Introduction:

A successful cafe owner in Melbourne (Australia) wants to open up another cafe. Instead of deciding based on instinct, she wants to have a data driven decision on where to open up the new cafe and seeks a scientific answer.

Since the current cafe is a success, and she has already learnt how to run such a business, she wants to find a similar suburb, and open up a similar cafe. She will also use the same furniture, same machinenery, other assets, which will lower her investment financially. On the other hand, this kind of a cafe, has become a business that she knows how to deal with the customer base.

Data:

Data used will be of two parts. First will be a Wikipedia page scraped for suburbs and corresponding postcodes. Using geo packages, i will add the latitude and longitude for each suburb. And then I will connect to the FourSquare API and pull the data for the venues for each coordinate.

And then i will pull Melbourne house prices via Kaggle API. After cleaning this dataset, i will use the mean of housing prices per suburb as a rough income distrubution guide to the dataframe holding venues.

Methodology

My idea for the advice required is

- to find suburbs similar to Brunswick with all other venues -will prove demand from a similar customer base-
- and suburbs similar to Brunswick with the housing prices. -this will prove the affordability of a new rental shop and again will prove demand from a similar customer segment-

For this, I have found a Wikipedia page that holds Melbourne suburbs and postcodes. (https://en.wikipedia.org/wiki/List_of_Melbourne_suburbs) And using the BeautifulSoup package, I have scraped the table. Also to be able to leverage the Foursquare API, I needed latitude and longitudes, so I have used GeoPy package for this. And pulled coordinates for each postcode.

	Suburb	Postcode	Latitude	Longitude
544	Woori Yallock	3139	NaN	NaN
545	Yarra Glen	3775	NaN	NaN
546	Yarra Junction	3797	NaN	NaN
547	Yellingbo	3139	NaN	NaN
548	Yering	3770	NaN	NaN

```
[8] geo_melb.head(2)
```

	Suburb	Postcode	Latitude	Longitude
0	Bellfield	3081	-37.7357	145.05
1	Briar Hill	3088	-37.7046	145.11

+

After getting the coordinates, I have connected to the FourSquare API and pulled nearby venues and venue types for each suburb in my dataframe and created a new dataframe.

```
[20] melb_venues.head()
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bellfield	-37.735707	145.049696	Latitude	-37.733320	145.049530	Recreation Center
1	Bellfield	-37.735707	145.049696	Waterdale Road Foodworks	-37.736362	145.048651	Convenience Store
2	Bellfield	-37.735707	145.049696	Waterdale Fish & Chips	-37.736390	145.048500	Fish & Chips Shop
3	Bellfield	-37.735707	145.049696	Fowl Play Charcoal Chicken	-37.737430	145.046907	Fried Chicken Joint
4	Briar Hill	-37.704640	145.109652	WaterMARC	-37.704519	145.104846	Swim School

After getting the nearby venues of each suburb, I have labelled and encoded these venue types and then grouped them on suburb and taken mean of each venue type per suburb, which gives me a frequency of occurrence. And my first dataset was ready.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Arcade
0	Abbotsford	-37.803555	144.995194	0.037037	0.0000	0.0	0.0	0.0	0.0	0.0	0.0
1	Aberfeldie	-37.749436	144.919553	0.000000	0.0000	0.0	0.0	0.0	0.0	0.0	0.0
2	Aintree	-34.873065	-56.267905	0.000000	0.0000	0.0	0.0	0.0	0.0	0.0	0.0
3	Airport West	-37.723395	144.857632	0.000000	0.0000	0.0	0.0	0.0	0.0	0.0	0.0
4	Albanvale	-37.740779	144.798206	0.000000	0.0000	0.0	0.0	0.0	0.0	0.0	0.0
5	Albert Park	-37.845448	144.958343	0.000000	0.0000	0.0	0.0	0.0	0.0	0.0	0.0
6	Albion	-37.784963	144.827122	0.000000	0.0625	0.0	0.0	0.0	0.0	0.0	0.0
7	Alphington	-37.778057	145.017671	0.000000	0.0000	0.0	0.0	0.0	0.0	0.0	0.0
8	Altona	-37.863230	144.828730	0.000000	0.0000	0.0	0.0	0.0	0.0	0.0	0.0
9	Altona Meadows	-37.870456	144.763801	0.000000	0.0000	0.0	0.0	0.0	0.0	0.0	0.0

	Neighborhood	Argentinian Restaurant	Art Gallery	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Australian Restaurant	Austrian Restaurant	Auto Workshop	BBQ Joint	Badminton Court	Bagel Shop	Bakery	Bar	Baske
0	Abbotsford	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.037037	0.0	0.0
1	Aberfeldie	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
2	Aintree	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
3	Airport West	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
4	Albanvale	0.0	0.0	0.0	0.111111	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.111111	0.0	0.0
5	Albert Park	0.0	0.0	0.0	0.000000	0.222222	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.
6	Albion	0.0	0.0	0.0	0.062500	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.062500	0.0	0.0
7	Alphington	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
8	Altona	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
9	Altona Meadows	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0

For the second dataset, I've found the Melbourne house prices dataset on Kaggle datasets. (anthonypino/melbourne-housing-market) Luckily, this dataframe contained suburb names and postcodes as well. So I've cleaned up the dataframe. Filled the missing price of houses by mean price of the houses with that many rooms in that suburb. For this, I have created a temporary dataset grouped by Rooms and Suburbs with mean prices and used it as a dimension table for missing values.

