MELBOURNE SUBURBS ANALYSIS

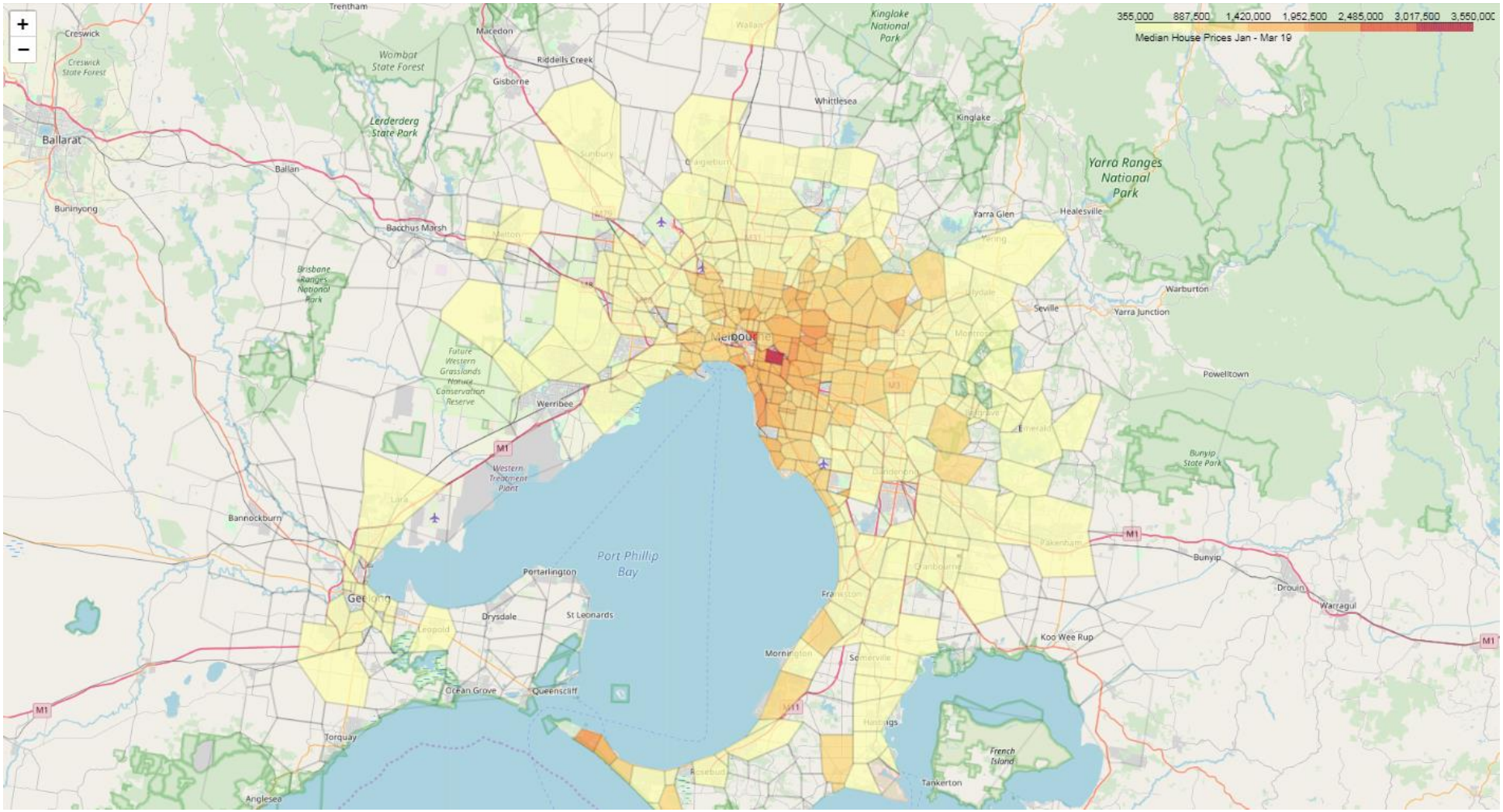
Analysing suburbs in Melbourne will be useful for those looking to move to Melbourne

- The business problem we are trying to solve, is "Where should someone unfamiliar with Melbourne buy a house in metropolitan Melbourne?".
- This question is relevant for large groups of people looking to purchase property in Melbourne.
- There are number of elements to this problem, for example:
    - What are average property prices in different suburbs?
    - What amenities are available in different suburbs?
    - What are different suburbs "like" in terms of character?
    - What are some of the drivers of property prices?
    - Each of these questions will be explored in this analysis

