TO: Haiyun He (hh743)
FROM: Emily Mei (em664), Jonathan Yun (jly37)
DATE: 19 March 2023
SUBJECT: Team 33 Project Proposal

I. QUESTION

With a desire to improve both local health and communities, we propose a project taking a closer look at the intersection between the two. In doing so, the main question for our project is: *How does walkability impact health?*

More specifically, we will build a classification model to determine whether walkability can be used as a predictor for various health outcomes, and if it is a stronger predictor for certain health conditions. We will also consider the other direction with a regression model: can we predict an area's walkability score if we are given health data? Finally, the walkability score for a community is calculated using a variety of factors; we will investigate whether any of these other subfactors (eg. race, income, etc.) are stronger predictors of health outcomes than just the walkability score itself.

II. DATASET

We are considering using a combination of two datasets that are widely available from the U.S. Federal Government: a Walkability Index dataset, from the EPA, and the Local Data for Better Health dataset, from the CDC. The former has data on income, employment, and demographic mixes (ordinal), residential / commercial density, transit availability, and the overall "walkability" of specific neighborhoods (as a score from 1-15); the latter has various measures of health outcomes (such as prevalence of high blood pressure), preventative care availability, and chronic disease risk.

While these datasets are technically disparate (one uses the "census block" for each observation, while the other uses ZIP code), these can be merged together using relationship files provided by the Census Bureau; we thus have data about both topics for each ZIP code, and can form the basis for regression and model-building from there.

III. RELEVANCE & FEASIBILITY

We believe this analysis is relevant because, while current research points to increased walkability improving health outcomes (Howell et. al, 2019)[1], a complete neighborhood-by-neighborhood analysis has not been done, and a truly rigorous means of predicting health outcomes does not currently exist. In the interest of improving public health outcomes nationally, especially for underrepresented communities, this is a relevant question to answer: understanding more about this relationship can, for example, lead to policy recommendations to improve nationwide health outcomes.

Given that our analysis largely focuses on 3 main models / questions to answer, and that the data we have chosen to use is both manageable in size (~ 400 MB) and mostly complete (few missing / NA values), we think the analytics in this project are more than feasible over the next few months.

---

[1] Howell et al., "Association Between Neighborhood Walkability and Predicted 10-Year Cardiovascular Disease Risk: The CANHEART (Cardiovascular Health in Ambulatory Care Research Team) Cohort", *Journal of the American Heart Association* 8, no. 21 (November 2019), https://doi.org/10.1161/JAHA.119.013146.