IBM Data Science Capstone

Opening a French Restaurant in Wellington, New Zealand

INTRODUCTION

Wellington is New Zealand's capital making it the world's southernmost capital of a sovereign state. It is the third-largest city by population in New Zealand. Wellington consists of the central historic town and certain additional areas within the Wellington metropolitan area, extending as far north as Linden and covering rural areas such as Mākara and Ohariu. It has an urban population of 215,400 over an urban area of 442 km^2.

Wellington is characterised by small dining establishments, and its café culture is internationally recognised, being known for its large number of coffeehouses. Restaurants offer cuisines including from Europe, Asia and Polynesia; for dishes that have a distinctly New Zealand style, there are lamb, pork and cervena (venison), salmon, crayfish (lobster), Bluff oysters, pāua (abalone), mussels, scallops, pipis and tuatua (both New Zealand shellfish); kumara (sweet potato); kiwifruit and tamarillo; and pavlova, the national dessert. (source: Wikipedia)

There is definitely a community of foodies in Wellington and a demand for European cuisine. French cuisine being the most famous and appreciated European cuisine, it is also highly regarded internationally for its supposed refinement and fanciness. French restaurant in non-French speaking countries are thus generally quite fancy and pricy.

Hence, to open a restaurant, one has to look at:

- Areas where people have a sufficient standard of living to frequent the restaurant regularly,
- Areas not overcrowded with restaurants, but still lively, having enough venues around,
- Areas where the housing price is not too expensive since one want to reduce the initial cost of the acquisition of the restaurant building thus reducing the risk of the investment.

DATA

The data that have been successfully collected and their use are listed below:

- **Wellington's Wards and Suburbs Dataset** scraped from the Wikipedia Page of Wellington (https://en.wikipedia.org/wiki/Wellington#Economy, table: *Official suburbs of Wellington City, New Zealand*). The coordinates of the Wellington City suburbs are then collected thanks to the Python library GeoPy.
- **Median house price per Suburb of the Wellington Region as of September 2019 Dataset** compiled manually into a csv file from the website https://blog.homes.co.nz/wellington-median-house-price-by-suburb/#https://blog.homes.co.nz/wellington-median-house-price-by-suburb/. This dataset will be use to find a suburb where the acquisition cost of a restaurant is appropriate.
- **Victimisations Numbers for Wellington City per Area Unit (7/1/2014 to 3/1/2020) Dataset** found on the New Zealand police website (https://www.police.govt.nz/about-us/publications-statistics/data-and-statistics/policedatanz /victimisation-time-and-place). This dataset permits to assess the insecurity level of each Area Unit as we do not want workers or customers of the restaurants to feel uneasy to go to the restaurant as night/dinner time is a prime time for restaurants as well as victimisations. Moreover, we want to avoid putting the restaurants into an area where it will be likely to get stolen or vandalised.
- **Census - Median household income per Area Unit 2001, 2006, 2013 Dataset** for all Area Unit of New Zealand. The data is provided by Stats NZ and can be found on the following website: https://figure.nz/table/SSkBckhaaTU3hRqA. This dataset is used to explore the areas where the households have a sufficient income and lifestyle to frequent a French restaurant regularly.
- **Area Unit 2013** provided by Stats NZ and available at https://datafinder.stats.govt.nz/layer/25743-area-unit-2013/. This is the definitive set of area unit boundaries for 2013 as defined by Statistics New Zealand as at 1 January 2013. The data defines the boundaries of all Area Unit of New Zealand. First, The Wellington City's Area Unit boundaries are extracted. Then, the data is converted to GeoJSON and used to create choropleth map.
- **Wellington Venues**, obtained using the **Foursquare API**. Foursquare API is used to get the common venues of the different areas, including the restaurants and their numbers.
- **Wellington French Restaurants**, obtained using the **Foursquare API**. Foursquare API is used to get the French Restaurants of the different areas.

Suburb vs Area unit

The datasets collected are either grouped by Area Unit or by Suburb. Wellington have 57 suburbs and is divided into 78 Area Unit. Some Suburbs contain several Area Units and some Area Units contain several Suburbs.

- Suburb are official denomination representing the different parts of Wellington City, each having their own center.
- Area units are aggregations of adjacent meshblocks with coterminous boundaries to form a single unbroken surface area (land and/or water). Exceptions to this rule are some area units comprising collections of geographically related inlets and marinas. In an urban location, an area unit is often a collection of city blocks, while in rural situations area units may be equated to localities or communities. Area units must either define or aggregate to define urban areas, rural centres, statistical areas, territorial authorities, and regional councils. Each area unit must be a single geographic entity

with a unique name. The area unit pattern is revised once every five years in the year immediately before a Census of Population and Dwellings. There may also be changes in other years, in conjunction with local body boundary changes.

The question is, do we want to open a French restaurant in the best Area unit or in the best Suburb? This is a question that is worth asking. Indeed, looking solely at the numbers, it seems that Area Unit correspond to a thinner decomposition of the city that could show patterns inside suburbs. However, this might lead to overfit our analysis and results. Suburbs might also make more sense since it does not depend on the specific geography of the area.

Bearing these consideration in mind, we choose to select the Suburb slicing of Wellington as it seems more appropriate for our problem.

