1. Problem Statement

The Australian real estate market is thriving right now, and every individual is fascinated by investing properties. However, there are over 15,000 suburbs in Australia and which suburbs we should choose to invest in is a real problem that everyone will encounter. Therefore, this project focuses on helping investors choose more suitable suburbs for their tailored investment from a rational perspective, which is based on facts and data.

2. Data source

The URL is as follow: https://www.domain.com.au/?mode=sold

Domain Group encompasses a portfolio of brands that support it to be a leading property marketplace in Australia for consumers, agents and organisations with an interest in the Australian property market.

Domain's purpose is to inspire confidence for all of life's property decisions and they deliver to this by offering a suite of products and solutions for every step of the property journey. Through their digital and print solutions they reach an audience of over 9.6 million Australians in a month.

At the beginning of the project, Python request.get() Function was used to get the content from the above URL and then write into a CSV file.

The data was crawled once for all and it is for personal use only, and has been deleted after this analysis.

Data profiling

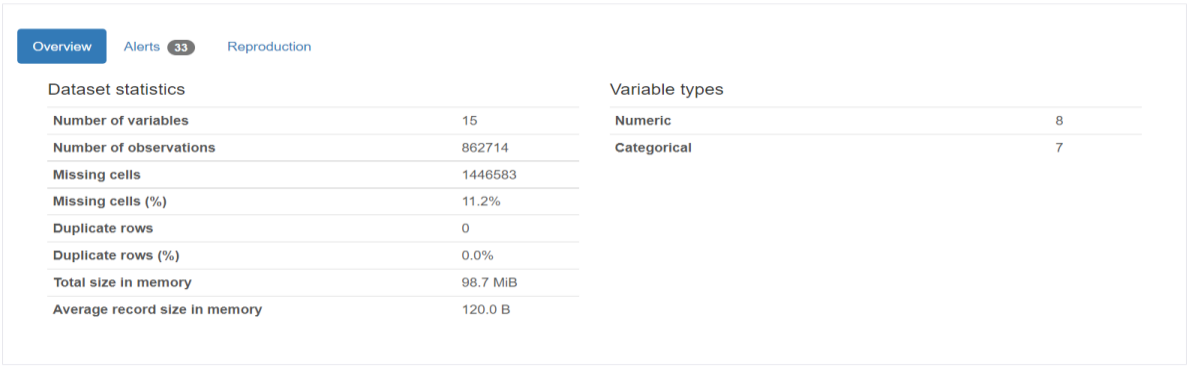
Pandas profiling is an open source Python module which generates interactive reports in HTML format and gives us an insight of the overall image of the csv file.

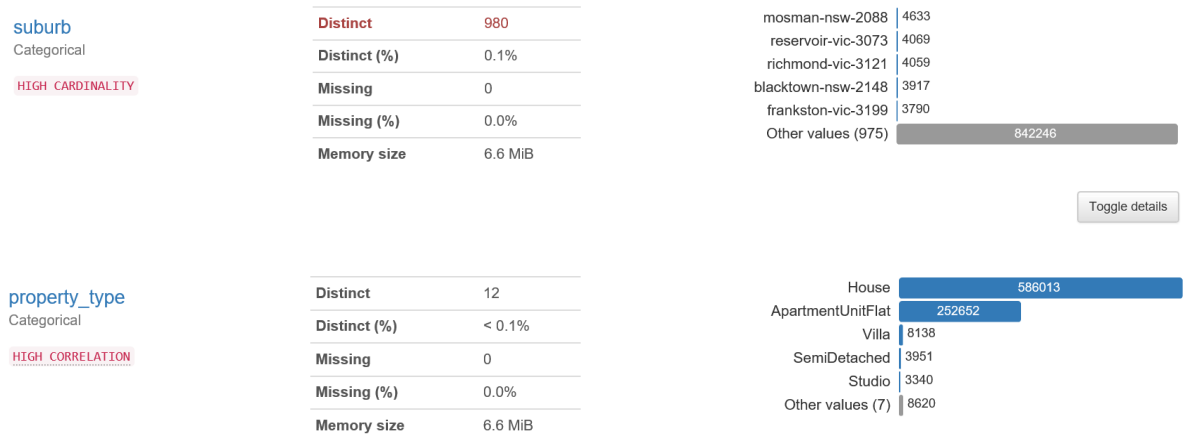
```
import  pandas as pd
import pandas_profiling

df = pd.read_csv('properties_cleaned.csv')
pfr = pandas_profiling.ProfileReport(df)
pfr.to_file(output_file ='C:\\Users\\zt583\\Desktop\\dataset\\report.html')
```

Report: file:report.html

Overview





The data covers the information of sold-properties from 2000 to July 2021. The dataset has 862714 rows and 15 columns.

