Categorizing Melbourne, Australia Suburbs for Potential Real Estate Buyers

Wilhem Kornhauser

May 14, 2020

# Introduction

A challenge many people may come across at some point in their lives is the requirement to decided upon a place of residence. A quick Google search may bring up results for "Top 10 cities to live in today", but each city is as equally diverse within itself as cities are diverse with regard to other cities.

The aim of this project is to use Foursquare data along with crime data and potentially real estate data to identify and categorize different neighborhoods which meet the different wants and needs of a diverse set of people.

Once this model has been developed, it can be used with new data to quickly put a new real estate listing into one of these categories, and then with targeted advertising, can be quickly shared with potential buyers who would be interested in that type of neighborhood.

For this problem, the city of Melbourne, Australia was chosen. Melbourne is a popular city, frequently quoted as being "most livable" in Australia, and ranking high globally. Given it’s popularity, I decided that it would be an interesting location to try and model.

# Data Acquisition and Cleaning

Listed below is the data which was gathered for this project, along with a brief explanation as to the importance/relevance of these features.

1. Location data: A set of postal codes for Melbourne, which will be required to query data from Foursquare.
2. Foursquare location data: Through the iterative process which was followed, this data was eventually drawn as the top 100 venues for each neighborhood. This include stop picks, trending locations, food, coffee, nightlife, fun, shopping, breakfast spots, public schools and high schools, hospitals, universities and colleges, parks, groceries, libraries, and beaches, as well as many more. The idea is to select as many non-redundant features as possible which are comprised of interests for people from all perspectives.
3. Crime data: Foursquare was unable to provide this kind of data, so some research would be required for this. The reason why I wish to include this is because low crime neighborhoods are obviously more desirable, and high crime neighborhoods may reflect characteristics of the real estate data. Because of this, the inclusion of this data is important. I was able to find burglary rates for many neighborhoods from the following location: <https://www.racv.com.au/in-your-home/in-your-home/burglary-statistics.html>. This data was saved to a csv file which can be found here: <https://github.com/Wkornhauser/Coursera_Capstone/blob/master/crimes.csv>. In the future, a more comprehensive set of crime data, which includes more than just simple burglary rates should be incorporated, however this single feature will have to serve the intended purpose for now.
4. Real estate data: This data can be used to segment the different neighborhoods based on incomes and wealth, and help to create a model which can be used to give potential home buyers options which are realistic and achievable. The median house prices for many neighborhoods was found with the following link: <https://www.propertyvalue.com.au/>.

Many steps were taken during the data collection to be able to clean the data. When first gathering the location data, by visual inspection it was found that some of the neighborhood name strings returned latitude and longitude coordinates that were clearly not within Australia. By sorting the Latitude column of this data in descending order, it was easy to quickly identify four observations where this was the case, and then remove these from the data.

The next step had me getting postal codes for each neighborhood name, which was to be required in querying the real estate data. Here, the webscraping returned neighborhood strings where the each string was preceded and followed by a “\n”, and also delimited by a “\n”. To clean this, I leveraged the ability to index strings to select the 2nd to 2nd last elements, and then replace instances of “\n” to “, “.

I also noted that there were some instances where postal codes could not be returned, or a single neighborhood would span several postal codes. In instances where no postal code was returned, the observation was dropped. Where multiple postal codes were returned, the first postal code was arbitrarily assigned to that neighborhood.

Once the postal code data was gathered, I iterated through the data to create strings which became unique url’s for web scraping of real estate data, based on each postal code. This returned string values for median house prices, in either the form of $X.Xm or $XXXk, representing the price in millions or thousands respectively. Using conditional statements, I iterated through the returned data, removing the unwanted $, m, and k characters, and returning float type, numeric values instead.

To get the crime data, I tried numerous websites to find either data sets I could use, or good references for web scrapping. Unfortunately, I was unable to do this, and had to by hand enter the burglary rates for each neighborhood I had by hand. This took a little extra time, but it was good to be able to include this data, as it was a key feature in my opinion.

From here, the data was all acquired, cleaned, and ready for analysis.