Three Main DataFrames Collected (two secondary DataFrame exist for Choropleth Map) :

Master DataFrame with the coordinates and the relevant figures for each suburbs,

	Ward	Suburb	Latitude	Longitude	Median household income (2001)	Median household income (2006)	Median household income (2013)	AVG Median household income 2001-2013	MEDIAN HOUSE PRICE (NZD)	Number of Victimisations
0	Lambton Ward	Aro valley	-41.295328	174.766580	34900.0	41900.0	62300.0	46366.0	769875.0	1026.0
1	Southern Ward	Berhampore	-41.323264	174.774090	36250.0	47400.0	63350.0	48999.5	769112.0	367.5
2	Eastern Ward	Breaker bay	-41.334316	174.827818	77200.0	100000.0	136300.0	104500.0	987443.0	158.0
3	Onslow-Western Ward	Broadmeadows	-41.233961	174.796556	69000.0	81400.0	100500.0	83633.0	717413.0	249.0
4	Southern Ward	Brooklyn	-41.306574	174.762354	68100.0	82800.0	107800.0	86233.0	853524.0	295.0

Table 1. Master DataFrame

Venues DataFrame with the number of venues and indicator variables for each venue categories,

	Suburb	Nb Venues	Airport	Airport Terminal	American Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	...	Turkish Restaurant	Vegetarian / Vegan Restaurant	Video Store	Vietnamese Restaurant	Wat
0	Aro valley	62.0	0	0	0	1	0	0	0	1	...	0	1	1	0	
1	Berhampore	6.0	0	0	0	0	0	0	0	0	...	0	0	0	0	
2	Breaker bay	5.0	0	0	0	0	0	0	0	0	...	0	0	0	0	
3	Broadmeadows	1.0	0	0	0	0	0	0	0	0	...	0	0	0	0	
4	Brooklyn	11.0	0	0	0	0	0	0	0	0	...	0	0	1	0	

Table 2. Venues DataFrame

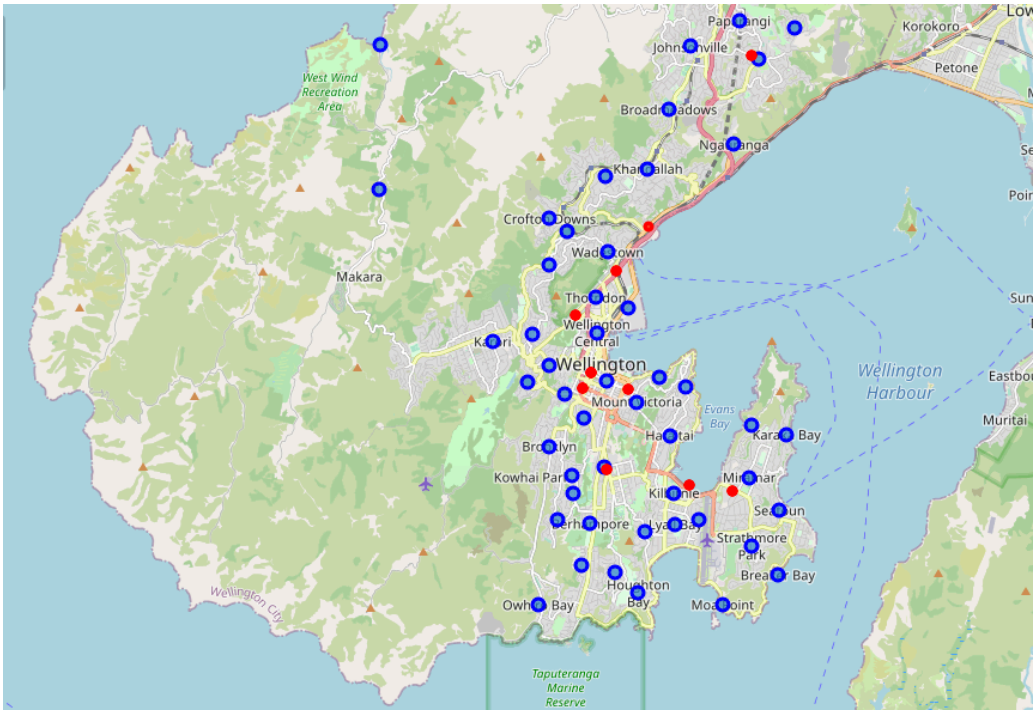
French Restaurants DataFrame associating French restaurants and suburbs information.

	Suburb	Suburb Latitude	Suburb Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Aro valley	-41.295328	174.766580	Le Mirage Patisserie & Cuisine Centrale	-41.294144	174.772037	French Restaurant
1	Aro valley	-41.295328	174.766580	La Recette	-41.290730	174.774325	French Restaurant
2	Kelburn	-41.289205	174.762393	Le Mirage Patisserie & Cuisine Centrale	-41.294144	174.772037	French Restaurant
3	Kelburn	-41.289205	174.762393	La Recette	-41.290730	174.774325	French Restaurant
4	Kilbirnie	-41.316646	174.798093	La Rotisserie Du Canard	-41.314972	174.802528	French Restaurant

Table 3. French Restaurants DataFrame

METHODOLOGY

Before beginning, let's have a sense of the localisation of the suburbs and the French restaurants using the *Folium* library to display a map:



In this section, we are going to describe the exploratory data analysis and inferential statistical testing that were conducted. Then, we are going to present how we used these data in order to solve our problem, including what machine learning were used and why.

Little reminder:

The Pearson Correlation measures the linear dependence between two variables X and Y. The resulting coefficient is a value between -1 and 1 inclusive, where:

- **1**: Total positive linear correlation.
- **0**: No linear correlation, the two variables most likely do not affect each other.
- **-1**: Total negative linear correlation.