```
house_prices.head()
```

	Suburb	Address	Rooms	Type	Price	Method	SellerG	Date	Postcode	Regionname	Propertycount	Distance	CouncilArea
0	Abbotsford	49 Lithgow St	3	h	1490000.0	S	Jellis	1/04/2017	3067	Northern Metropolitan	4019	3.0	Yarra City Council
1	Abbotsford	59A Turner St	3	h	1220000.0	S	Marshall	1/04/2017	3067	Northern Metropolitan	4019	3.0	Yarra City Council
2	Abbotsford	119B Yarra St	3	h	1420000.0	S	Nelson	1/04/2017	3067	Northern Metropolitan	4019	3.0	Yarra City Council
3	Aberfeldie	68 Vida St	3	h	1515000.0	S	Barry	1/04/2017	3040	Western Metropolitan	1543	7.5	Moonee Valley City Council
4	Airport West	92 Clydesdale Rd	2	h	670000.0	S	Nelson	1/04/2017	3042	Western Metropolitan	3464	10.4	Moonee Valley City Council

```
avg_prices = df_nona.groupby('
avg_prices.head()
```

Postcode	Price
3000	7.731029e+05
3002	1.409505e+06
3003	1.041723e+06
3006	6.520588e+05
3008	6.954444e+05

After the dataset had no more missing prices, I've binned the prices into 5 categories, low, mid-low, mid, mid-high and high. Ideas was to use it not for the ML model but to use it to find the suburbs in the same price bracket at the end.

	Price Bracket	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Arcade	Argentinian Restaurant	Art Gallery	Arts & Crafts Store	Asian Restaurant
0	Mid-Low	Abbotsford	-37.803555	144.995194	0.037037	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
1	Mid	Aberfeldie	-37.749436	144.919553	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
3	Low	Airport West	-37.723395	144.857632	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
4	Low	Albanvale	-37.740779	144.798206	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.111111
5	High	Albert Park	-37.845448	144.958343	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000

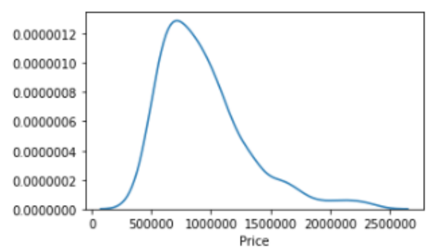
When the house prices table was set, I wanted to check if the prices are distributed normally.