We have drawn on a number of data sources

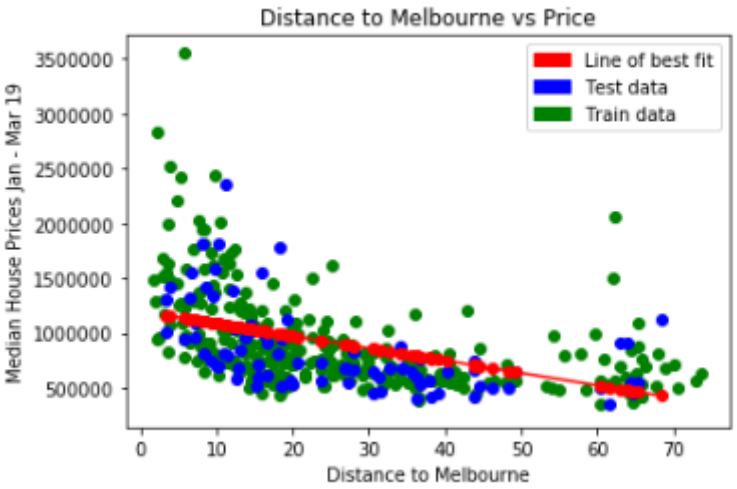
- **Australian postcode location data**: to plot different suburbs on a map of Melbourne
    - Data source: http://www.corra.com.au/australian-postcode-location-data/
- **Postcode remoteness data**: to narrow down our data to metropolitan Melbourne only and remove regional and remote areas
    - Data source: https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1270.0.55.005July%202016?OpenDocument
- **Suburb geojson files:** to create a choropleth map of Melbourne
    - Data source: https://github.com/tonywr71/GeoJson-Data/blob/master/australian-suburbs.geojson
- **House price data by postcode:** to append average house prices to add to our analysis
    - Data source: https://discover.data.vic.gov.au/dataset/victorian-property-sales-report-median-house-by-suburb
- **Foursquare API data:** to cluster suburbs with similar characters

House prices seem to be related to proximity to the city….



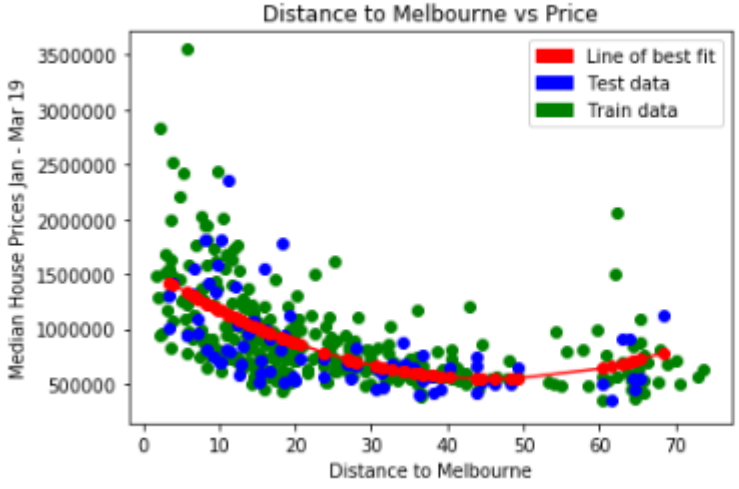
…however regression results are not strong…

Linear Regression




```
Train Data:
Coefficients:  [[-11269.45031909]]
Intercept:  [1200970.43359469]
Mean absolute error: 276091.47
Residual sum of squares (MSE): 149105072301.40
R2-score (train): 0.22

Test Data:
Mean absolute error: 269783.74
Residual sum of squares (MSE): 117531255091.75
R2-score (test): 0.22
Text(0.5, 1.0, 'Distance to Melbourne vs Price')
```

Polynomial Regression (Degree 2)



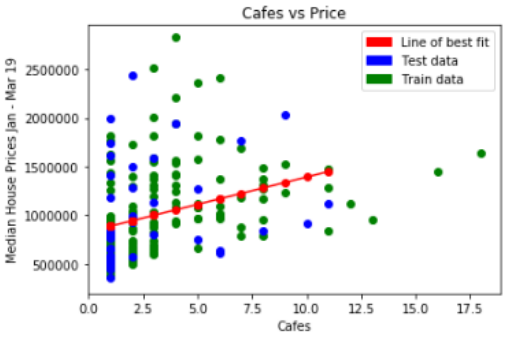

```
Train Data:
Coefficients:  [[     0.         -44459.61780876    481.74478014]]
Intercept:  [1572817.25302619]
Mean absolute error: 244237.25
Residual sum of squares (MSE): 122418372226.05
R2-score (train): 0.36

Test Data:
Mean absolute error: 227879.11
Residual sum of squares (MSE): 94867144840.90
R2-score (test): 0.37
Text(0.5, 1.0, 'Distance to Melbourne vs Price')
```