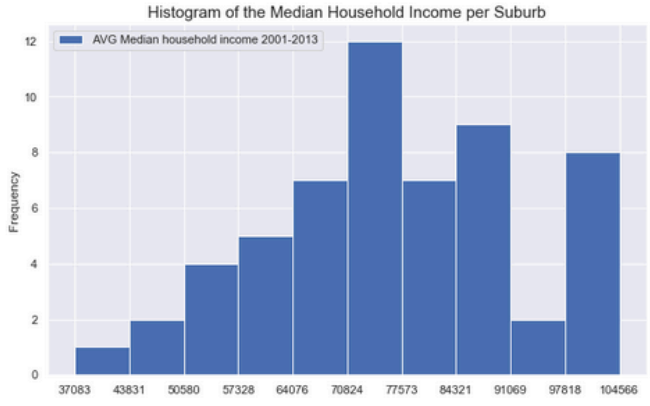
The P-value is the probability value that the correlation between these two variables is statistically significant. Normally, we choose a significance level of 0.05, which means that we are 95% confident that the correlation between the variables is significant. By convention, when the

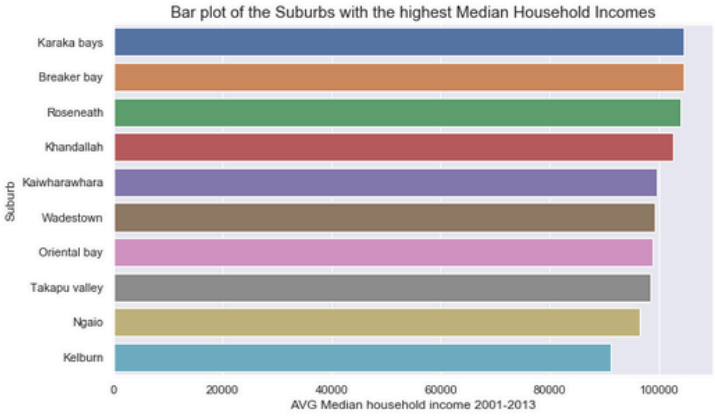
- p-value is < 0.001: we say there is strong evidence that the correlation is significant.
- p-value is < 0.05: there is moderate evidence that the correlation is significant.
- p-value is < 0.1: there is weak evidence that the correlation is significant.
- p-value is > 0.1: there is no evidence that the correlation is significant.

Exploratory data analysis

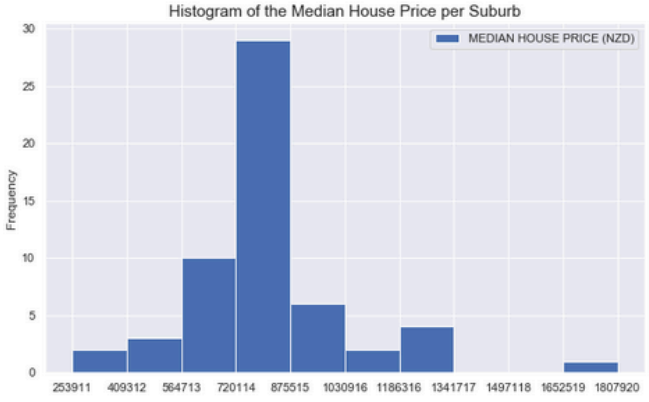
Median House Price & Median Household Income

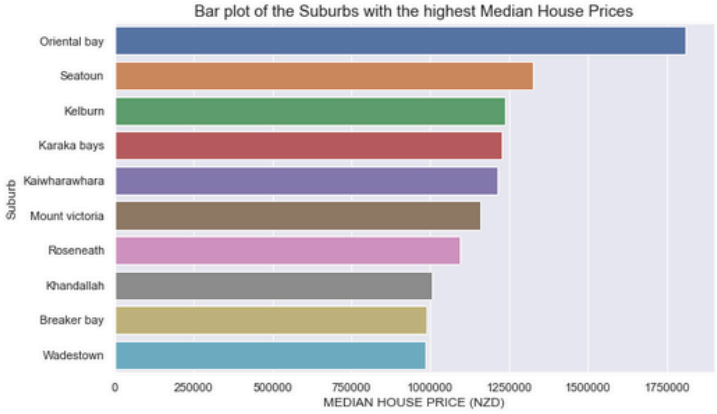
We plot the histogram of the mean household income and the top 10 suburbs having the highest median household income.



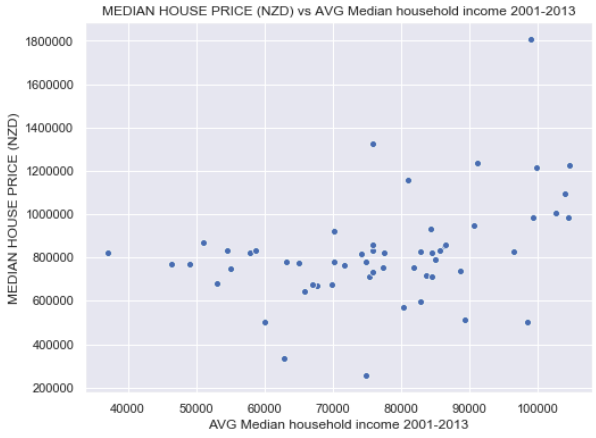


We plot the histogram of the mean house price income and the top 10 suburbs having the highest median house price.



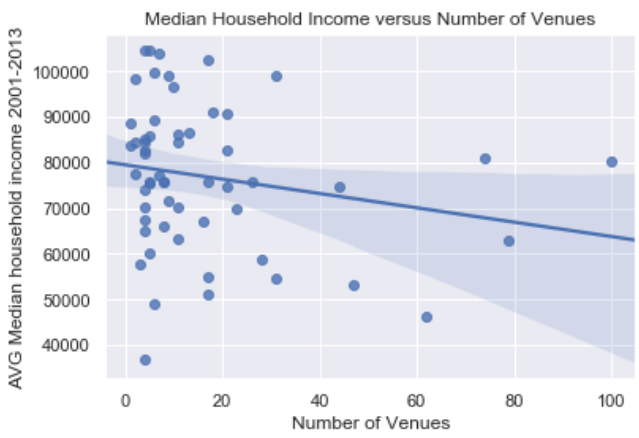


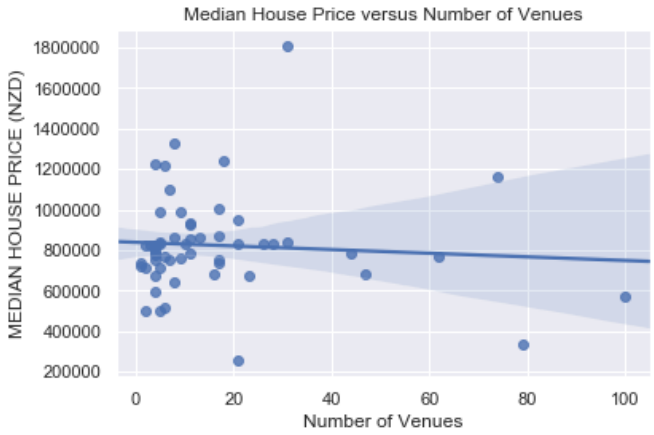
Here is a scatter plot of the two variables:



The Pearson Correlation Coefficient is 0.3897 with a P-value of P = 0.0027. Since we have: 0.001 < P-value < 0.05, there is moderate evidence that the correlation is significant. We could have thought that the correlation would have been more significant.

