Where to open a coffee shop?

Thomas Kohlborn

March 21, 2021

Introduction

Anyone, who wants to open a venue in a certain city is faced with the question of "Where should I open my venue?". Obviously, location is very important as it is usually the main driver in retail and gastronomy.

Imagine I am a owner of a coffee shop franchise and I would like to understand in what suburb should I open a new coffee shop in Vancouver. My question is the following:
- In what suburb should I open a new coffee shop?

This problem is relevant for any person interested in opening a venue in a certain city. The methodology taken for the rest of the project, can easily be modified to work for any location and any venue category listed on Foursquare.

Data

What data will be used?

In order to answer the question above, I will need information about existing venues and I will need some information about Vancouver's suburbs. In particular, it is not enough to simply look at the number of coffee shops and a suburb and decide on the suburb with the lowest number of coffee shops. Why?

Maybe the suburb with the lowest number of coffee shops has the highest unemployment rate, so disposable income in that are will be quite restricted. We need to understand what variables or characteristics of the suburbs have a statistically significant impact on the number of coffee shops.

Therefore, we need two data sets:

- **Census data:** Census data for Vancouver is available here: https://opendata.vancouver.ca/explore/dataset/census-local-area-profiles-2016/information/

  Unfortunately the census data is from 2016, but we need to assume for this exercise that the demographics have not changed until today. The census data lists 22 suburbs (or local areas) and provides demographic information about the people, who are living there. For example, there are 4000 people between the age of 0 and 14 living in 'Downtown' at the time of the census. The data can freely be downloaded in a csv or xls format.
- **FourSquare data:** Using the API from FourSquare allows us to get information about venues (amongst other things) and use that information to identify any trends or clusters.

Methodology

1. Load publicly available census data for Vancouver
2. Extract only relevant variables from the census data that could have a significant impact on the number of coffee shops (e.g., language spoken at home is deemed to have minimal impact on the number of coffee shops in a suburb)
3. Derive the number of coffee shops for each suburb in Vancouver through using the Foursquare API
4. Merge the two different data sets
5. Run a correlation analysis against all variables
6. Understand if any of the calculated correlation coefficients are statistically significant with p<=0.05
7. Results: Identify suburb based on the values of the statistically significant variables

   The suburb identified in step 7, will be the one that is proposed to be chosen to investigate further re a new coffee shop development.

Analysis

Load publicly available census data for Vancouver¶

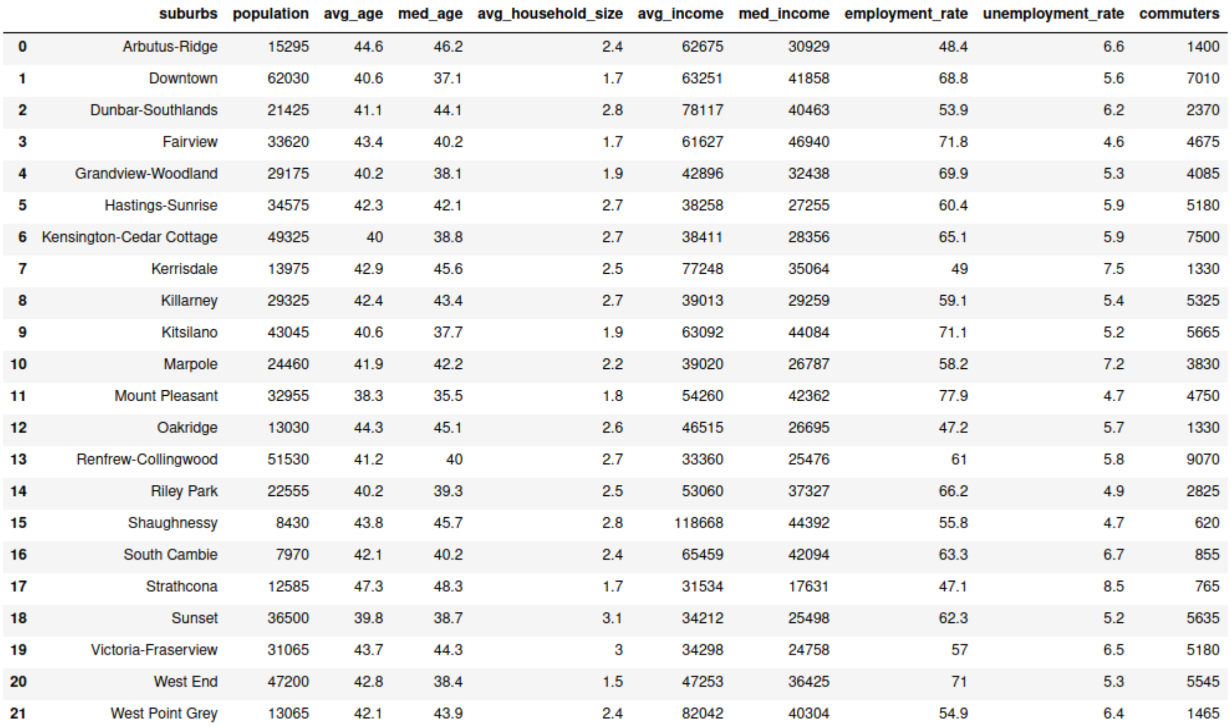
Vancouver government provides free and open access to a myriad of different data sets. The data set that interests us contains demographic information structured by the different suburbs (i.e., local areas). This information can be found here:

