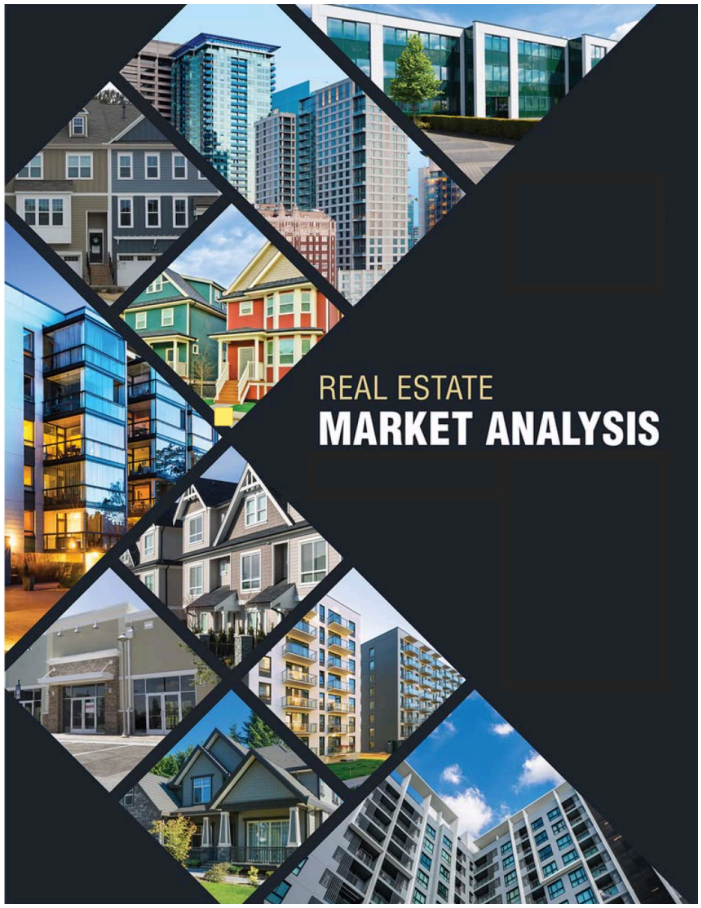


DATA EXPLORATION PROJECT (MELBOURNE REAL ESTATE MARKET ANALYSIS)

Name: Parul

Student id: 29507960

Tutorial Number: 01-P1

Tutor Name: Farah Tasnuba Kabir

Table of Contents

Introduction

Description:

For living, Melbourne is considered as the number one city according to the global liveability surveys because of its thriving art and welcoming culture. Melbourne is beloved by young and independent people for the buzz in the city, bustling cosmopolitan atmosphere and everything at doorsteps. Melbourne offers a unique network of inner-city laneways with one of the most extensive tram systems, suburbs on boundaries with waterfront appeal. This makes the Real Estate Business in Melbourne to be one of the most busy and successful Industry.

Motivation:

The Real Estate business is growing, the customers are no doubt struggling to find and choose a perfect house according to their budget and needs. Housing prices are fluctuating over past few years. There are numerous agents in the market trying to provide best deals with a percentage of brokerage. Some well-known brokers have set foots well in the market. In order to get a better understanding of the market and housing prices in different areas, one needs to perform some exploration and visualisation on data publicly available.

Questions:

1. Which is the most expensive area in Melbourne?
2. How does the prices fluctuate over time?
3. How much commission some of these real-estate brokers are bringing in over the course of the year?
4. What is the price difference between old and newly build houses?
5. Does parking space have anything to do with pricing?
6. How does crime incidence affect the housing price in that area?

Data Wrangling

Two data sets will be used to perform exploration and analysis. One contains data about housing market in Melbourne and other contains information about the crime incidents in Victoria. Answering some questions require merging of two data.

Description of data sources:

A. Melbourne housing clearance data for 2016-2017. This data is in tabular form with 34857 rows × 22 columns

- `Unnamed: 0` `int64` – **drop column of no use**
- `Suburb` `object`
- `Address` `object (address of property)`
- `Rooms` `int64 (number of bedrooms)`
- `Type` `object (type h-house,u-unit,t-townhouse)`
- `Price` `float64 (selling price)`
- `Method` `object (method of selling)`
- `SellerG` `object (sellers name)`
- `Date` `object (date sold)`
- `Bedroom2` `int64`
- `Distance` `float64 (dist. from CBD)`
- `Postcode` `float64`
- `Bathroom` `float64 (number of bathrooms)`
- `Car` `float64 (number of parking)`
- `Landsize` `float64`
- `BuildingArea` `float64`
- `YearBuilt` `float64`
- `CouncilArea` `object (local government area)`
- `Lattitude` `float64`
- `Longtitude` `float64`
- `Regionname` `object`
- `Propertycount` `float6` – **drop column of no use**