Let's see how these data compare with the numbers of venues:

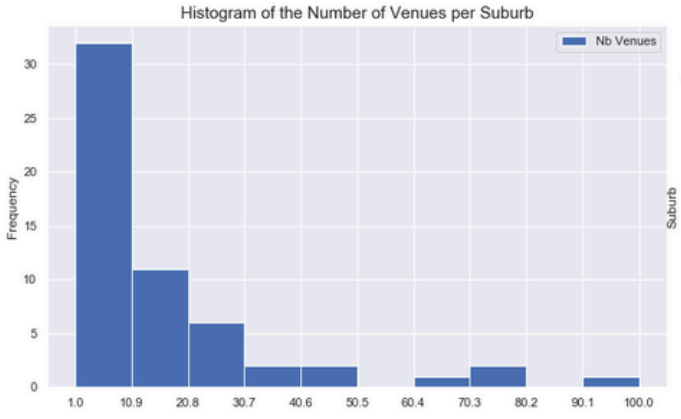


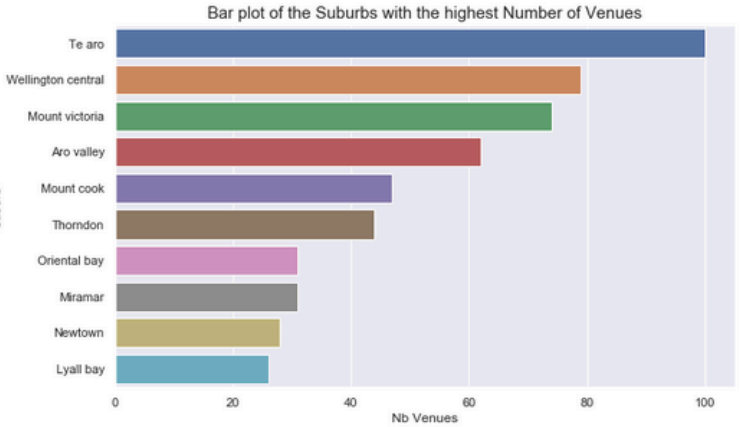


The Pearson Correlation Coefficient are - 0.2 and - 0.08 respectively with a P-value of P = 0.13 and P = 0.57 respectively. Therefore, there is no evidence of significant correlation.

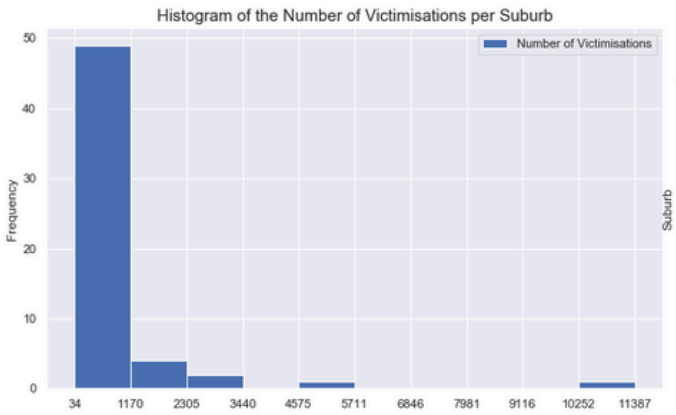
Number of Victimisations & Number of Venues

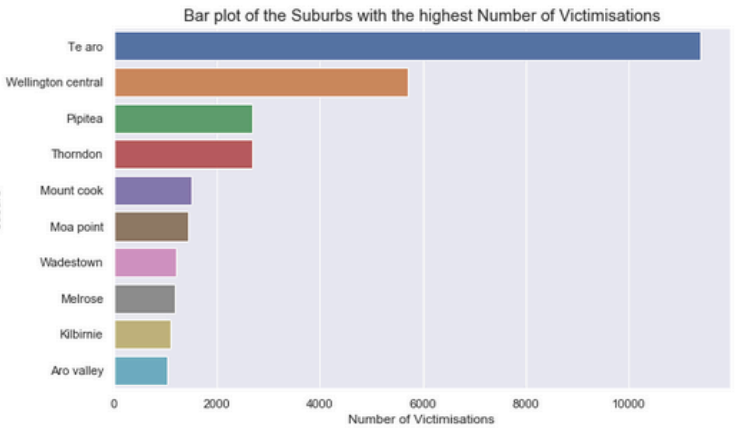
We plot the histogram of the number of venues and the suburbs having the most venues.



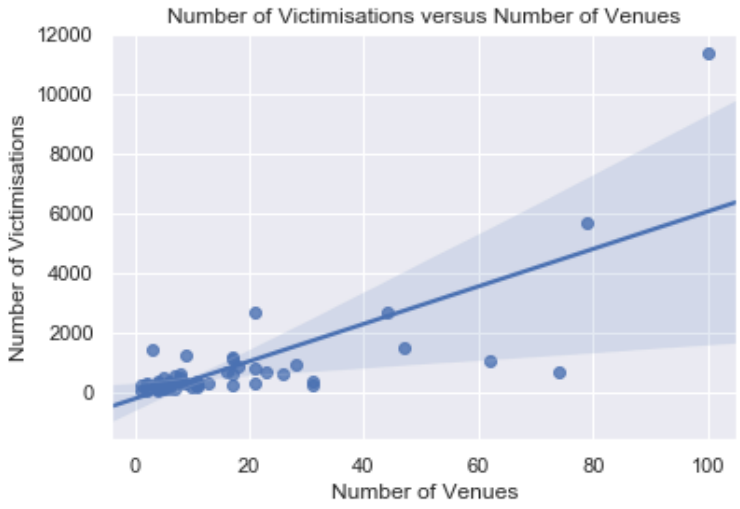


We plot the histogram of the victimisations and the suburbs having the most victimisations.