https://webtransfer.vancouver.ca/opendata/csv/

Extract only relevant variables from the census data

Based on the Census information, I have decided to choose the number of people living in each suburb ('population', the average age of the respective population in each suburb ('avg_age'), as well as median age('med_age'), average household size ('avg_household_size'), average income ('avg_income'), median income ('med_income'), employment rate ('employment_rate') and the number of people commuting to other suburbs ('commuters') as variables that could potential impact on the number of coffee shops in a suburb.

The data set also needed to be cleaned to a certain extent, as certain rows were empty and some information was not relevant. After extracting and cleaning the data set, the following table was used for further analysis (please see Table 1).



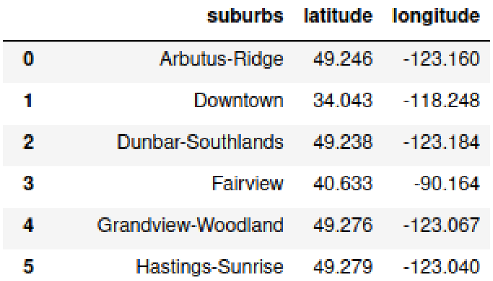
	suburbs	population	avg_age	med_age	avg_household_size	avg_income	med_income	employment_rate	unemployment_rate	commuters
0	Arbutus-Ridge	15295	44.6	46.2	2.4	62675	30929	48.4	6.6	1400
1	Downtown	62030	40.6	37.1	1.7	63251	41858	68.8	5.6	7010
2	Dunbar-Southlands	21425	41.1	44.1	2.8	78117	40463	53.9	6.2	2370
3	Fairview	33620	43.4	40.2	1.7	61627	46940	71.8	4.6	4675
4	Grandview-Woodland	29175	40.2	38.1	1.9	42896	32438	69.9	5.3	4085
5	Hastings-Sunrise	34575	42.3	42.1	2.7	38258	27255	60.4	5.9	5180
6	Kensington-Cedar Cottage	49325	40	38.8	2.7	38411	28356	65.1	5.9	7500
7	Kerrisdale	13975	42.9	45.6	2.5	77248	35064	49	7.5	1330
8	Killarney	29325	42.4	43.4	2.7	39013	29259	59.1	5.4	5325
9	Kitsilano	43045	40.6	37.7	1.9	63092	44084	71.1	5.2	5665
10	Marpole	24460	41.9	42.2	2.2	39020	26787	58.2	7.2	3830
11	Mount Pleasant	32955	38.3	35.5	1.8	54260	42362	77.9	4.7	4750
12	Oakridge	13030	44.3	45.1	2.6	46515	26695	47.2	5.7	1330
13	Renfrew-Collingwood	51530	41.2	40	2.7	33360	25476	61	5.8	9070
14	Riley Park	22555	40.2	39.3	2.5	53060	37327	66.2	4.9	2825
15	Shaughnessy	8430	43.8	45.7	2.8	118668	44392	55.8	4.7	620
16	South Cambie	7970	42.1	40.2	2.4	65459	42094	63.3	6.7	855
17	Strathcona	12585	47.3	48.3	1.7	31534	17631	47.1	8.5	765
18	Sunset	36500	39.8	38.7	3.1	34212	25498	62.3	5.2	5635
19	Victoria-Fraserview	31065	43.7	44.3	3	34298	24758	57	6.5	5180
20	West End	47200	42.8	38.4	1.5	47253	36425	71	5.3	5545
21	West Point Grey	13065	42.1	43.9	2.4	82042	40304	54.9	6.4	1465