```
[110] # want to how the data is distributed before fitting to machine learning

      import matplotlib.pyplot as plt
      import seaborn as sns
```

```
[112] plt.figure(figsize = (5,3))

      sns.distplot(final_df['Price'],hist=False)
      plt.show()
```



Also, wanted to check if there's any easy to see correlation between house prices and total number of venues in the suburb. Since more venues would mean more population and this would creat a natural inflation.

```
[132] # will see if any easy-to-see correlation exist between price and number of venues
      venue_count = melb_venues.groupby('Neighborhood')[['Venue']].count()
      venue_count.rename({'Venue':'Count'},axis=1,inplace=True)

      temp_df = pd.merge(left = venue_count, right = final_df[['Neighborhood', 'Price']], on='Neighborho
      temp_df.head()
```

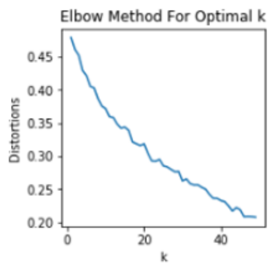
	Neighborhood	Count	Price
0	Abbotsford	27	1.065107e+06
1	Aberfeldie	8	1.351314e+06
2	Airport West	2	7.760442e+05
3	Albanvale	9	5.468430e+05
4	Albert Park	9	2.054282e+06

```
[133] temp_df.corr()
```

	Count	Price
Count	1.000000	0.245037
Price	0.245037	1.000000

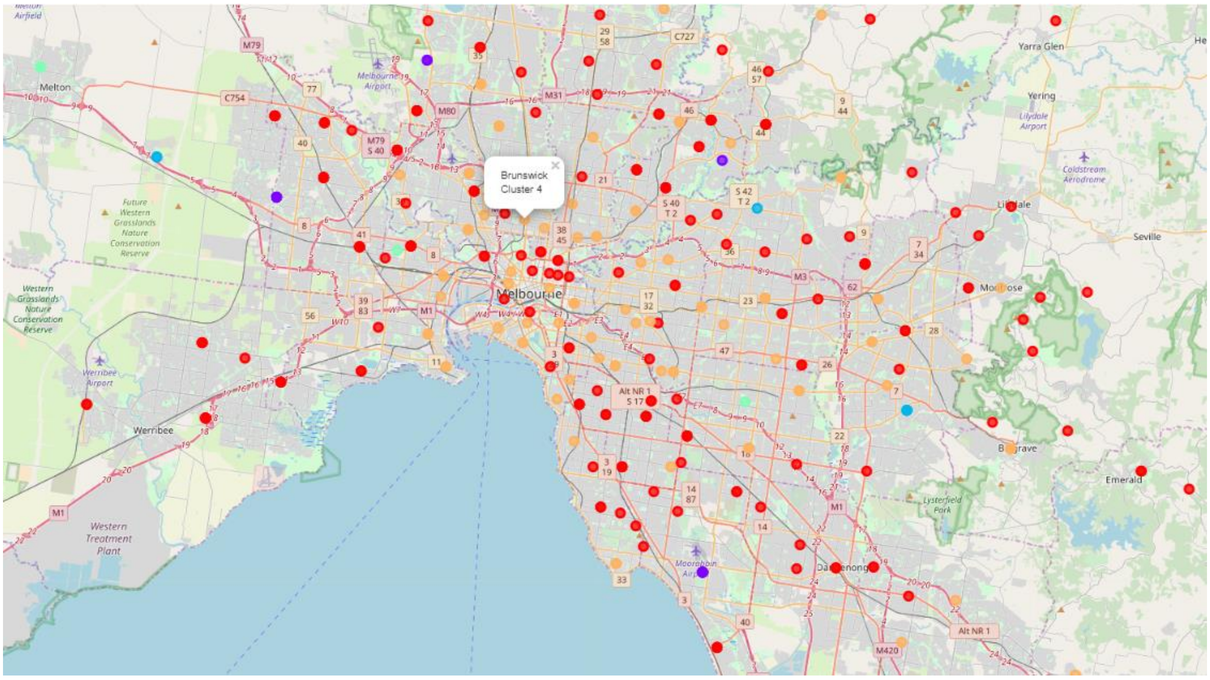
But the correlation between the number of venues in a suburb and the pricing is very low.

After I made sure, dataset was all good to be trained in a ML model, I've used a elbow method to find a optimal number of clusters. But the methodology was very ambigious for this model. And since I have around 350 observations, I've decided to go with 5 as the cluster number. The same number of price bracket bins. Also did not want to overfit to this small dataset.



After fitting a K Means Cluster model with 5 clusters, I've inserted the labels to the dataframe and visualised the clusters on a map

	Cluster Labels	Price Bracket	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Servic
0	0	Mid-Low	Abbotsford	-37.803555	144.995194	0.037037	0.0	0.0	0.(
1	0	Mid	Aberfeldie	-37.749436	144.919553	0.000000	0.0	0.0	0.(
3	0	Low	Airport West	-37.723395	144.857632	0.000000	0.0	0.0	0.(
4	0	Low	Albanvale	-37.740779	144.798206	0.000000	0.0	0.0	0.(
5	4	High	Albert Park	-37.845448	144.958343	0.000000	0.0	0.0	0.(



Since the advice required was a new suburb to open up a café, which was already successful In Brunswick, Ive filtered to see Brunswick's price bracket and cluster label. After that, ive sliced all dataframe for these values and ended up with around 40 suburbs. Then I went back and regrouped the venues dataframe with a count value for the cafes per suburb.