Let's study the correlation of the number of venues and victimisation:



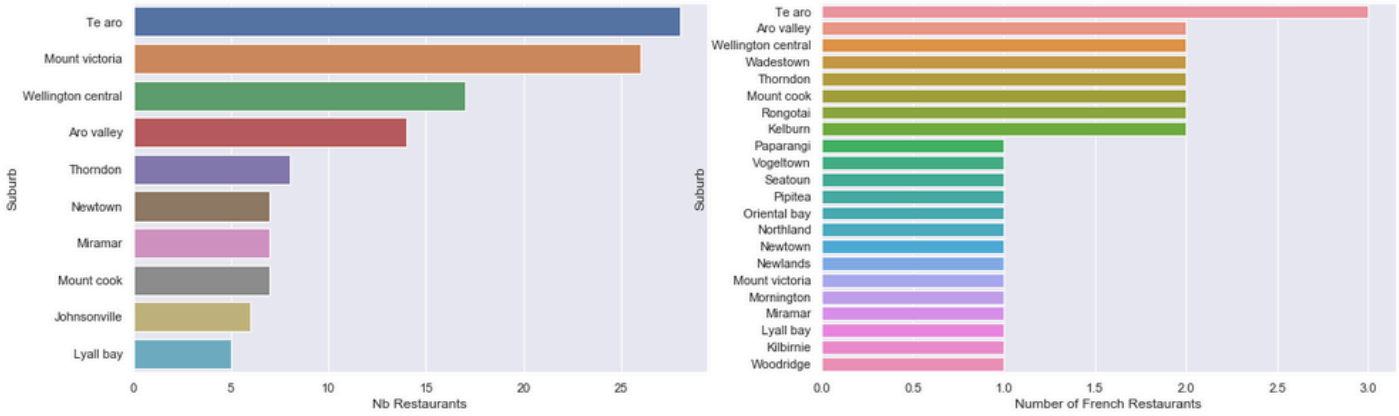
The Pearson Correlation Coefficient is 0.76 with a P-value of P = 5e-12

The P-value being less than 0.001, there is a strong evidence that the correlation is significant. This is logical as you expect the places having the most venues being places where victimisations are more likely to happen as they are more venues to steal from and they would concentrate a higher density of people in the street.

As we can see on the map and as a rapid search on internet reveals, the place with the most venues and the most victimisations, Te Aro, is in the city center and "is New Zealand's largest entertainment district and thrives at night" (source : Wikipedia Page of Te Aro). This clearly explains the particular high volume of victimisations and venues.

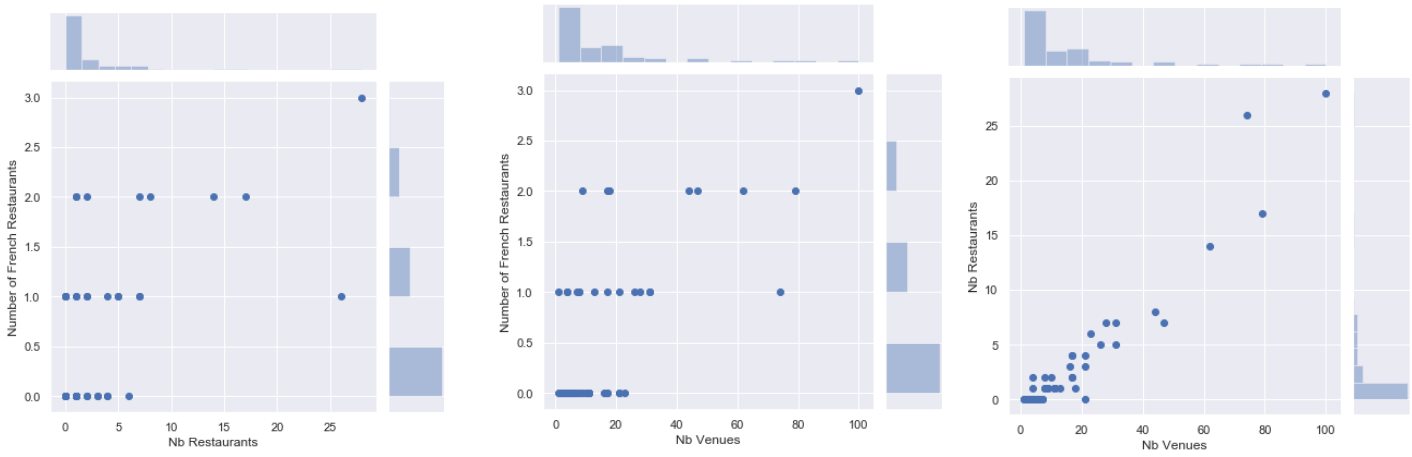
Number of Restaurants & French Restaurants

Let's plot the bar plots of the Number of Restaurants and Number of French Restaurants per Suburb:



We can see that without surprise, Te Aro and the other central suburbs have the most restaurants. Moreover, we can note that a suburb does not have more than three French restaurants and the only suburb having three of them is Te Aro.

Let's visualize the number of French restaurant against the number of restaurants and the number of venues. We expect them to be of course correlated, and we are particularly interested in the outliers (having a more venues in comparison to the number of restaurants and French restaurants).