Table 1: Data extraction of census data

Derive the number of coffee shops for each suburb

As a first step, we need to get the longitude and latitude for each suburb, which will then be used in our requests to the FourSquare API.

Therefore, we have copied the values of the first column in Table 1 into a list and used the list to derive the coordinates via GeoPy[1] The extract of a table containing the name of the suburb as well as latitude and longitude can be found in Table 2.



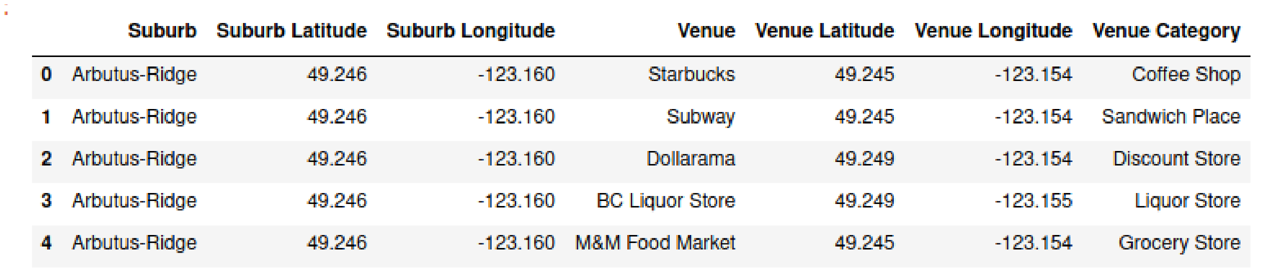
	suburbs	latitude	longitude
0	Arbutus-Ridge	49.246	-123.160
1	Downtown	34.043	-118.248
2	Dunbar-Southlands	49.238	-123.184
3	Fairview	40.633	-90.164
4	Grandview-Woodland	49.276	-123.067
5	Hastings-Sunrise	49.279	-123.040

Table 2: Deriving latitude and longitude through GeoPy

[1] Please see https://geopy.readthedocs.io/en/stable/ Last accessed 21/03/2021

Get Foursquare data

I have then used the list above to derive each of the venues in the suburbs of Vancouver and saved the data in a separate data frame. An extract of that data frame can be found in Table 3.

