A BIT OF BACKGROUND

PROJECT OVERVIEW

Cities in the US have long been segregated by variables such as race, ethnicity, and income. We wanted to investigate whether these demographic variables had any relationship with people's access to critical infrastructure like schools, banks, hospitals, libraries, supermarkets, and parks. With New York City, Chicago, Boston, and Rhode Island as our targets for analysis, we collected data from the US census to understand the demographic makeup of cities at the block group level. We also used the Google maps Places API to collect a list of all places of interest (POIs) available within each block group. We used this information to compute a connectivity score for each block group, measuring how well connected it is to POIs.

OUR ML PROCESS

We used K-Means Clustering, a type of machine learning, to provide a baseline overview of similarities between block groups within a given city or state. We clustered twice, with one clustering based on demographic census data, and the other based on connectivity scores. We then looked at the overlap between the two collections of clusters to validate our machine learning process. If the clusters had high overlap, it would suggest a strong relationship between census data and connectivity data. On the next page is an example of the overlap between clusters for the city you chose to analyze. "True" represents block groups connectivity cluster and census cluster matched, whereas "False" represents when they didn't match.

HYPOTHESIS TESTING

We prepared two kinds of hypothesis tests to help analyze the data. A **chi-square independence test** can tell us whether there is a statistically significant correlation between a demographic variable and the connectivity of block groups to infrastructure by comparing the value of the demographic variable to the connectivity cluster labels. The **two-tailed independent t-test** does something similar, but at a more granular level - it can help analyze whether there is a statistically significant difference between the connectivity of block groups with differing demographics to a certain type of infrastructure.