…and there is also no clear relationship between some of the top venues in Melbourne and house price either

Linear Regression

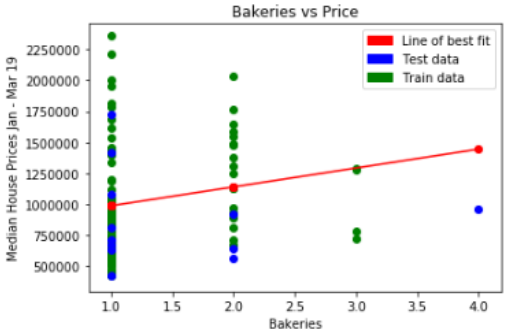
Venue Category	Venue Count
Café	553
Pizza Place	154
Supermarket	135
Grocery Store	134
Bakery	123
Fast Food Restaurant	120
Thai Restaurant	93
Coffee Shop	87
Train Station	84
Pub	79
Sandwich Place	78
Shopping Mall	74
Bar	73
Park	72
Italian Restaurant	72

Linear Regression (Cafes vs Prices)



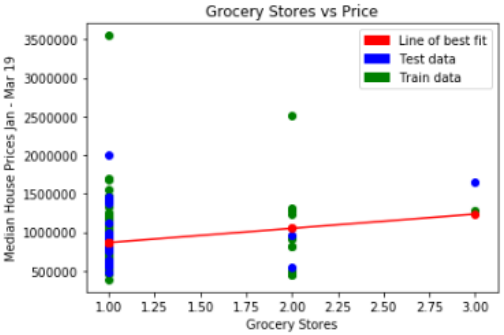

```
Train Data:
Coefficients:  [[56216.47558003]]
Intercept:  [830614.35406699]
Mean absolute error: 312249.09
Residual sum of squares (MSE): 179163293377.03
R2-score (train): 0.14

Test Data:
Mean absolute error: 424317.92
Residual sum of squares (MSE): 273998018724.98
R2-score (test): 0.00
Text(0.5, 1.0, 'Cafes vs Price')
```

Linear Regression (Bakeries vs Prices)



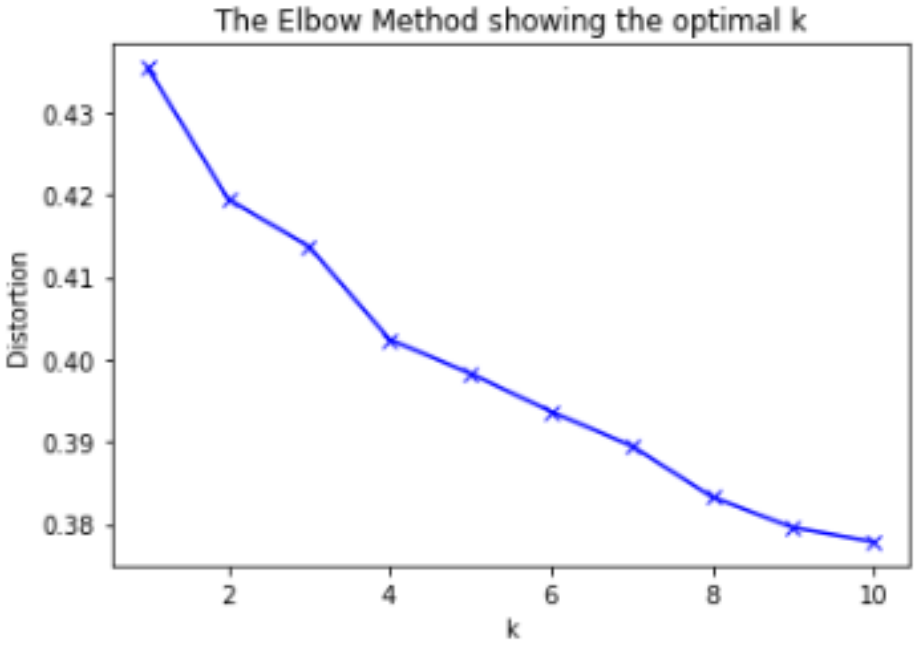

```
Train Data:
Coefficients:  [[152496.3681592]]
Intercept:  [834527.6119403]
Mean absolute error: 367676.98
Residual sum of squares (MSE): 208640804286.73
R2-score (train): 0.03

Test Data:
Mean absolute error: 394610.17
Residual sum of squares (MSE): 188421013593.18
R2-score (test): -0.45
Text(0.5, 1.0, 'Bakeries vs Price')
```

