Predicting An Ideal Sydney Suburb For Property Purchase

Vidyadhari Prerepa

IBM Data Science Professional Certificate Capstone Project

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

3. Methodology

After performing required data acquisition and cleaning, the next step is to analyze the collected data in a way to extract the answer to the main project question. There are a variety of plots that have been generated in order to filter out suburbs which do not satisfy the condition of being an ideal suburb. Following is a screenshot of the final dataframe after data wrangling:

```
Shape of final dataset :  (3896, 14)
```

	Suburb	PostCode	PropertyType	NumberOfBeds	FirstQuartileWeeklyRent	MedianWeeklyRent	ThirdQuartileWeeklyRent	NewBondsLodged	TotalBondsHeld	AnnualChangeInNewBonds	AnnualCha
0	abbotsbury	2176	3	Total	420	500	570	133	2,422	-20.83%	
1	abbotsbury	2176	3	2 Bedrooms	370	380	400	33	374	-26.67%	
2	abbotsbury	2176	3	3 Bedrooms	480	510	545	58	1241	-10.77%	
3	abbotsbury	2176	3	4 or more Bedrooms	570	620	660	32	617	-25.58%	
4	abbotsbury	2176	1	Total	485	530	600	89	1659	-9.18%	

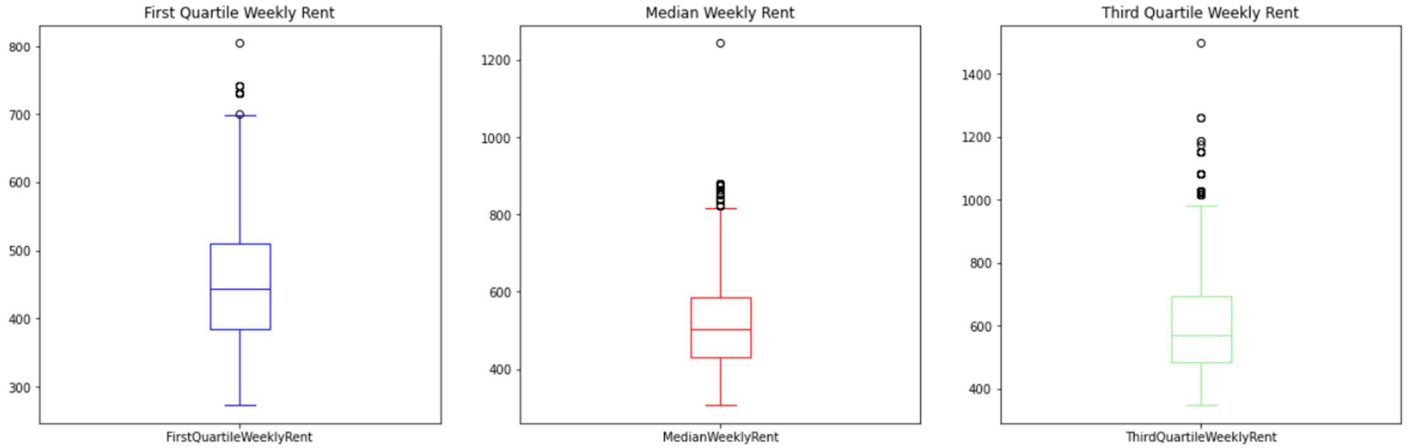
The following subsections describe each of these plots in brief.

3.1 **Box Plot Of Weekly Rent**

As we can see from the main dataframe there are three columns which contain the weekly rent, which are "FirstQuartileWeeklyRent", "MedianWeeklyRent" and "ThirdQuartileWeeklyRent". We wish to generate a box plot such that we can extract the minimum and maximum rent that the ideal property's rent can fall between.

```
Median for FirstQuartileWeeklyRent : 443.88
Median for MedianWeeklyRent : 504.71
Median for ThirdQuartileWeeklyRent : 569.29

75th percentile for FirstQuartileWeeklyRent : 509.44
75th percentile for MedianWeeklyRent : 585.62
75th percentile for ThirdQuartileWeeklyRent : 694.86
```

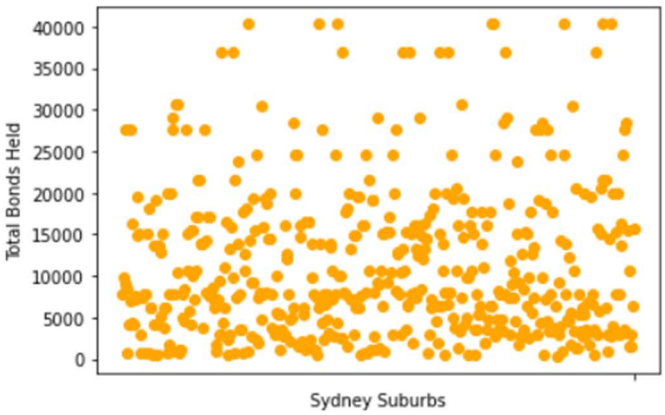


From the first boxplot which is for the column : "FirstQuartileWeeklyRent", we can observe a median value of 443.8 and a maximum value of 700. Defining our ideal property's rent to fall within this range would not be an optimal solution as the maximum value seems closer to the outliers. Hence, we define the minimum rent t be the median and maximum rent to be the 75th percentile. This rule is applied to all the box plots and their respective median and 75th percentile values are printed.

3.2 Scatter plot of total number of bonds lodged in each suburb

In order for the property purchase to be of maximum profit to the owner, it is necessary that the ideal suburb has a history of many bonds lodged. This will make it more likely for there to be at least one tenant in the property in an year which will yield good profit to the owner annually. Hence, a scatter plot has been generated with suburbs on the x-axis and total number of bonds lodged from 2009-2020 on y-axis.