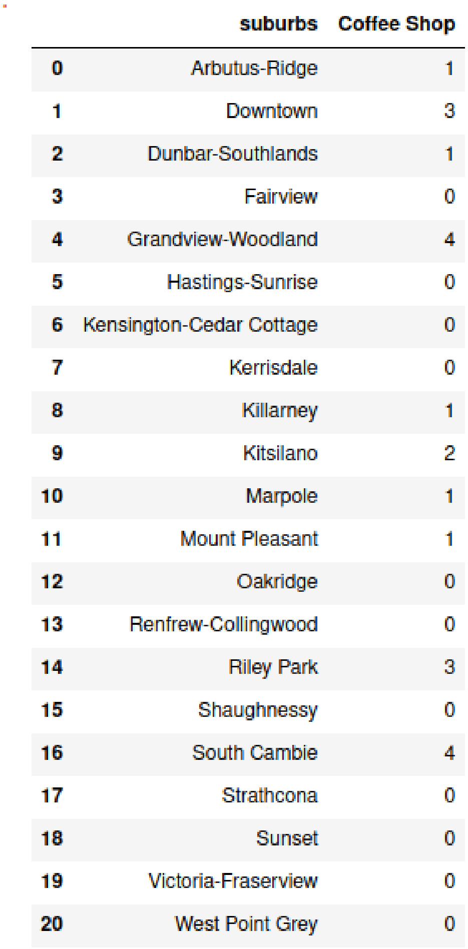


	Suburb	Suburb Latitude	Suburb Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Arbutus-Ridge	49.246	-123.160	Starbucks	49.245	-123.154	Coffee Shop
1	Arbutus-Ridge	49.246	-123.160	Subway	49.245	-123.154	Sandwich Place
2	Arbutus-Ridge	49.246	-123.160	Dollarama	49.249	-123.154	Discount Store
3	Arbutus-Ridge	49.246	-123.160	BC Liquor Store	49.249	-123.155	Liquor Store
4	Arbutus-Ridge	49.246	-123.160	M&M Food Market	49.245	-123.154	Grocery Store

Table 3: Retrieving each of the venues by suburb through FourSquare

Overall Vancouver has 419 venues with most venues being in Riley Park (61 venues), Killarney (52 venues) and Kitsilano (50 venues).

In this exercise, though, we are preliminary focused on ‘Coffee Shops’. Therefore, I have extracted only the venue category ‘Coffee Shops’ and discarded the remaining data set. The number of coffee shops by suburb can be found in Table 4.



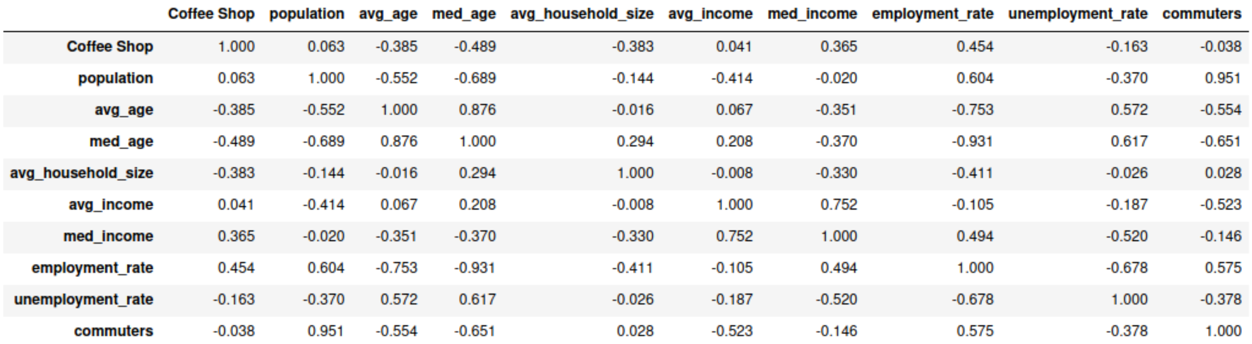
	suburbs	Coffee Shop
0	Arbutus-Ridge	1
1	Downtown	3
2	Dunbar-Southlands	1
3	Fairview	0
4	Grandview-Woodland	4
5	Hastings-Sunrise	0
6	Kensington-Cedar Cottage	0
7	Kerrisdale	0
8	Killarney	1
9	Kitsilano	2
10	Marpole	1
11	Mount Pleasant	1
12	Oakridge	0
13	Renfrew-Collingwood	0
14	Riley Park	3
15	Shaughnessy	0
16	South Cambie	4
17	Strathcona	0
18	Sunset	0
19	Victoria-Fraserview	0
20	West Point Grey	0

Table 4: Number of coffee shops in each suburb

Merge the two different data sets¶

In order for us to investigate any correlation between the number of coffee shops and the variables chosen out of the census data, I have merged both data sets together.

There are different methods to derive correlation coefficients – one of the most popular method being Pearson, which has been chosen in this case as well. The correlation matrix can be seen in Table 5.



