**Literature review (min 3 sources) 300**

* Sources for context vs. sources for methods
* How does this inform my approach? My methodology?
* Demonstrates the importance of your investigation
* Concise quotes
* Critical approach
* need to be relevant!

Introduction

According to the Understanding Inequalities Project ''the housing market has a key role in "sorting" poorer households into areas with the worst pollution, schools, crime, and employment.'' (*Housing | Understanding Inequalities*, n.d.). While housing is a basic human need, for some, homeownership is an unachievable goal and for others another great investment opportunity. Housing affordability has long been a very pressing matter for cities around the world and although housing discrimination and practices such as Redlining in the USA have been condemned decades ago, the housing crisis in states such as California is still present (Chew and Muñoz Flegal, 2020).

This study will investigate whether house index price variations are affected by environmental and socioeconomic factors and test the premise that pollution, unemployment, and health statistics all contribute to the overall house price trends. For this, California is chosen as a case study, and a variety of environmental, health and socioeconomic indicators for its census tracts are used to predict house price index variations for the same area using and comparing different machine learning algorithms.

--------------------

House prices, like the economy, are in constant flux and are of course affected and driven by many factors, which is why house price estimation and the study of the drivers behind has been and still is, a broad field of study. For decades many of these studies were based on hedonic regression models (Jafari & Akhavian, 2019) that quantify the influence of the various factors on the good (i.e. price), estimating the influence attributes such as the number of bedrooms or the location have on the demand or price of the house (Hedonic Regression Definition, 2022).

(Hanink et al., 2012) and (Montero et al., 2018) however criticise these models as they do not incorporate the spatial parameter inherent in property data. They argue that spatial heterogeneity and spatial autocorrelation need to be accounted for and therefore proceed their analysis based on Geographically Weighted Regression (GWR) and spatial hedonic models such as spatial error models (SEM) and global spatial autoregressive (SAR) models.

Machine learning methods have been employed to model housing prices since the early 2000 and as Park and Kwon Bae indicate, multiple studies have tested their performance against traditional hedonic models. The authors themselves proceed to comparing various classifiers and conclude that these can significantly contribute towards accurate house price predictions (Park & Kwon Bae, 2015).

Similarly, but more recently, (Phan, 2019) compare different algorithms, among them, Regression Tree and Support Vector Machine (SVM), in combination with dimensionality reduction methods in order to predict house prices in Melbourne from the properties’ features.

Multiple studies have examined the relation between environmental and socio-economic factors and the real-estate market. (Barreca et al., 2018) showed the spatial correlation between housing and social vulnerability indicators and house prices while (Montero et al., 2018) focused on the ­­impact of the environment, concluding that house prices are strongly affected by pollution and odours. However, their measure of environmental factors was based on the resident’s perception, which is a somewhat subjective measure and arguably difficult to be generalised. A paper by Boyle and Kiel, is a comprehensive review of hedonic model studies that considered the environmental factors of air quality, water quality and distance from toxic sites (Boyle & Kiel, 2001).

This study will be conducted not for a specific city but on the whole of California using the spatial unit of the U.S Census tracts. Differently from previous studies and partially driven by the unavailability of data in the public domain, this research will not focus on individual house prices but the change in the House Price Index for each census tract which will also enable to shift the focus to the influence of environmental and socioeconomic factors rather than the properties’ attributes.