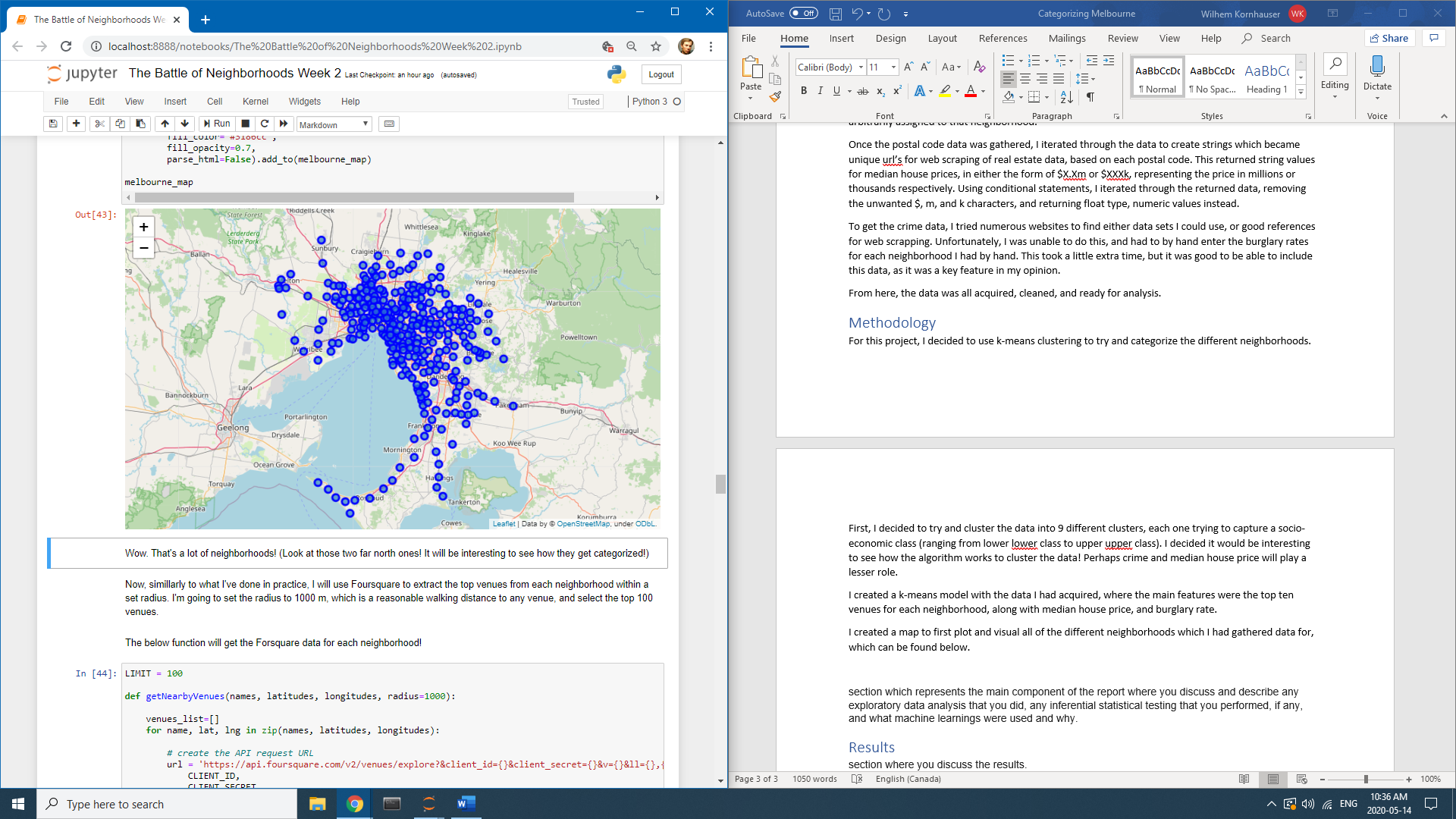
# Methodology

For this project, I decided to use k-means clustering to try and categorize the different neighborhoods.

First, I decided to try and cluster the data into 9 different clusters, each one trying to capture a socio-economic class (ranging from lower lower class to upper upper class). I decided it would be interesting to see how the algorithm works to cluster the data! Perhaps crime and median house price will play a lesser role.