```
[166] # Cafe is in Brunswick, so will see the cluster and price bracket for housing
      final_df[final_df['Neighborhood'] == 'Brunswick']
```

	Cluster Labels	Price Bracket	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Arcade	Argentinian Restaurant	A Galle
63	4	Mid-Low	Brunswick	-37.767725	144.959508	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

1 rows × 276 columns

```
[169] # will slice Similar suburbs to Brunswick
      advice_df = final_df[(final_df['Cluster Labels'] == 4) & (final_df['Price Bracket'] == 'Mid-Low')]
      advice_df.shape
```

```
(40, 276)
```

```
[189] venue_type_count = melb_venues.groupby(['Neighborhood','Venue Category'])[['Venue']].count()
      venue_type_count.reset_index(inplace=True)
      cafe_count = venue_type_count[venue_type_count['Venue Category']=='Café']
      cafe_count.drop(labels = 'Venue Category',axis=1, inplace =True)
      cafe_count.rename({'Venue':'Cafe Count'},axis =1, inplace =True)
      cafe_count.reset_index(drop=True,inplace=True)
      cafe_count.head(1)
```

Results

And when merging that café counting dataframe to the advice dataframe (that naturally looks like Brunswick), I have just sorted out the 10 suburbs with least number of cafes around.

```
[191] advice_counted = pd.merge(left = advice_df[['Neighborhood']], right = cafe_count, on='Neighborhood')
```

```
[193] advice_counted.sort_values(by = 'Cafe Count').head(10)
```

	Neighborhood	Cafe Count
36	Wantirna South	1
2	Bentleigh East	1
35	Wantirna	1
16	Keilor East	1
23	North Warrandyte	1
8	Burwood East	1
17	Knoxfield	1
19	Montmorency	2
15	Heathmont	2
24	Notting Hill	2

Discussion:

I have scraped data from websites, merged them and grouped a city’s suburbs using machine learning algorithm based on the venues in the suburb. As above, my advice to this owner of the Brunswick café would be opening up a new branch in one of ‘Wantirna South’, ‘Bentleigh East’, ‘Wantirna’, “Keilor East’, ‘North Warrandyte’, ‘ Burwoord East’ or ‘Knoxfield’.

Results:

There are plenty of cafés, and competition among them in Melbourne. But using scientific methods, the odds would be better for anyone who is about to invest in a venue.