Title: Active Living Areas in Madison
Name: Andrew Schaefer

The Active living index is a score out of 100 assigned to an area depending on a number of mobility components. The City of Madison has a dataset (https://data-cityofmadison.opendata.arcgis.com/datasets/active-living-index-composite-score) for its own city that considers seven components; bicycle facilities, bicycle level of service, street intersection density, population density, employment density, transportation services, and walking destinations. Each of these individual components are assigned a score, then summed up to calculate the Active living index score (ALI), for most of the rows.

The goal is to determine which of the factors have the most influence on a high ALI score. This can identify the most important component that could be improved in the Madison area to promote a healthy, active environment. I added a column called "High ALI" to capture areas with ALI scores of over 35.

Figure 1 shows the summation of the scores assigned to each component. This summation is split between areas that are active (areas that have a ALI of over 35) and areas that are not as active. As we can see, the intersection density contributes to the most activity in low-ALI areas. For areas with high activity, walking destinations and intersections make up more than 65% of all the scores. Walking destinations is the largest contributing component (32.7%) for active areas. It also has a considerable increase when compared to its contribution in areas with lower activity (20%). Population density is the lowest at 2.8% in active areas.

In **figure 2**, I performed a principle component analysis on the components. At first, I found that 1 component was enough to capture 87% of the variance, and 2 components could capture 94%. Upon further analysis, I found that the components had a different range of values (for instance, destinations had a range of 0 – 36 and population had a range of 0 – 4). After scaling the data with StandardScaler, I observed a more gradual slope for number of components vs explained variance. This tells me that each of the components has a meaningful contribution to the variance of the data. This also revealed that the dataset is reduceable to 6 components, since it captures 98% of the variance. Even 5 components is enough to capture 91% of the variance.

I split the data into train/test sets (with a 50% split), then used a pipeline with StandardScaler and LogisticRegression to predict the 'High ALI' column. The predictor has an accuracy score of 99.9% for the train and test sets, which is expected, considering how you could make the same prediction by summing up the components and checking if it is above 35.

Figure 3 shows the variable coefficients for each of the components. We can see that walking destinations (with a weight of 15.6) is the most important factor in determining if an area has a high ALI score or not. Intersection density is the second most important factor. This is consistent with figure 1, since these two components contribute the most to active areas. The increase in walking destinations that we discussed earlier could have contributed to a weight higher than that of the intersection component. As expected, population density has the lowest weight (which is 1.0) since it has the lowest summation of scores in active areas.

Overall, the data suggests that more walking destinations and intersections will result in higher level of activity for an area. This could also mean that more crowded areas are correlated with a higher number of walking destinations or intersections. An easy way to improve activity would be to install more bicycle shops around the area, since this is a combination of an extra walking destination and a bicycle facility.

Figure 1: Distribution of Component Scores

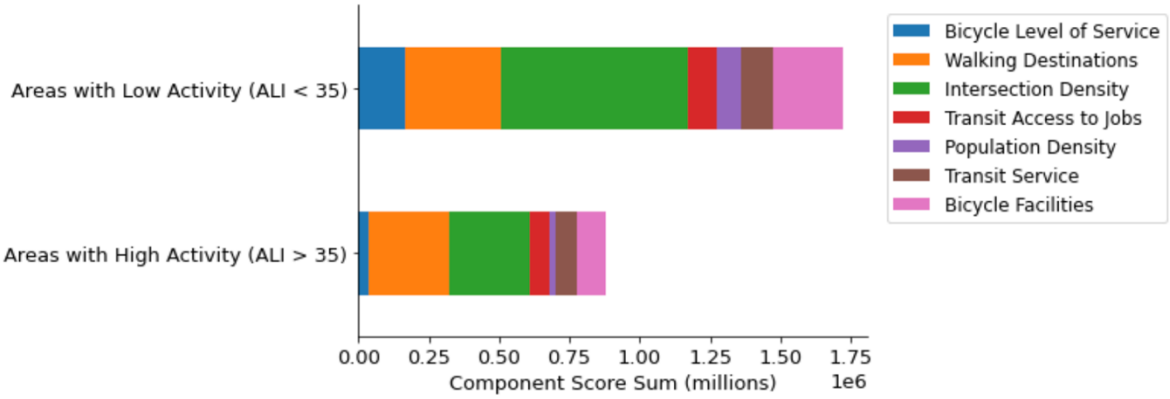

Figure 2: ALI Principle Components

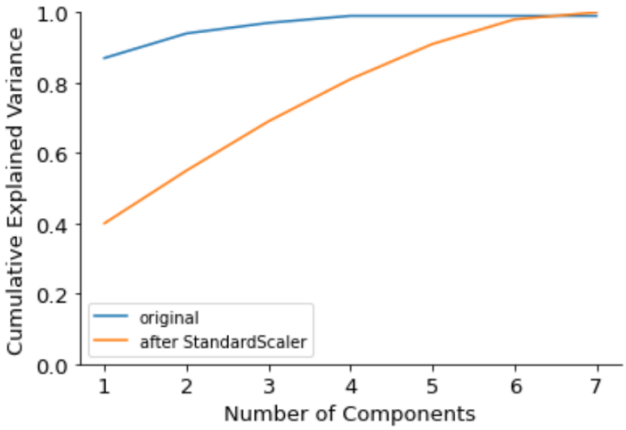

Figure 3: Logistic Regression Coefficients