B. Criminal offence recorded by police region and local government area, since 2010, counted and included in the data where it was reported to Victoria Police and first recorded in LEAP within the reference period. This data is available as tableau sheet which is used to get the data of size 34857 rows × 22 columns

- `Local Government Area` `object`
- `Year` `int64`
- `Rate per 100000 population` `float64(rate of incidents reported)`

Data Transformation

The two data sets merged to form a dataset containing information about housing and crime rate in that area. Data sets will be merged on the year of selling and the council area of the property to get the crime rate in the area in that year. The council area names are in different format in both the file which needs to be made similar. In housing data, area names are

followed by a space and in crime data names are followed by “council area” which needs to be striped to perform merging.

Housing data : `['Yarra ','Moonee Valley ','Port Phillip ',…`

Crime data: `['Total', 'Council Area Yarriambiack', 'Council Area Yarra …`

This comparison was done in python using pandas. “council Area” replaced by blank space in text editor and stripping the space in python.

```
In [44]: import pandas as pd
         data = pd.read_csv("Melbourne_housing_FULL.csv")
         df = pd.read_csv('criminal_incident.csv')

         a = list(pd.unique(data[['CouncilArea']].values.ravel('K')))
         b = list(pd.unique(df[['Local Government Area']].values.ravel('K'))

In [ ]: data['CouncilArea'] = data['CouncilArea'].apply(lambda x: x.strip())
        df['Local Government Area'] = df['Local Government Area'].apply(lambda x: x.strip())
```

Fig 1: python coding for wrangling

For year in housing data, use the date column to extract year of selling and add the values to the new column “Year”.

```
data['Year'] = data['Date'].apply(lambda x: int(x[-4:]))
```

merging two data sets provide a tabular data of size 34857 rows × 21 columns

```
new_df = pd.merge(data, df,  how='left', left_on=['CouncilArea','Year'], right_on = ['Local Government Area','Year'])
```

Data Checking

For good results of exploration and visualisation we need quality data, which needs to be representable and reliable. There is no use of having millions of entries with null values.

```
Unnamed: 0          0
Suburb              0
Address             0
Rooms               0
Type                0
Price            7610
Method              0
SellerG             0
Date                0
Distance            1
Postcode            1
Bathroom         8226
Car              8728
Landsize        11810
BuildingArea    21115
YearBuilt       19306
CouncilArea         3
Lattitude        7976
Longtitude       7976
Regionname          3
Propertycount       3
Bedroom2            0
dtype: int64
```

data.isnull().sum()

Use python to check null values and see if there are any incompletes in the data. The table below shows that filtering out null values for Price will get rid of many null values for other variables too. Anyhow, Landsize, BuildingArea and YearBuilt will still have thousands of null values left which are tolerable, as these columns has not much in our visualisations to perform. Coding to drop rows with null values in Price column:

data = data[data['Price'].notna()]

From information about the data, showed Date is in string type which needs to be in Date type format.

Change data type:

data['Date'] = data['Date'].astype('datetime64[ns]')

```
Method          34857 non-null object
SellerG         34857 non-null object
Date            34857 non-null object
Distance        34856 non-null float64
```

There are two columns 'Room' and 'Bedroom2', both contains the information about the number of bedrooms. One should always remove the duplicity in data. One column should be removed if there is high correlation between the two columns. Deleting column 'Bedroom2':

del data['Bedroom2']

```
In [20]: data[['Rooms','Bedroom2']].corr()
```

Out[20]:

	Rooms	Bedroom2
Rooms	1.000000	0.823836
Bedroom2	0.823836	1.000000

On exploring the YearBuilt column in Tableau, it appears that all of the houses were built from year range 1820-2019. But there was one row with entry of 1196 in YearBuilt column. This could be a wrong entry, but we cannot be so sure as it is possible to have this old property. We have considered it as an outlier and removed the row manually.

