Exploring Residential Segregation through Machine Learning: An Examination of Language in Real Estate Listings and its Spatial Distribution in Durham, North Carolina

I. Abstract

This study examines the spatial distribution of "residential" and "investment properties" in Durham, North Carolina, and their relationship to racial and socioeconomic demographics. We categorize and map different property kinds using machine learning techniques, Natural Language Processing (NLP), and geospatial analysis based on the descriptions in online real estate listings. According to preliminary research, 'residential homes' are mostly found in wealthy, white communities. This study offers an investigation of the linguistic dynamics in real estate advertisements and their possible bearing on racial inequalities and residential segregation. We provide a novel method to investigate the intricate interactions between real estate marketing language, housing discrimination, and community transformation by merging NLP, machine learning, and geospatial analysis. Our research adds a new perspective to the ongoing discussion on housing inequality by deepening awareness of the systemic forces influencing urban environments. Our capacity to recognize and battle subtle forms of existing manifestations of housing discrimination might be improved by applying the technique and findings presented as a model for comparable investigations in other metropolitan contexts to find the divide among racial groups and identify a dual housing market.