The features of the dataset are shown as follow:

Field Name	total records	mismatch/null
listing_id	862714	0
suburb	(unique)980	0
sold_date	862463	251
timeframe		26/04/2000 - 24/07/2021
sold_type		
Auction	285741	0
private treaty	573601	0
blank	3372	0
property_type	(unique)12	0
is_rural	0	862714
price	$807,899,110,204.00	440(under 50000)
beds	(unique)20	3716
baths	(unique)20	723
parking	(unique)20	74732
landsize	384005	478709
address_lat	861761	953
address_lng	861760	954
address_street	838883	23831

The description of numeric features is shown as follow:

data description

	listing_id	is_rural	price	beds	baths	parking	land_size	address_lat	address_lng	rn
count	862714	2997	862714	859069	862007	788019	384005	861761	861760	862714
mean	not applied	not applied	936462.27	3.09	1.7	1.82	709.83	not applied	not applied	1.0
std	not applied	not applied	1753469.92	1.67	0.8	1.78	3570.29	not applied	not applied	0.0
min	not applied	not applied	1.0	1.0	1.0	1.0	1.0	not applied	not applied	1.0
25%	not applied	not applied	503000.0	2.0	1.0	1.0	451.0	not applied	not applied	1.0
50%	not applied	not applied	705000.0	3.0	2.0	2.0	613.0	not applied	not applied	1.0
75%	not applied	not applied	1075000.0	4.0	2.0	2.0	752.0	not applied	not applied	1.0
max	not applied	not applied	1010000000.0	1112.0	48.0	1289.0	1877743.0	not applied	not applied	1.0
is_null	0	859717	0	3645	707	74695	478709	953	954	0

From the data description, we notice that there are many missing values across the dataset. The **"is_rural"** and the **"land_size**"features have the most missing values. Thus, these two features are not considered in the analysis.

Then use SQL to do the further filtering to get the total unique suburbs per year, total sales per year, and total revenue per year.

```
CREATE TABLE step_1_filtering AS

WITH get_year AS( SELECT * , substr(sold_date, 8) AS yearFROM properties_cleaned)

SELECT year , COUNT (DISTINCT suburb) AS 'How many unique suburbs are involved each year'

                , COUNT (listing_id) AS 'How many properties are sold each year'

                , SUM (price) AS 'Total price of each year' FROM get_year WHERE year Between '2000' AND '2021'

GROUP BY year

ORDER BY year
```

year	How many (unique)suburbs involed each year	How many properties are sold each year	Total price of each year
2000	8	11	6643000
2001	14	46	23996000
2002	27	104	66164211
2003	30	156	137539575
2004	36	133	104013500
2005	137	721	468914414
2006	178	1091	811797902
2007	310	1979	1816467846
2008	626	4229	3019073007
2009	885	28810	20286543727
2010	900	34713	27398907617
2011	904	39830	29039956488
2012	910	44673	32772255335
2013	929	65276	50513980523
2014	946	79728	67207800528
2015	949	90260	85687093940
2016	948	85599	85346382777
2017	950	84814	89075799567
2018	949	79724	81419700313
2019	951	85451	86295182494
2020	952	83806	91360481885
2021	929	51309	54914801139

From the above table we can clearly see that **Domain.com.au** was expanding around 2009, which involved 885 suburbs and still increases from then on. On the other hand, since 2021 is still ongoing and the data ends at July 24, 2021, which cannot represent a whole year.

In the meanwhile, removed all the null and mismatching data to get a more precisely result.

3. Assumptions

Our analysis is based on the following assumptions:

Assumption 1

The data scrapped from the Domain website, which can represent the overall trend of the housing market in Australia.

Assumption 2

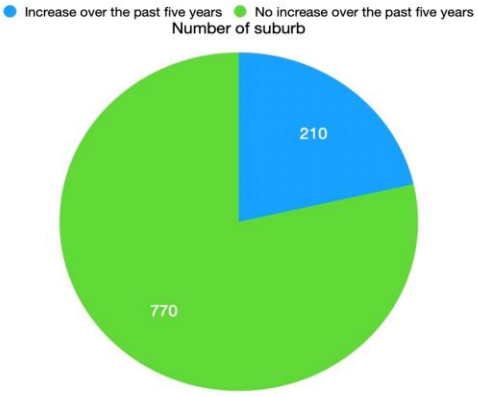
There are many irrational prices for some sold properties in our data. According to our data description above, the minimum price of sold property is $1. I assume that this problem was caused by some technical issues, and the real transaction prices were not input properly. Thus, properties with irrational price are not included in our analysis.

4. Persona

Investors who would like to buy houses with a budget of 1,000,000 AUD. Aged between 35-45, and prefer 3 beds-House located 15-25km to CBD with the median price growth in the recent 1,3,5 year.