The latitudes range from -90 to 90. On exploring the data, there were 5 rows found with wrong entries in latitudes and longitudes. The values for latitudes and longitude were in opposite columns.

```
data.loc[data['Lattitude']>90]
```

	Method	SellerG	Date	Distance	...	Car	Landsize	BuildingArea	YearBuilt	CouncilArea	Lattitude	Longtitude	Regionname	Propertycount
	SP	Biggin	17/06/2017	23.2	...	10.0	993.0	128.0	1966.0	Knox	145.25632	-37.84688	Eastern Metropolitan	5030.0
	S	Barry	22/07/2017	24.7	...	10.0	734.0	NaN	NaN	Greater Dandenong	145.21043	-37.96969	South-Eastern Metropolitan	10894.0
	S	RW	26/08/2017	12.0	...	10.0	1002.0	170.0	1985.0	Darebin	145.03086	-37.70671	Northern Metropolitan	21650.0
	PI	Douglas	6/01/2018	10.5	...	10.0	980.0	NaN	NaN	Brimbank	144.81960	-37.79207	Western Metropolitan	6763.0
	SP	Greg	9/12/2017	18.4	...	10.0	NaN	NaN	1988.0	Wyndham	144.69513	-37.85998	Western Metropolitan	13830.0

These values in latitude and longitude columns need to be swapped. Using pandas:

```
idx = data['Lattitude']>90
data.loc[idx,['Lattitude','Longtitude']] = data.loc[idx,['Longtitude','Lattitude']].values
```

Data Exploration

After performing wrangling, transformation, and cleansing on dataset, it is ready to use be used for exploration. Data exploration will help us answer some questions and have a better understanding about the market and housing prices using appropriate visualisations and statistical tests. Explore the data to find something interesting can be done using Tableau or R.