Indeed, we observe that the datasets are somewhat positively correlated, which is what was expected.

	Nb Restaurants	Number of French Restaurants	Nb Venues
count	57.000000	57.00000	57.000000
mean	2.964912	0.54386	16.263158
std	5.728890	0.78080	20.363243
min	0.000000	0.00000	1.000000
25%	0.000000	0.00000	4.000000
50%	1.000000	0.00000	8.000000
75%	3.000000	1.00000	18.000000
max	28.000000	3.00000	100.000000

We also describe the data on the numbers of restaurants, French restaurants and venues.

Most common venues per Suburb

Using the Venues DataFrame, we can extract the 10 most common venues per Suburbs, giving us the following DataFrame:

	Suburb	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Aro valley	Café	Bar	Coffee Shop	Italian Restaurant	Restaurant	Pizza Place	Fish Market	Burger Joint	Brewery	Bakery
1	Berhampore	Park	Recreation Center	Laundromat	Skate Park	Hockey Arena	Electronics Store	Fish Market	Fish & Chips Shop	Fast Food Restaurant	Farmers Market
2	Breaker bay	Playground	Smoke Shop	Beach	Scenic Lookout	Nudist Beach	Train	Exhibit	Food	Fish Market	Fish & Chips Shop
3	Broadmeadows	Train Station	Train	Food & Drink Shop	Food	Fish Market	Fish & Chips Shop	Fast Food Restaurant	Farmers Market	Falafel Restaurant	Exhibit
4	Brooklyn	Convenience Store	Pie Shop	Fast Food Restaurant	Gastropub	Burger Joint	Park	Pharmacy	Café	Deli / Bodega	Indie Movie Theater

Table 4. Most Common Venues DataFrame

This DataFrame will enable us to get a feel of the atmosphere of each suburb and to know which amenities are present around the possible location of the French restaurant we want to open.

Using the data to find the best suburbs to open a French restaurant

To solve our problem, we proceed in several steps. As a reminder, we want to find areas where people have a sufficient standard of living to frequent the fancy French restaurant regularly, areas not overcrowded with restaurants but still lively, having enough venues around; areas where the housing price is not too expensive since we want to reduce the initial cost of the acquisition of the restaurant building.

First Step : To answer the first and last criteria, we cluster our suburbs in function of their median household income and median house price in order to find suburbs where the median household income is high (or very high), but with a moderate (or less) median house price. The K-Means clustering algorithm is used to do this, and we use the 'elbow' method to choose the right K.

Second Step : After we have reduced the pool of interesting suburbs, we eliminate the ones with too high of a number of victimisations.

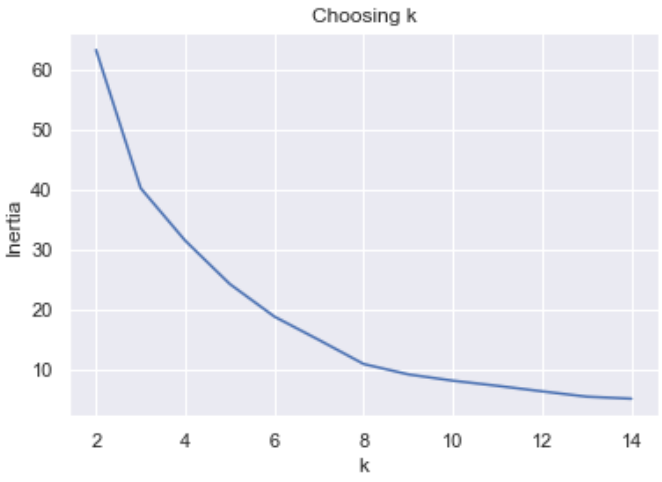
Third & Last Step : Then, we analyse the remaining suburbs in function of their number of restaurants, French restaurants, and venues in comparison of all the suburbs of Wellington in order to find an area overcrowded with restaurants but still lively, having enough venues around.

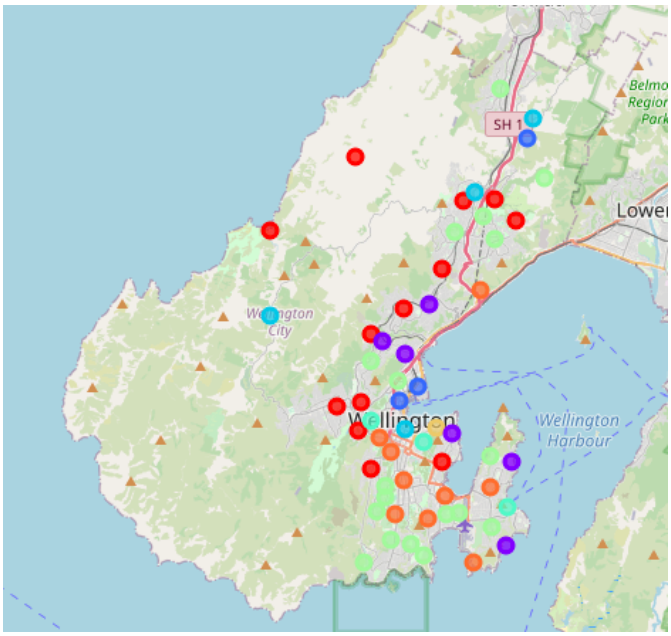
At that point, we have found all the suburbs that meet all the main criteria listed above.