5. Feature extractions

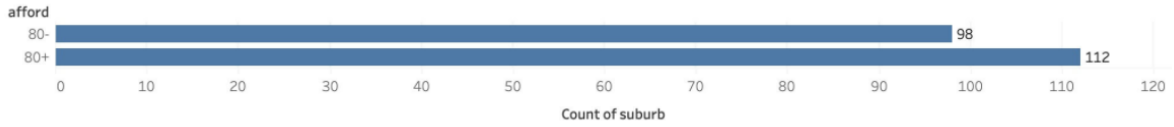
(1) Among all the suburbs 980 in total, client would like to chose those mediean price with a growth trend in the recent 1, 3, 5 year. The figure is shown as follow, the blue one with 210 suburbs are those selected ones.




```
select *
,lag(price,1) over (PARTITION by suburb, new_type order by year) as pre1_price
,lag(price,3) over (PARTITION by suburb, new_type order by year) as pre3_price
,lag(price,5) over (PARTITION by suburb, new_type order by year) as pre5_price
from table5), table7 as( select suburb ,new_type ,affordability ,price
,median_price - pre1_price as pre1_diff
,median_price - pre3_price as pre3_diff
,median_price - pre5_price as pre5_diff
```

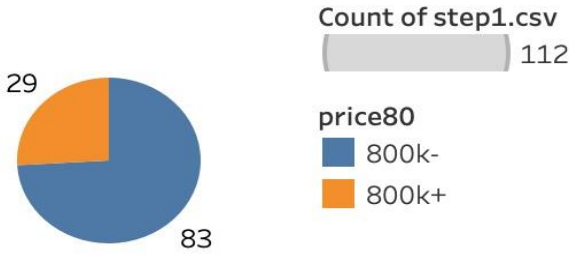
from table6 where year ='2021' and pre3_price>= pre5_price and pre1_price>= pre3_price and median_price >= pre1_price

(2) Among those suburbs with increasing mediean price (210 in total), chose those affordability over 80%. There are various way to define affordability. Within the context of this project, we measure affordability by the total number of affordable ones(price under 1,000,000 AUD) devide total amount of properties, and the result would be a percentage, which is the affordability of a suburb. As the figure shown, we will only focus on the 112 suburbs for the further analysis.



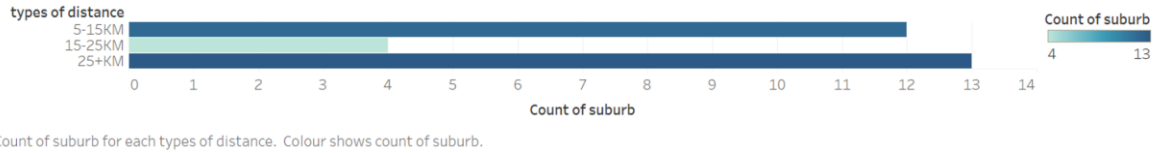
SELECT * FROM all WHERE affordability >= 80%

(3) By now, we have 112 suburbs left for our client. However, the budget of the client is 1,000,000 AUD, in this step, the price under 800,000 AUD which is the blue part will be regarded as irrelevant and then the orange one with 29 suburbs remain.



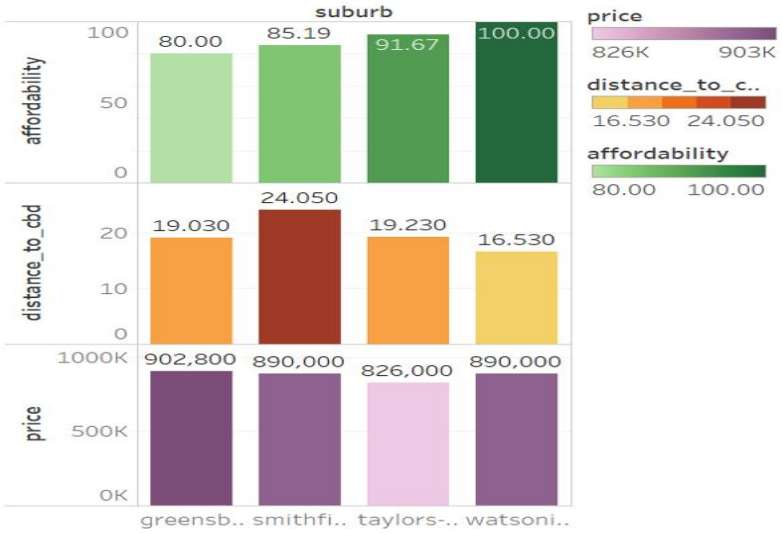
SELECT * FROM step1 WHERE price >= 800000

(4) Finally, the client mentioned that the best location would be near CBD but not living in it, thus the most suitable distance would be 15-25km far away from CBD.