MOST EXPENSIVE AREA IN MELBOURNE

By Council Area

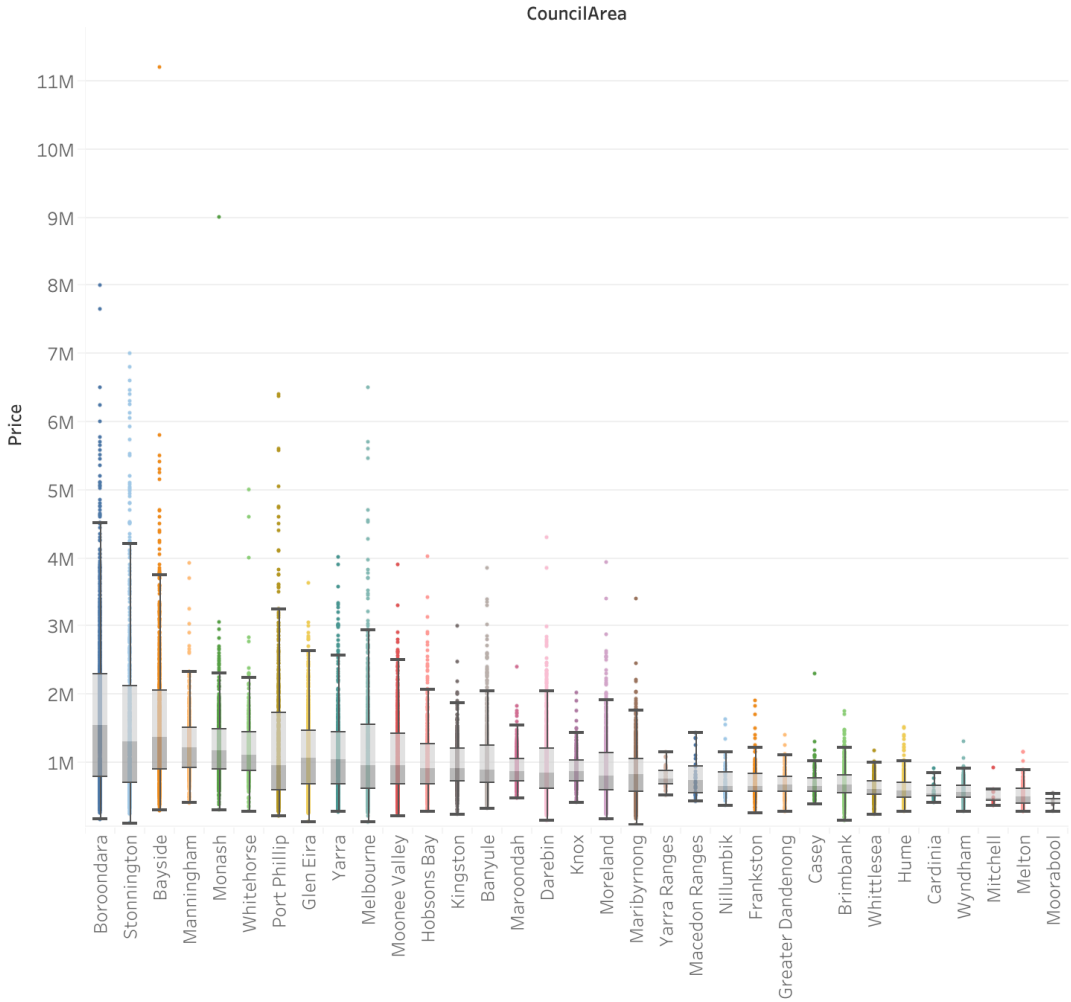


fig 2: Melbourne housing price by council

Created using Tableau, boxplot above shows a clear picture about the average housing price in each local government area. Boroondara is the most expensive council area with highest price range followed by Stonnington and Bayside. We can have clearer and deeper understanding by examining the prices by suburb. One can use choropleth map created in Tableau to get easy perception of comparison of average price by each suburb. The map below clearly shows Canterbury with darkest area to be the most expensive suburb in most expensive Council.

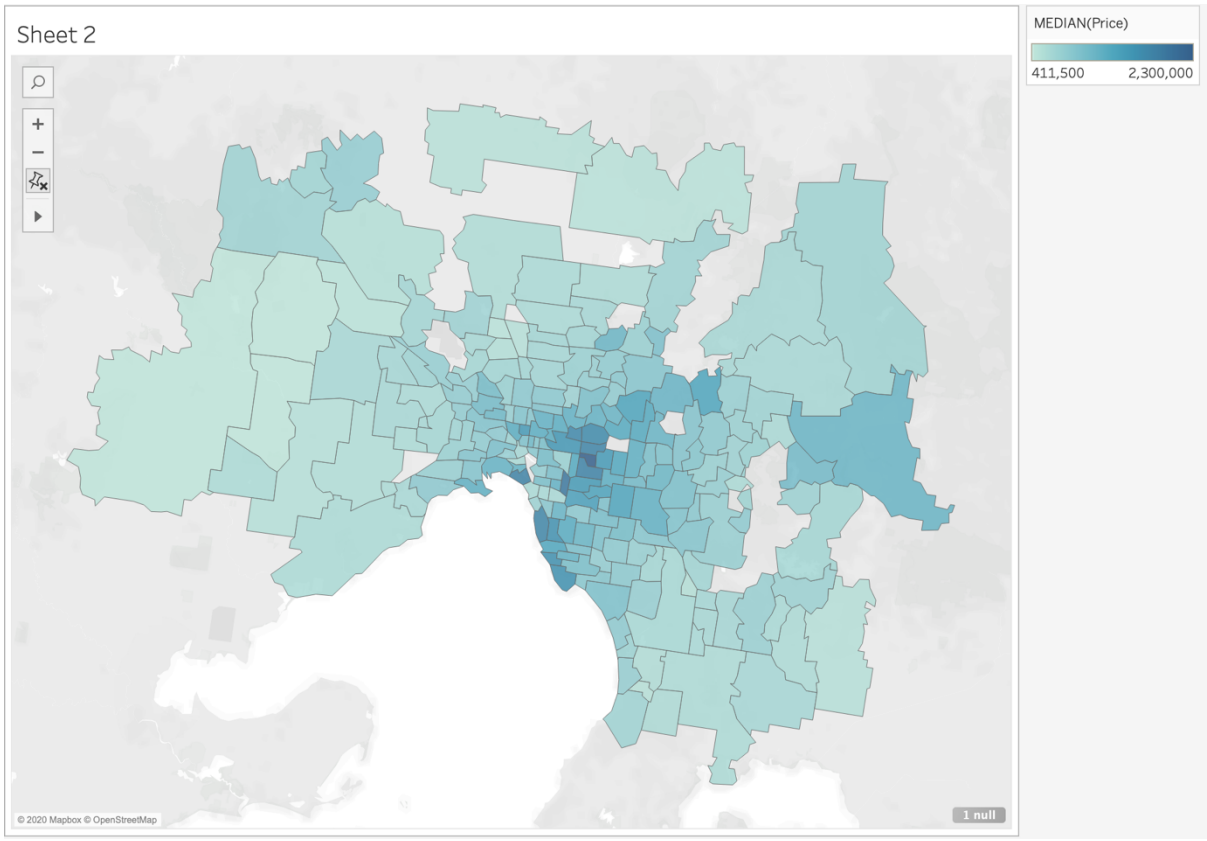


fig 3: average housing price by suburb

PRICE FLUCTUATION OVER TIME

Data contains information about housing from Jan 2016 to March 2018. Using Tableau, the line graph below shows fluctuation of average price over quarter each year.

