Students: Un Leng Kam 1178863
Lang (Ron) Chen 1181506
Yujie Li 1174055
Aoxiang (Sean) Xiao 1174270

Proposal

The research question to be investigated by our team is, "**Is there a correlation between housing prices and living standards in different parts of Victoria?"**. Housing prices vary between different suburbs in Victoria and the factors contributing to such differences are to be investigated in this project. Our team proposes that the number of public facilities such as sporting centres and parks, as well as access to services such as hospitals, restaurants and shopping centres will lead to higher house prices. We believe that the existence and dense distribution of the aforementioned facilities contribute to increasing the local residents' standard of living; however, we also believe that an area with higher housing prices will lead to over-development of an area, thus, degrading the sustainability of the location and affecting future living standards. Therefore, by exploring our research question we will be able to gain a better understanding of health and sustainability of communities in Victoria.

This project is worth tackling because the quality of living is important for every individual, and when a large sum of money is spent on purchasing a property, it is important for the purchased house to generate a higher welfare than the sum that was spent. Hence, by investigating this project our aim is to determine if house prices have a correlation with living standards.

Stakeholders and people whom to which the results of this investigation would be of interest include the government, investors and potential home buyers. For the government, through this project they could gain a better understanding of the actual value of house prices, and subsequently target areas for development to improve the people's lifestyles. For example, if the project returned a positive correlation between housing prices and the existence of a certain infrastructure, the government could use this information to moderate house prices by building more of it and thus making it less unique to certain suburbs, which has the effect of regulating house prices and improving living standards, making Victoria a better place to live. For real estate investors, they could use the result to match up watchlist properties with infrastructure being built around it to predict which of them would likely grow the fastest in value and hence give them the highest rate of return. It could also help them understand what renters may want and thus make investment choices accordingly. For home buyers, they could use this project to understand if spending a large sum is worth it compared to spending a similar amount in other suburbs. If the project comes to the conclusion that high-priced areas will over-develop in the future and hence affect sustainability, home buyers can use the result to make more informed choices when making their one-in-a-lifetime decision.

The two or more open dataset we plan to use are Aurin and Data Vic. We plan to use multiple open data from each dataset, where most of them are csv's and the remaining text documents. The information that these data will contain includes the median housing prices in different suburbs of Victoria, the number of facilities in different suburbs, development plans, and pollution data across Victoria. We believe they can be linked together after pre-processing of the suburbs (our common key) and aggregation of individual data files, then for each suburb we will have their respective statistics such as house price, pollution, number of facilities and presence (boolean) of major developments in the area.

The result of our wrangling would likely produce different data frames where the data are then processed according to each's characteristics and ultimately aggregated into one dataframe, using Pandas. The value of this processed data compared to most of the raw data files is that we would have the facilities in quantity rather than listed individually (numerical data rather

than categorical), and some statistics could be normalised to allow for visualisation. Also, we will extract keywords that match development projects in different suburbs, and try to utilise this information to evaluate whether they will degrade sustainability.

Challenges and risks we foresee is in trying to convert codes into suburb names and unifying them, as different datasets we have found either has different sets of suburbs (electoral vs suburb; different abbreviations) and/or different types of codes to represent each (e.g. 5-digit LGA code vs 4-digit postcode). Another challenge that we may face is the inability to find data on certain topics. In the early stages of database searching, we have already been forced to eliminate certain valuable datasets because they did not contain the key column of suburb. A final challenge is to find a way to measure living standards in different areas using a quantitative representation which may lead to subjections of assuming what constitutes a high living standard.

Word count: 788