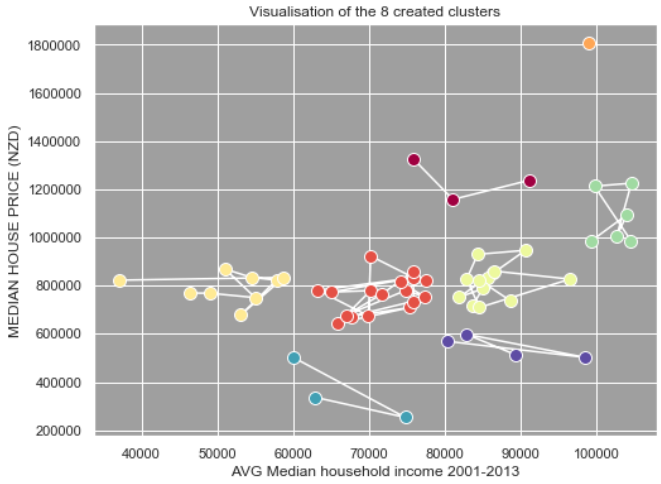
RESULTS

1 - Clustering the Suburbs on median household income and median house price with the K-Means algorithm

To run the K-Means algorithm we find the best value of K. Using the 'elbow' method, the optimal value of K is found to be 8. This value results in the following clusters:





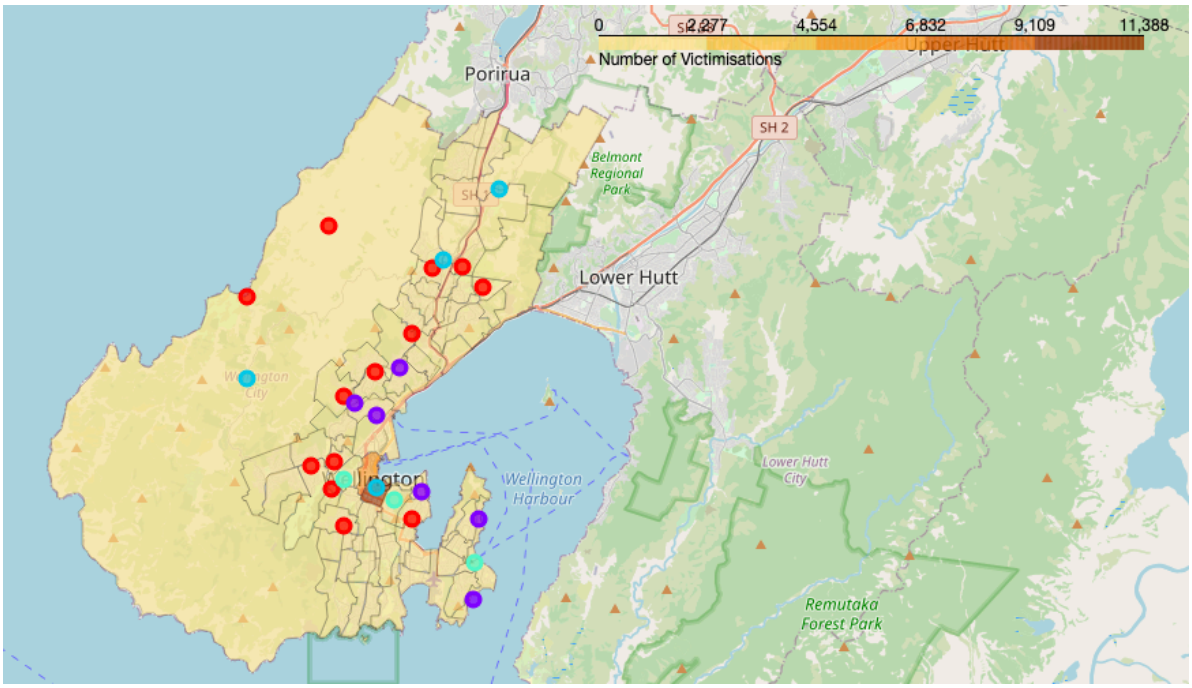


	Household Income	House Price
Cluster 0	HIGH	MODERATE - HIGH
Cluster 1	MODERATE	LOW - MODERATE
Cluster 2	VERY HIGH	VERY HIGH
Cluster 3	VERY LOW	LOW
Cluster 4	HIGH	LOW - MODERATE
Cluster 5	VERY HIGH	MODERATE - HIGH
Cluster 6	LOW - MODERATE	VERY LOW
Cluster 7	HIGH	VERY LOW - LOW

From this decomposition, we can clearly see which clusters we are interested in and which we can eliminate. We want clusters where the household income binned is at least 'High' and a correct house price in comparison. We are especially interested in Cluster 7, Cluster 4, Cluster 5 and Cluster 0 in this order of interest. Now, there is 26 out of 57 suburbs considered.

2 - Eliminating suburbs using Victimisations Data

Choropleth map of the selected suburbs with the number of victimisations shown per area unit:



We can see that there is only two area unit where we don't want to open a French Restaurant due to a high number of victimisations. Among our selected suburbs, Te Aro belong to the area unit with the highest number of Victimisations. For this reason, Te Aro is not considered as an option anymore. Furthermore, besides from Te Aro, the suburbs in Cluster 7 are quite far from the city center. This is something to take into consideration for the future. Now, there is 25 out of 57 suburbs considered.

3 - Filtering the selected suburbs regarding their numbers of venues, restaurants, and French restaurants

Statistical description of the considered suburbs:

	Nb Restaurants	Number of French Restaurants	Nb Venues
count	25.000000	25.000000	25.000000
mean	1.720000	0.320000	10.720000
std	5.168172	0.627163	14.455449
min	0.000000	0.000000	1.000000
25%	0.000000	0.000000	4.000000
50%	0.000000	0.000000	6.000000
75%	1.000000	0.000000	11.000000
max	26.000000	2.000000	74.000000

The numbers of restaurants in our considered suburbs are in average quite inferior to the maximum number of restaurants in one suburb, 28. One exception is Mount Victoria which has almost as many restaurants that has 26 restaurants. Therefore, we consider Mount Victoria to be overcrowded with restaurants and we no longer consider it.

Ideally, we also want a suburb not too close to too many French restaurants. Let reduce our considered suburbs to those with less than two French restaurants.

We also want the French restaurant to be in a suburb lively enough, so it needs to be in the 75% of the numbers of venues of our considered suburbs (i.e. number of Venues >= 11).