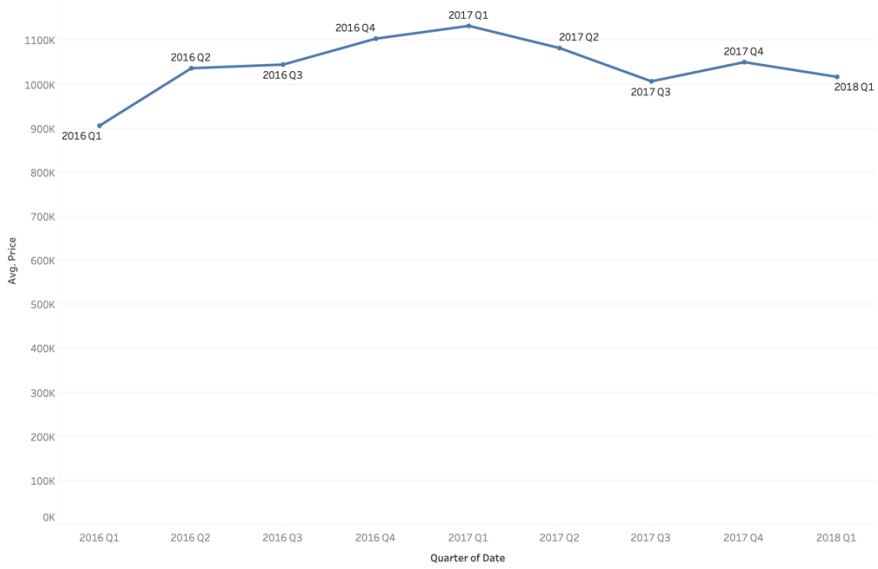


fig 4: average price by quarter of date

There was a gradual increase in prices by the end of year 2017 which took a downfall in year 2017. Melbourne housing cooled of and could be seen slowing of by the mid of 2017. No expectation of growth can be seen by the year 2019.

COMISSION THAT REAL ESTATE BROKER BRING IN

Since the dataset has no indication about the commission each of the listed agents received, a little research shows that the typical commission in Melbourne ranges from 1.5% - 3.0%. Assuming a commission of 2.0% can help us to get an idea of how much each agent brought in over these couple of years.

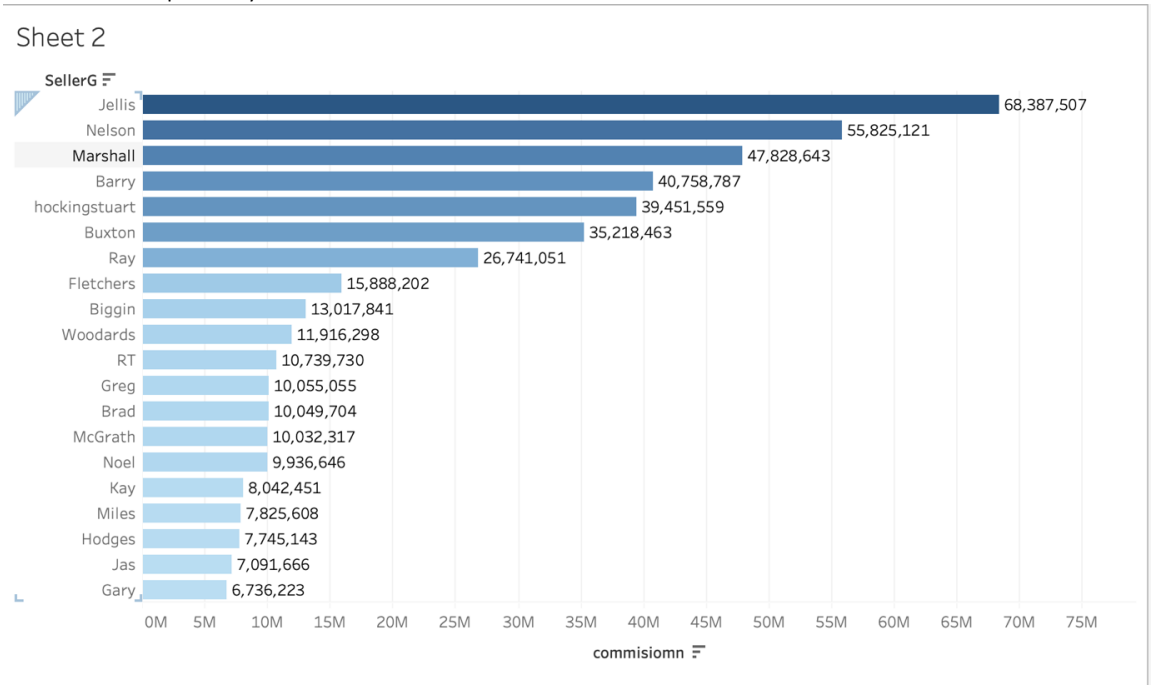


Fig 5: top 20 total commission by agents

Based on the bar plot above, it appears that some of these sellers are real-estate agencies with multiple agents working with them (e.g. Jellis, Hockingstuart). Even so, these top 20 agents and companies are making most of the commissions in Melbourne. Talking about only these top 20 sellers, they are comprising almost 81% of all of the commission found in the dataset and, for example, Jellis bringing in over 30,000,000 AUD in less than two years.