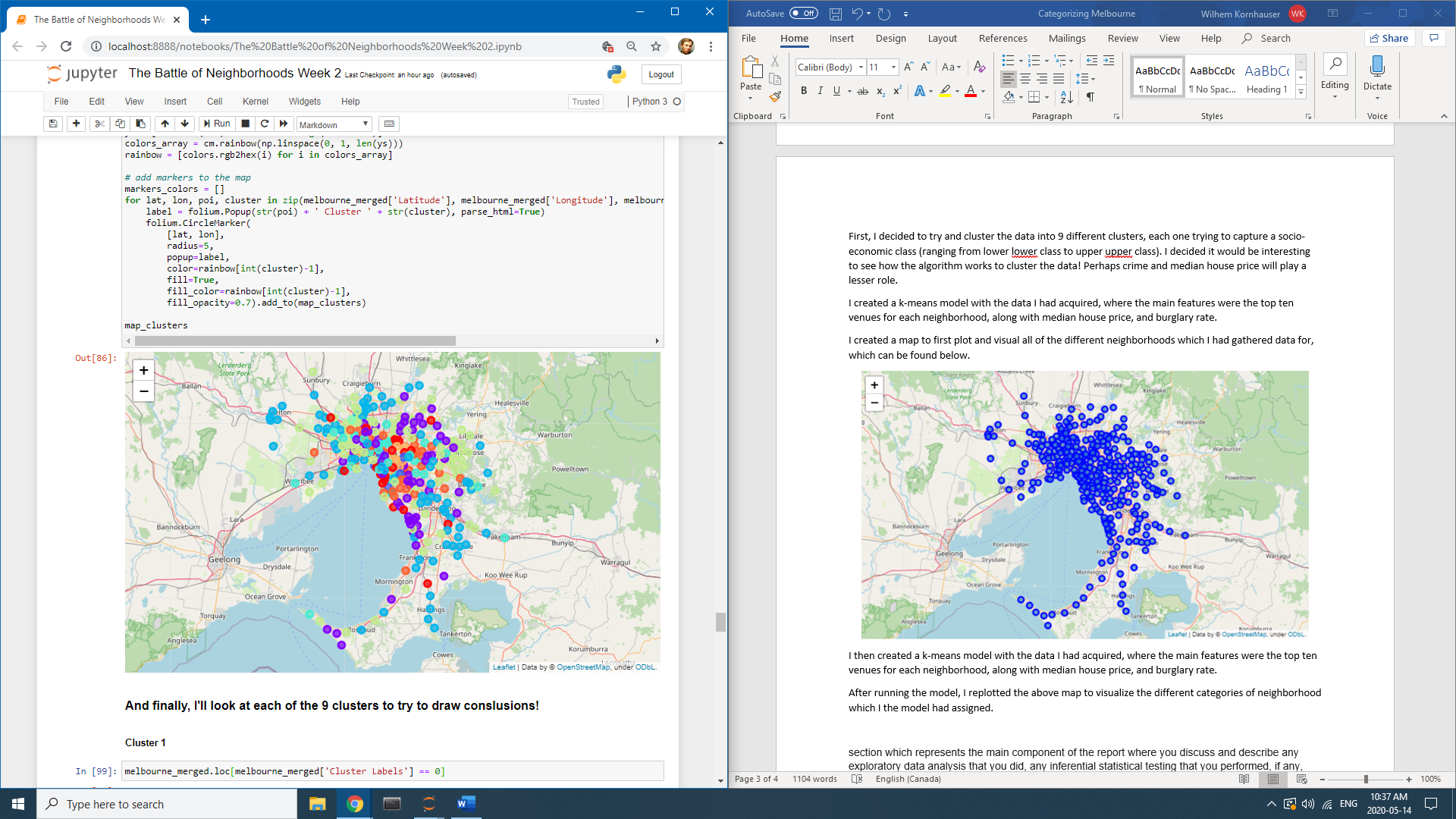
I created a k-means model with the data I had acquired, where the main features were the top ten venues for each neighborhood, along with median house price, and burglary rate.

I created a map to first plot and visual all of the different neighborhoods which I had gathered data for, which can be found below.



I then created a k-means model with the data I had acquired, where the main features were the top ten venues for each neighborhood, along with median house price, and burglary rate.

After running the model, I replotted the above map to visualize the different categories of neighborhood which I the model had assigned.



For this project, I did not perform any inferential statistical testing, as I felt it was an extra step, and given the fact that the data is unlabelled, and the training is unsupervised, I felt that it would not be relevant or possible to properly evaluate how well the model can assign a category to each neighborhood.

# Results

In addition to the above plot, each class of neighborhood was investigated to look for trends. This data is found in the Jupyter workbook. Given the large size of each respective data frame, it does not make sense to append this data into this report.

In the discussion section below, I will outline these results and implications.

# Discussion

First and foremost, when looking at the second map plot, with the color coded neighborhood data points, it is hard to draw any kind of conclusions about the categories of each neighborhood based on the visualization. Most points are well spread out, and seemingly randomly strewn across the Melbourne map. However, it can be noted that orange and red clusters tend to be more centralized, while the blue, light blue, green, and purple neighborhoods tend to be found throughout the entire plot.

When looking at the cluster 1 data print out (found in the Jupyter notebook), it quickly becomes apparent that the majority of neighborhoods in this cluster have the #1 most common venue being a café, and those which do not have it as 1st are likely to have it within their top 10 venues. There is also a very large disparity between the median house prices and crime rates in these neighborhoods, so it is likely that these features are less important when assigning a neighborhood to this class.

In cluster 2, again café is a very common top 10 feature, but in addition it was seen that many neighborhoods here also include parks and beaches within their top 10 venues. Again, the median house price and crime rates display a large variance, and are unlikely to have played a major role here.

