however, for Asian and Hispanic, we notice an opposite pattern: communities with a high population of Asian and Hispanic report much fewer cases regarding sanitation as shown in Figure 3. The reason also could be complicated.

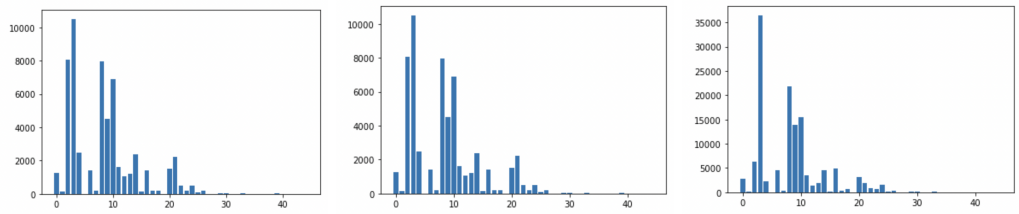

Figure 3. The left plot is the number of different cases for low hispanic people population communities; The middle plot is the number of different cases for middle hispanic people population communities; The right plot is the number of different cases for high hispanic people population communities.

- **Another angle we considered:**
  1. **Geographic angle**: here we tried to bar where we can see the proportion of total cases according to the location zip code in Boston. Where we where we found that most of the requests were from 2128, 2127 and 2124, which is mostly dominated by hispanics, whites and blacks. The percentage of case_status closed within the targeted date is really good. But also we can see in our analysis that 2108 has the most cases with 82% closed rates. There is much more we can do with this angle as the dataset has a lot more to say geographically when we take different resources too where they have defined the population(race, age, status) wrt the location zip code.