PRICE DIFFERENCE BETWEEN OLD AND NEWLY BUILD HOUSES

Categorising the houses as old if build before 1950 and newly build otherwise, line graph can be used to see the difference in average price in each council. It appears that there is not much difference in prices in most of the council areas except some near the CBD. These areas seem to have higher price rate for old houses than the new ones.

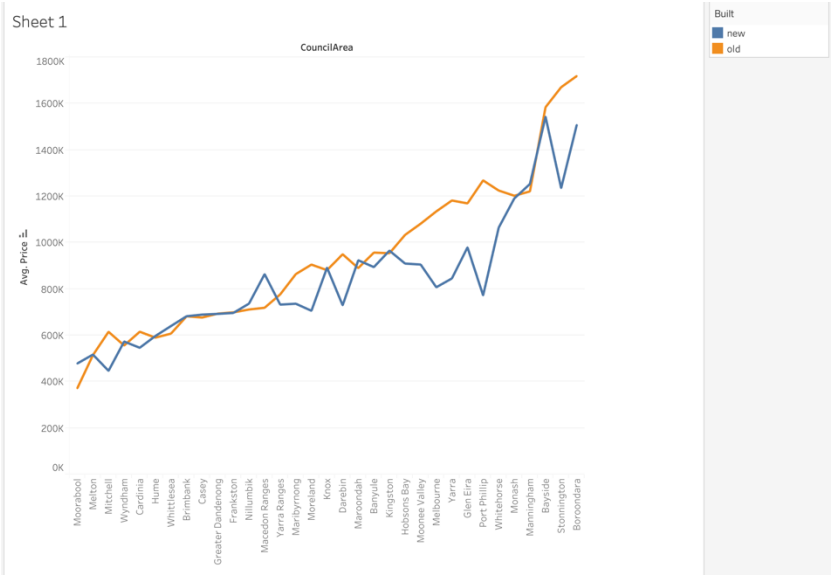
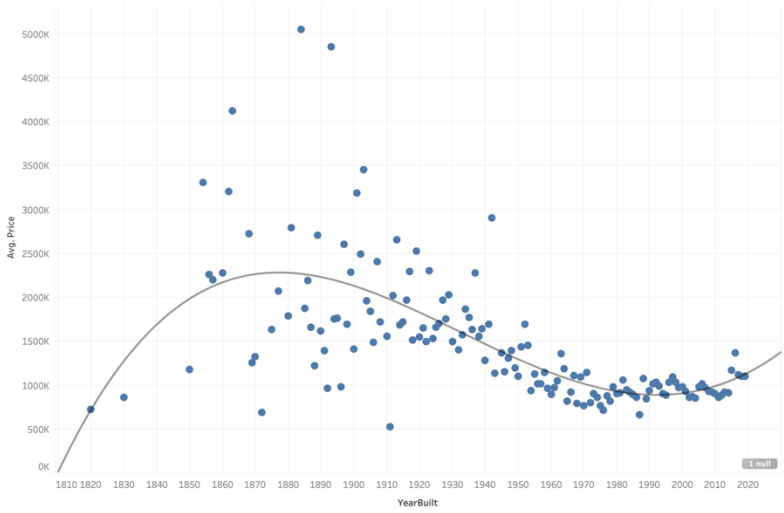


Fig 6: avg. price of old & new houses by council area



A negative correlation can be seen between year build and the average price from 1890 to 1990 which implies that the old houses have higher rate than the new ones. But the trend changes after 2000 with increase in the price with time. As newly built houses are expected to be expensive.

Fig 7: average price by year built

Further exploration resulted in a conclusion that most of the old houses are near the CBD where the house prices comparatively higher. The dot map below shows the darker colour in the centre with more old houses there.