Cluster 3 is small, containing only 5 neighborhoods…however it is quick to see that this cluster is for the more wealthy. The lowest median house price here is 1.7 million dollars, and ranges up to 3.4 million. This is the first instance where it appears that the clustering was done based on house prices. Hilariously, once again cafes were a #1 venue for 3 of the five neighborhoods, and within the top 5 venues for every neighborhood.

Cluster 4 seems to be the first cluster where cafes do not dominate as the #1 venue of the top 10 venues. These houses also tend to be lower on the median price, with only a few exceeding $1 million, and many falling within $500,000 to $600,000. Once again, the crime rates here display a high variance, however a potential home buyer could quickly query this list and find a neighborhood which is both affordable, and lower on the crime scale.

Cluster 5 is another king of the cafes type of cluster. However, most house prices here exceed $1 million, and crime rates are again highly varied.

Cluster 6 was so unique that but only one neighborhood was selected for it! This is clearly an outlier, however upon a quick inspection of the top venues, median house price, and crime rate, it is difficult to accurately state what cause this type of cluster to be populated with just the one neighborhood.

Cluster 7 is like many other clusters with high levels of cafes within the top #1 venue, and a highly varied crime rate and median house price.

Cluster 8 is similar to cluster 3, representing a smaller number of neighborhoods with very high median house prices, a large number of cafes within the top #1 venue spot, and highly varied burglary rate.

And finally, cluster 9 is again displaying a high number of cafes as the top venue for the neighborhood, but median house prices here frequently exceed $1 million.

In review of all of this, there are several point to be discussed:

1. Cafes in Melbourne
2. Cluster 6
3. Ideal number of clusters
4. Crime rates and clusters
5. Crime rates and median house prices

Foremost, it is quite clear that cafes dominate in Melbourne as a top venue! No matter which neighborhood someone looks at, they are almost guaranteed to be close to a café. Perhaps observation will be changed by reducing the radius from 1km to 500m for the Foursquare data query, which could be explored in a future iteration of the model. The high presence of cafes within the clusters also makes it appear to produce a bias in the clustering, and this should be eliminated.

Cluster 6 is a very interesting case, as it is difficult to discern why the model decided that Tullamarine was so unique that it belonged to its own cluster, especially given that the data set was sufficiently large enough to have many neighborhoods within each cluster. Also, the neighborhood does not seem very unique, having café as the 4th top venue, a median price of $620,000, and crime rate of 1.01. This is similar enough to Broadmeadows in cluster 4 where café is 4th top venue, median house price is $526,000, and crime rate is 1.33. However, upon a closer inspection, it is noted that 8 of the 10 top venues are explicitly restaurants, so perhaps this is the “foodie” neighborhood cluster!

As noted, there are many clusters where there is a great deal of similarities, or no discerning attributes for which the cluster was made. For this reason, perhaps the number of clusters is not ideal, More clusters could be used to refine this, or, perhaps there are already too many clusters and the number should be reduced. In a future iteration, this is an important factor to consider!

It does not appear that the crime rate played a major role in any single cluster! This is not what I was anticipating. I would like to be able to have some level of clustering performed based on the crime data, and in the future it would be nice to iterate the model to achieve this. Perhaps, if the number of clusters were to change this feature would be more important. Alternatively, with a broader set of features representing crime data this may also be achieved.

Finally, an unexpected observation which I noted through this is that there does not seem to be a strong correlation between the crime rates and median house prices! I was fully anticipating seeing strong negative correlation between house prices and crime rates, however this was not the case! There appears to be a high amount of variance between house prices and crime rates. This can be further explored by plotting the median house price vs crime rate, and would be an interesting investigation in the future. Perhaps, with a better set of crime data a stronger correlation would be seen in the observations.

# Conclusion

In conclusion, this project has done an acceptable job of categorizing some of the neighborhoods of Melbourne, Australia for potential home buyers. However, it has done an amazing job of reinforcing concepts which I have learned over the last few weeks during this data science course. I had to draw upon all my new knowledge to be able to develop this project, which did wonders for cementing the knowledge. As I’ve said above, and initially learned in an earlier data science course, this is clearly an iterative process! On the first development of this, I have accomplished some of what I wanted to (being able to classify neighborhoods for potential buyers), however, it has also become clear that there are many tweaks to my model and data that I should experiment with to further refine the product, making it better at developing unique, intuitive classes! I also observed new instances of data which piqued my curiosity, and made me want to further explore and try to understand why certain relationships are seen (or more specifically, why no clear relationship between crime rates and median house prices are seen). Overall, this project has taken me one step further to becoming a competent data science, and laid the foundation for a model which can be used for interested buyers in Melbourne. As a final note, this model can be easily adapted to just about any city in the world, with some minor changes to code. This implies that, at a future date, if this model were to be refined to a point of high proficiency in classifying unique neighborhoods, then it could be used world wide for anyone interesting in moving to a new, unknown city.

# 