Predicting the ideal suburb for property purchase

Vidyadhari Prerepa

February, 2021

1. Introduction

Sydney is a city filled with buildings and high skyscrapers. Everyone wishes to live in this city for its extreme exposure to several amenities, attractions and nature. There are a variety of suburbs in Sydney which consists mainly of independent houses (suburbs away from the ocean) to apartments in the Sydney CBD. Due to its high population, there is an extreme demand for specific types of houses that residents wish to reside in, especially the ones with a sea view or the ones that are close to train station. Due to its growing population and many expats visiting the city and falling in love with it, it has become a crucial aspect for such people to decide to buy a home in such a beautiful city. However, such a purchase comes with a lot of uncertainty such as high rates of annual rent reduction or a decline in the number of bonds lodged.

There are several suburbs that are away from the city (Sydney CBD) including blue mountains, penrith and many others. It is important to understand/analyze the differences between purchasing a home in the city and purchasing a home away from the city. Such an analyses will help the purchaser in reaping the maximum profit out of his/her property purchase.

1.1 Business Problem

This capstone project is inspired by the idea of **extracting the region where a house/apartment purchase can lead to maximum annual revenue for the owner through rental bonds**. This project also explores the chosen datasets by summarizing the features of such a home. It analyses existing data of homes in Sydney suburbs and decides on an ideal suburb.

1.2 Interest

Sydney is a beautiful city with the maximum population of Australia residing in this city. Personally, I have seen a variety of homes in various suburbs that are different from the homes in countries I have resided in so far. Hence, this project is of interest to those who are willing to reside in Sydney for a long time in the future or wish to purchase an ideal home which can give them.

2. Data Acquisition and Cleaning

The data required for this capstone is not readily available. There are a variety of aspects to cover for such an analysis. Most importantly, we need numeric data for the number of bonds lodged and the weekly rent in homes of different suburbs. Following this, we need the features of the house such as the number of bedrooms to suggest/recommend it in the final conclusion of an ideal home. Coming to the amenities point of view, we need the nearby train station (within 1 km, if present) and supermarkets (Woolworths and Coles top the list in Australian retail industry, so we will be fixing to these supermarkets) and others. Most importantly, since our problem is related to maximum profit to the owner, we require the weekly rents of all units in each of the suburbs in Sydney. Factors such as change in annual rent will also need to be considered in order to differentiate between units in different suburbs.

2.1 Data sources

This project included the following data which was obtained using the corresponding source:

1) **Sydney suburbs** – Using web scraping of Wikipedia source containing a bullet list of suburbs in alphabetical order (https://en.wikipedia.org/wiki/Category:Suburbs_of_Sydney)
2) **Suburbs postcodes** : Downloaded a dataset from Kaggle from which suburb postcodes were extracted (https://www.kaggle.com/htagholdings/nsw-suburb-median-price-years-20072020)
3) **Unit/House in each suburb** : Extracted this information from a dataset containing weekly rents which was available publicly on NSW communities & justice (https://www.facs.nsw.gov.au/resources/statistics/rent-and-sales/dashboard)
4) **First, Median, Third Quartile weekly rent** : Extracted from the dataset available on NSW communities & justice (https://www.facs.nsw.gov.au/resources/statistics/rent-and-sales/dashboard)
5) **Train stations in each suburb** : Using web scraping of Wikipedia source containing a table of train stations in Sydney (https://en.wikipedia.org/wiki/List_of_Sydney_Trains_railway_stations)
6) **Supermarkets in each suburb :** Used Foursquare API to get the latitude, longitude of each suburb and sent the venue request to get a boolean result of the presence of Woolworths or Coles within 1km.

2.2 Data Cleaning

The data was collected from a variety of sources which makes it challenging to filter and clean the data. Due to the unavailability of all the required columns in one location, data cleaning first involved creation of a dataframe with all the necessary columns. This required the following steps :

1) Collecting suburbs and their postcodes
2) Merging data from 'rent-and-sales' stats to the above table. This gave us most of the required information for data analysis such as weekly rent, total bonds lodged and many more
3) There were a lot of rows with missing values of weekly rent. These rows had to be removed as taking a mean and filling it for all the units would not be an optimal solution. Removing rows with missing values also helped in reducing the size of the dataframe (from 17,542 rows to 6,743 rows) which made the processing faster
4) The train stations data was obtained and merged with the above dataframe with an added column taking a bool value of 1 (if train station is present in that suburb) or 0 (train station not present in that suburb)
5) Finally, to make the choosing of an ideal property more realistic for tenants, the fact of presence/absence of a supermarket (Woolworths or Coles) in the suburb was considered. This was obtained using the Foursquare API and two new columns were added to the above dataframe