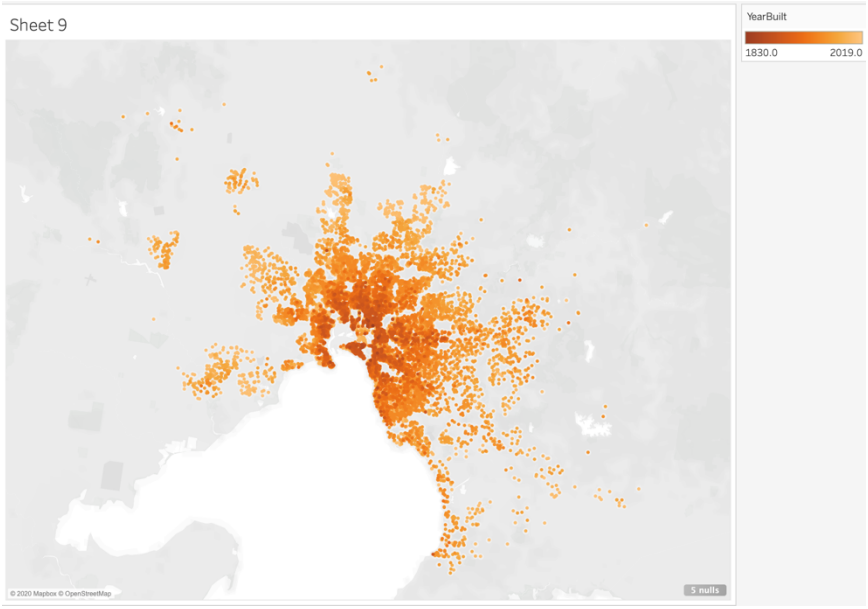


Fig 8: year build by location

PARKING SPACE AND HOUSING PRICE

In Tableau, exploring data using packed bubble plot showed that most of the houses have 1 or 2 parking spaces. Houses with 3,4 or no parking spaces are comparatively lower. We can see a positive correlation between the number of parking space available and the price of the house. The line graph depicts the relation between them. The price range increase with increase in the number of parking spaces. Parking space dose have an effect on the price estimation and cannot be ignored.

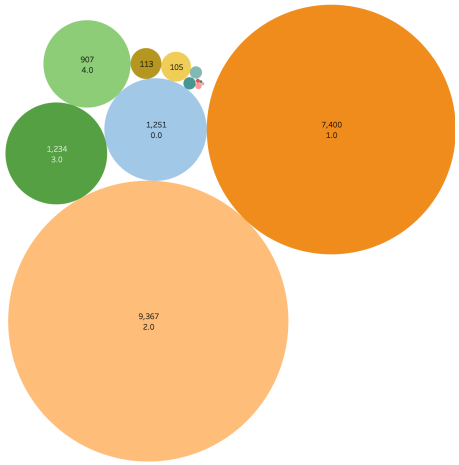


Fig 9: coun of houses by parking space

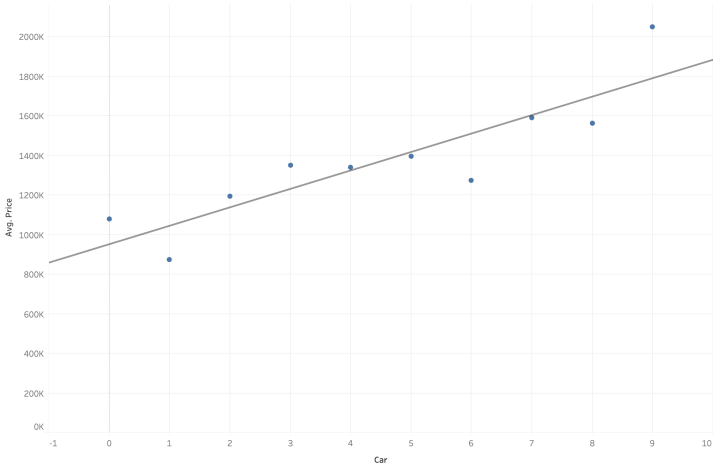


fig 10: average price by parking space

EFFECT OF CRIME INCIDENT ON HOUSING PRICE IN THAT AREA

In Tableau, dot map can be used to explore the relation between the crime rate and the housing price in that area. In the plot below, the size of the bubble depicts the average housing price (higher the price larger the size) and the colour concentration shows the offence reported rate per 100000 population (darker the colour, higher is the offence rate) in that council area. From the figure, it appears that the larger circles indicating higher price are lighter in colour indicating lower rate of offence showing negative correlation between the two.