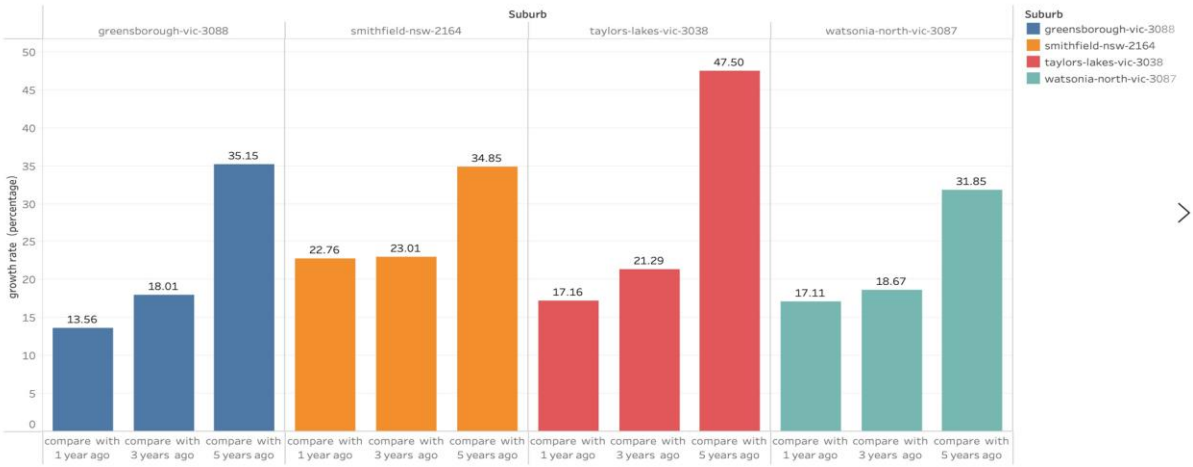
SELECT * FROM step2 WHERE distance_to_cbd BETWEEN 15 AND 25

(5) Among the above 29 suburbs, the following 4 suburbs meet all the criteria and affordability, distance to CBD, median price are shown separately in the figure.

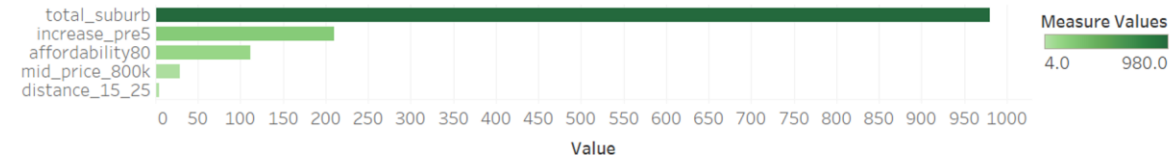


Affordability, sum of distance_to_cbd and sum of price for each suburb. For pane Affordability: Colour shows sum of affordability. For pane Sum of distance_to_cbd: Colour shows sum of distance_to_cbd. For pane Sum of price: Colour shows sum of price.

In the meanwhile, since the client is pretty interested in the median price growth in the recent 1, 3, 5 year. This figure gives a full image of the price changing of these selected 4 suburbs.



(6) Overall, this project focuses on helping investors choose more suitable suburbs for their tailored investment, therefore in order to meet all the criteria from our client, as the figure shown: it is a shrinking process of suburbs from the very beginning to the end.



Total_suburb, increase_pre5, affordability80, mid_price_800k and distance_15_25. Colour shows total_suburb, increase_pre5, affordability80, mid_price_800k and distance_15_25.

```
SELECT COUNT(total_suburb), COUNT(increase_pre5), COUNT(affordability80), COUNT(midPrice800K), COUNT(distance15_25) FROM all
```

The above content is a demo of how the project works, however, the main purpose is to help those people without Python and SQL skills to locate the suitable suburbs for them.

Thus, Pandas Module and Python For Loop were used to separate the type of the housing into different CSV files, which are useful tools for every individual, by using the filter function, they can easily get the suitable suburbs to do the investment.

6. Summary

To sum up the project process, the first step was to do the data ingestion from the above URL. The Python Request.get() function was used for the data collection and then write them into a CSV file.

The data generation process focused on data understanding, data cleansing and data mining. Pandas profiling Python module gives an insight of the overall distribution of the data. During this step, some redundant and irrelevant variables were excluded, and missing values were removed.

Then, a demo used persona was designed and figures are shown corresponce with the specific demands.

Finally, Python and SQL were used to implement a fully detailed spreadsheet, which can be a useful tool especially for non-technical users to choose suitable suburbs among the large amount thousands. Due to the various demands of individuals, this project is to help investors to locate their ideal suburbs.