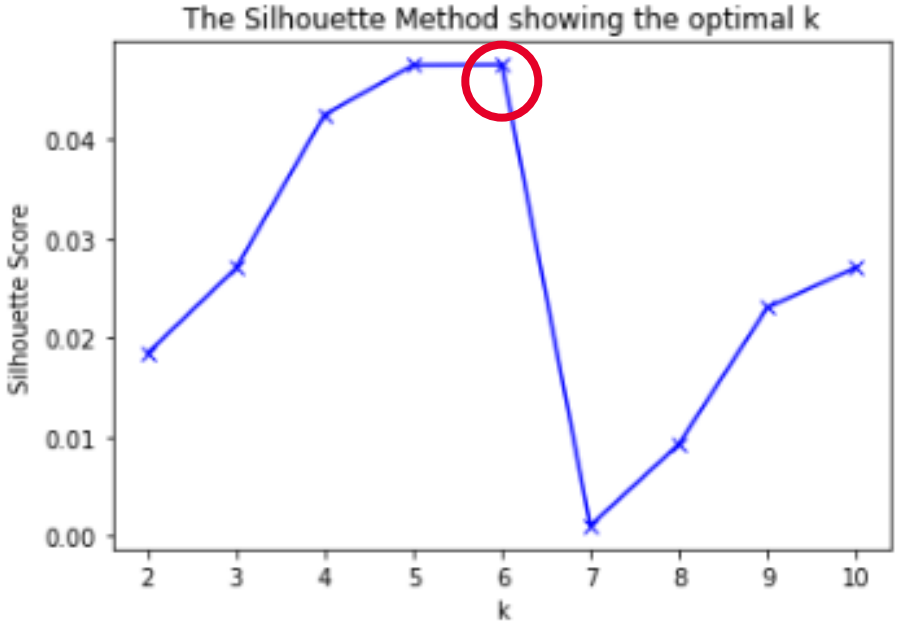
Linear Regression (Grocery Stores vs Prices)