Finally, there is 6 out of 57 suburbs selected : Hataitai, Karori, Khandallah, Northland, Brooklyn, and Highbury.

DISCUSSION

All of this 6 suburbs meet the initial criteria. Now, the choice of the suburb to open the restaurant will depend on the weight and the particular desire of the owner of the French restaurant. We can isolate three types of criteria that can influence a further choice of suburbs.

Criteria 1: the numbers

	Ward	Suburb	Latitude	Longitude	Number of Victimisations	Household Income Binned	House Price Binned	Nb Venues	Nb Restaurants	Number of French Restaurants
0	Eastern Ward	Hataitai	-41.304278	174.796780	772.000000	High	Moderate	21.0	4.0	0.0
1	Onslow-Western Ward	Karori	-41.284109	174.746052	300.666667	High	Low	21.0	3.0	0.0
2	Onslow-Western Ward	Khandallah	-41.246742	174.790589	226.000000	Very High	Moderate	17.0	2.0	0.0
3	Onslow-Western Ward	Northland	-41.282339	174.757356	276.000000	High	Low	13.0	1.0	1.0
4	Southern Ward	Brooklyn	-41.306574	174.762354	295.000000	High	Low	11.0	1.0	0.0
5	Lambton Ward	Highbury	-41.292798	174.756096	153.000000	High	Moderate	11.0	1.0	0.0

- If we want to be in a very safe suburb, 'Hataitai' can be dropped.
- If we want to minimise the cost of the acquisition of the restaurant, we would choose between 'Karori', 'Northland', and 'Highbury'.
- If we want to be in a suburb with enough other venues around the restaurants we would pick between 'Hataitai', 'Karori', and ‘Khandallah.

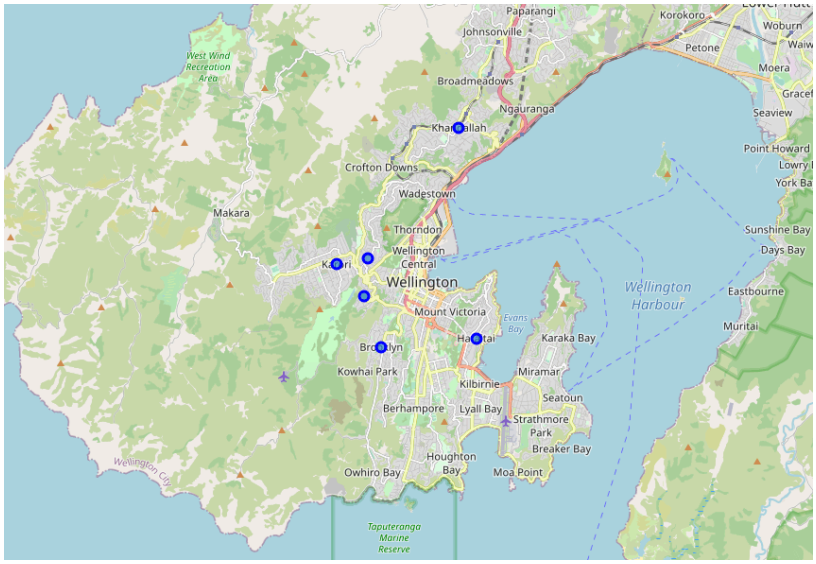
Criteria 2: the suburb ambiance

Looking at the most common venues of these neighborhood, we can get a feel of the ambiance of the suburb we want to be in.

	Suburb	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
4	Brooklyn	Convenience Store	Pie Shop	Fast Food Restaurant	Gastropub	Burger Joint	Park	Pharmacy	Café	Deli / Bodega	Indie Movie Theater
10	Hataitai	Park	Burger Joint	Café	Grocery Store	Indian Restaurant	Fish & Chips Shop	Moroccan Restaurant	Beach	Tennis Court	Thai Restaurant
11	Highbury	Nature Preserve	Café	Pub	Coffee Shop	Fish & Chips Shop	Malay Restaurant	Pizza Place	Tunnel	Bakery	Train
18	Karori	Café	Supermarket	Indian Restaurant	Bakery	Pool	Pub	Photography Studio	Deli / Bodega	Park	Sandwich Place
20	Khandallah	Café	Train Station	Park	Convenience Store	Deli / Bodega	Supermarket	Pizza Place	Asian Restaurant	Video Store	American Restaurant
37	Northland	Pizza Place	Tunnel	Convenience Store	Burger Joint	Thai Restaurant	Tourist Information Center	Park	Botanical Garden	Liquor Store	Playground

- If the restaurant owners and workers like getting to a café between the restaurants opening hours, we could pick between 'Hataitai', 'Highbury', 'Karori', and 'Khandallah.
- Being a French restaurant, we probably appreciate being close to a bakery to get fresh bread every morning. In that case, 'Karori' and 'Highbury' would be preferred.

Criteria 3: the localisation



- Depending where the restaurants owner and employees live, we can decide on the closest suburb.
- Moreover, we might want to be in a suburb close to a farmer market to get fresh fruits and vegetables everyday.
- Furthermore, we might want to be in a specific part of the city to have the custumers enjoy a great scenic view on Wellington and its Harbour.

These criteria would guide anyone to make a final decision depending of its specific needs. A suburb that meet most of the criteria mentioned above is Karori.

CONCLUSION

In this study, we analysed the suburbs of Wellington, New Zealand, based on their median house price, median household income, numbers of victimisations, numbers of venues, numbers of restaurants and French restaurant. Then, in order to fulfill our objective to find the best suburbs to open a French restaurant, we clustered the suburbs using the K-Means Clustering algorithm on the median house price and median household income. From there, we selected and reduced the number of suburbs using the rest of the data. This analysis has permitted to isolate 6 suburbs out of the 57 suburbs of Wellington. Finally, we gave clues and guidelines in order to make a thoughtful choice out of this 6 depending on the more specific need of the prospect French restaurant owners and employees.