Since there are a variety of units in each suburb, all the bonds of each unit (house, townhouse, apartment) have been summed up to obtain a total count for the suburb. This total count has been plotted against its suburb. The following graph was obtained:

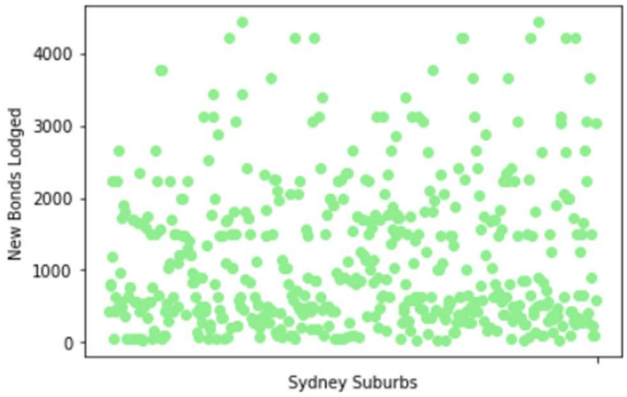


```
Suburbs with maximum number of total bonds lodged (>35000) :
casula
chipping norton
constitution hill
girraween
greystanes
hammondville
liverpool
lurnea
moorebank
mount pritchard
pemulwuy
pendle hill
prestons
south wentworthville
warwick farm
wentworthville
westmead
```

In the above scatter plot, there are a few suburbs which have a total number of bonds lodged > 40000. This is extraordinary when compared to total bonds lodged in other suburbs (few of them even having a count = 0). These suburbs have been printed and appended to a list for further consideration.

3.3 Scatter plot of new bonds lodged in each suburb

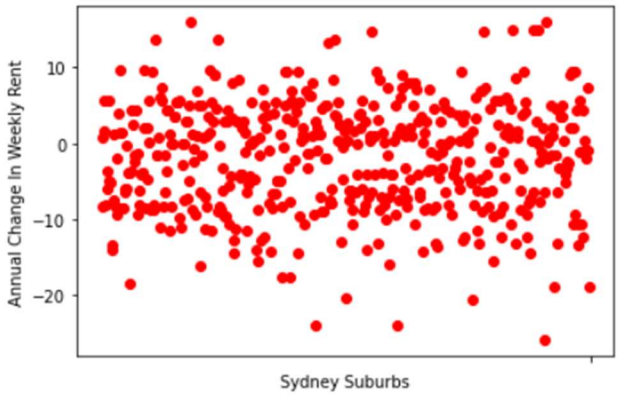
The final dataframe also contains a column for new bonds lodged. This includes new tenants and not those who have renewed existing bond with the property. Considering this feature is also important as to which suburbs are people preferring to sign new bonds in and which suburbs they prefer to reside in. After performing a few changes to the column and adjusting the 'object' type to integer, the following scatter plot was obtained.




```
Suburbs with maximum number of new bonds lodged (>4000) :
constitution hill
darlinghurst
girraween
greystanes
pemulwuy
pendle hill
south wentworthville
surry hills
wentworthville
westmead
```

3.4 Scatter plot of annual change in weekly rent in each suburb

Annual change in weekly rent is one of the most important factors as this will highly define the profit levels to the owner. If the change is positive which means that the rent has increased annually, it is highly beneficial to the owner. On the other hand, if the change is negative then it is not desirable as it means that the rents have reduced annually. Following is the scatter plot obtained:



```
Suburbs with maximum positive change(increase) in annual rent(>10%) :
bonnet bay
carramar
como
hunters hill
jannali
londonderry
richmond
st peters
sydenham
tempe
villawood
```

4. Results

The final dataframe to filter out suburbs will help in deciding which suburbs are ideal for property purchase. As we can see in the following dataframe, we observe that all the 8 suburbs have the same first quartile, median and third quartile weekly rents. These are also among those suburbs which have total number of bonds lodged > 35000 from 2009-2020 and new bonds lodged > 4000.

	Suburb	FirstQuartileWeeklyRent	MedianWeeklyRent	ThirdQuartileWeeklyRent	TrainStationPresent	Woolworths	Coles
0	constitution hill	406.736842	455.157895	502.473684	False	0	0
1	girraween	406.736842	455.157895	502.473684	False	0	0
2	greystanes	406.736842	455.157895	502.473684	False	1	0
3	pemulwuy	406.736842	455.157895	502.473684	False	1	0
4	pendle hill	406.736842	455.157895	502.473684	True	0	0
5	south wentworthville	406.736842	455.157895	502.473684	False	1	0
6	wentworthville	406.736842	455.157895	502.473684	True	1	1
7	westmead	406.736842	455.157895	502.473684	True	0	1

After obtaining the ideal suburbs with respect to weekly rents and bonds lodged, we can further filter them based on amenities such as near-by train station and near-by supermarket.

5. Discussion and Conclusion

Based on the results, we can recommend the following suburb as our answer to the question (after considering the amenities). This leaves us with 1 ideal suburb which is:

- **Wentworthville**

Though this is the ideal suburb on a higher perspective, we can find that the suburbs which have high positive annual change in weekly rent are quite different from these suburbs. We ignored those because the annual change might differ in certain years and this study is meant to provide an ideal suburb which satisfies most of the conditions for being a preferred suburb and not based on just one aspect. However, if the owner wishes to consider this aspect as the most important one in choosing an ideal suburb, then we can conclude that **Richmond** is the final answer as it has a train station as well as supermarket near-by.