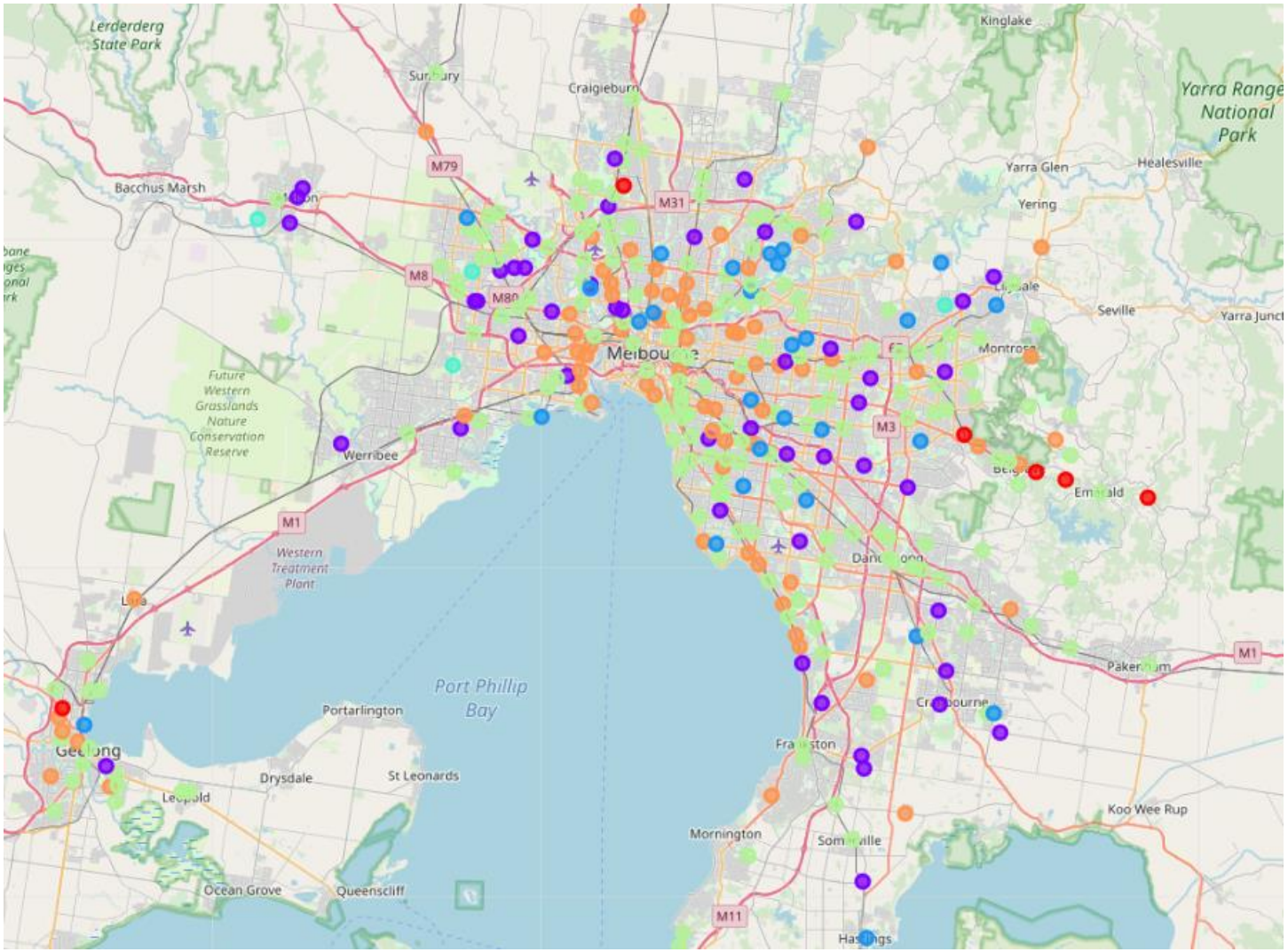

```
Train Data:
Coefficients:  [[184533.8762215]]
Intercept:  [681841.69381107]
Mean absolute error: 286416.68
Residual sum of squares (MSE): 194110764962.85
R2-score (train): 0.03

Test Data:
Mean absolute error: 316678.68
Residual sum of squares (MSE): 158697796484.30
R2-score (test): -0.02
Text(0.5, 1.0, 'Grocery Stores vs Price')
```

We conducted k-means clustering of suburbs using k = 6…





…and identified 6 distinct clusters which are shown below



- Cluster 0
- Cluster 1
- Cluster 2
- Cluster 3
- Cluster 4
- Cluster 5

Based on top venues in each cluster, as well as proximity to the city, we have named the clusters

Cluster	0	1	2	3	4	5
1st Most Common Venue	Pizza Place	Park	Furniture / Home Store	Café	Café	Grocery Store
2nd Most Common Venue	Café	Café	NaN	Supermarket	Pizza Place	Train Station
3rd Most Common Venue	Grocery Store	Tennis Court	NaN	Fast Food Restaurant	Grocery Store	Bakery
4th Most Common Venue	Bakery	Zoo Exhibit	NaN	Bakery	Bakery	Burger Joint
5th Most Common Venue	Fast Food Restaurant	Construction & Landscaping	NaN	Grocery Store	Train Station	Coffee Shop
6th Most Common Venue	Fish & Chips Shop	Train Station	NaN	Coffee Shop	Supermarket	Doctor's Office
7th Most Common Venue	Thai Restaurant	Light Rail Station	NaN	Thai Restaurant	Pub	Memorial Site
Distance to Melbourne	24.768535	23.572191	26.113992	25.111721	21.26592	39.515993

Cluster	0	1	2	3	4	5
Cluster Names	Pizza and Food	Parks and Nature	Other	Cafes and Shopping	Accessible Inner Suburbs	Accessible Outer Suburbs

This analysis provides a starting point for further work on suburbs and house prices in Melbourne

- This analysis provides useful information for someone considering purchasing a property in Melbourne
- Data allows for exploration of the types of venues in each suburb, as well as details on house prices
- A number of further analysis could be undertaken which may provide useful insight
    - Time-series analysis of suburb price over time (this is available in the data)
    - Polynomial regression of multiple venue types and suburb house prices, or even simple linear regression of the total number of venues in the suburb
    - Clustering based on other suburb characteristics, such as location, crime rate, etc.
    - Removing Geelong data and re-running existing analysis may prove useful, as this likely distorted some findings, particularly the regression analyses.