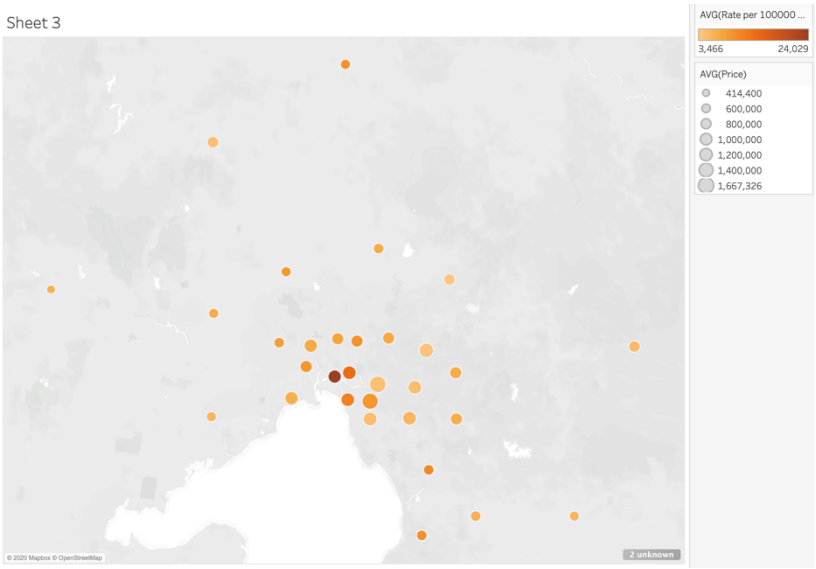


Fig 11a: average price vs incident rate by council area

The linear regression trend line (for year 2016) from Tableau clearly depicts the negative correlation between the incident rate and the housing price in that area. Though the relation is not very strong, so the crime rate affects the price but cannot be considered as an important factor with high effects.

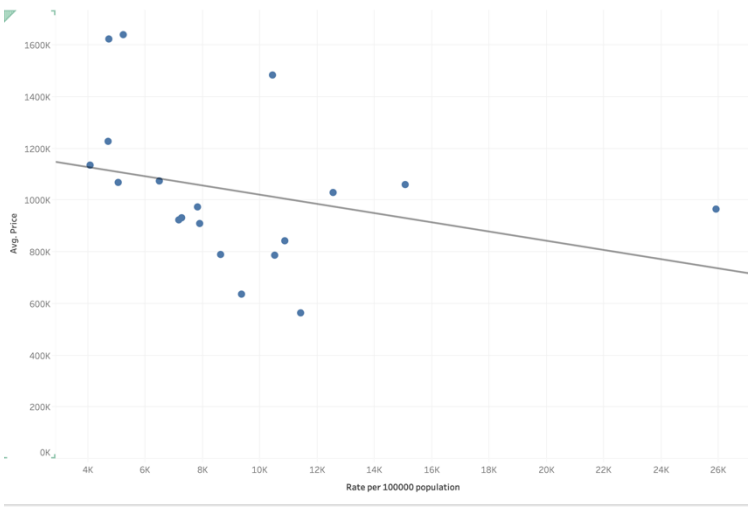


Fig 11b: average price vs incident rate by council area for year 2016

Conclusion

After successfully answering all the questions, we have a plenty of knowledge about the housing market to make decisions. Housing has a long-term role to play in investment portfolios, which needs a clear understanding. The data has enabled us to find some really interesting and important insights about Melbourne housing market. Suburb is an important predictive to find if the property is expensive. Southern and eastern suburbs seem to have more expensive properties. Further, crime incident rate has not much to do with the housing price in that area.
CBD, the central build up area of Melbourne seems to have the oldest buildings and the farther surrounding areas are often surprisingly new, with many buildings less than 20 years old. Old buildings are more expensive than the new once. Also, there is more trading of properties in and around the city centre and suburbs nearby, whereas there is a definite shift towards outer suburbs for less sales of properties.

Reflection

One of the he main key learning from this project is choosing the right representations to create an output that is relevant for solving a specific challenge or understand data. It is important to have proper knowledge of the grammar of graphics that govern the choice of chart. Some data transformation before actually using the dataset for exploration can make the process smoother and easier to answer questions. Preforming data checking upon the dataset can help finding errors or outliers if there are any. It is important to deal with the errors or outliers as they can have huge impact on the decisions made by visualisation.

Bibliography

- https://www.crimestatistics.vic.gov.au/explore-crime-by-location
- https://www.kaggle.com/anthonypino/melbourne-housing-market#Melbourne_housing_FULL.csv
- https://community.tableau.com/thread/265732
- https://www.youtube.com/watch?v=OQ3qBunXgp4
- https://community.tableau.com/thread/286895