	Coffee Shop	population	avg_age	med_age	avg_household_size	avg_income	med_income	employment_rate	unemployment_rate	commuters
Coffee Shop	1.000	0.063	-0.385	-0.489	-0.383	0.041	0.365	0.454	-0.163	-0.038
population	0.063	1.000	-0.552	-0.689	-0.144	-0.414	-0.020	0.604	-0.370	0.951
avg_age	-0.385	-0.552	1.000	0.876	-0.016	0.067	-0.351	-0.753	0.572	-0.554
med_age	-0.489	-0.689	0.876	1.000	0.294	0.208	-0.370	-0.931	0.617	-0.651
avg_household_size	-0.383	-0.144	-0.016	0.294	1.000	-0.008	-0.330	-0.411	-0.026	0.028
avg_income	0.041	-0.414	0.067	0.208	-0.008	1.000	0.752	-0.105	-0.187	-0.523
med_income	0.365	-0.020	-0.351	-0.370	-0.330	0.752	1.000	0.494	-0.520	-0.146
employment_rate	0.454	0.604	-0.753	-0.931	-0.411	-0.105	0.494	1.000	-0.678	0.575
unemployment_rate	-0.163	-0.370	0.572	0.617	-0.026	-0.187	-0.520	-0.678	1.000	-0.378
commuters	-0.038	0.951	-0.554	-0.651	0.028	-0.523	-0.146	0.575	-0.378	1.000

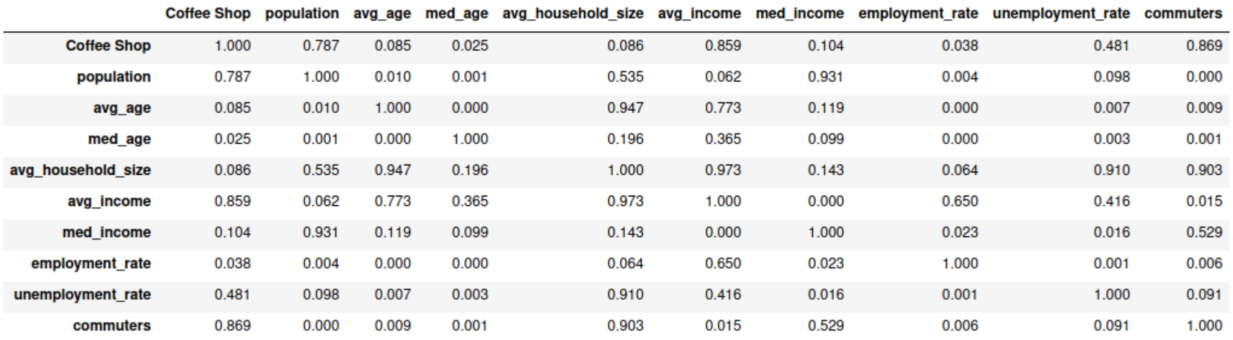
Table 5: Correlation matrix

A strong positive correlation would indicate that both variables vary to a similar degree. For example, the correlation coefficient between the median age and unemployment rate is 0.617, which indicates that the older the population in a suburb based on the median, the higher the unemployment rate.

However, just analysing the correlation coefficients is not sufficient, as we they do not indicate if the correlation can just happen by chance or if there is a deeper relationship between two variables.

In other words, if we define the null hypothesis to state that there is no relationship between two variables (results are due to chance), calculating p-values provides us with insights if the relationship is statistically significant. A p-value below 0.05 indicates strong evidence that the null hypotheses is wrong (there is a probability of less than 5% that the relationship between two variables is random).

If we analyse the p-values in Table 6, we can see that two values are statistically significant.



	Coffee Shop	population	avg_age	med_age	avg_household_size	avg_income	med_income	employment_rate	unemployment_rate	commuters
Coffee Shop	1.000	0.787	0.085	0.025	0.086	0.859	0.104	0.038	0.481	0.869
population	0.787	1.000	0.010	0.001	0.535	0.062	0.931	0.004	0.098	0.000
avg_age	0.085	0.010	1.000	0.000	0.947	0.773	0.119	0.000	0.007	0.009
med_age	0.025	0.001	0.000	1.000	0.196	0.365	0.099	0.000	0.003	0.001
avg_household_size	0.086	0.535	0.947	0.196	1.000	0.973	0.143	0.064	0.910	0.903
avg_income	0.859	0.062	0.773	0.365	0.973	1.000	0.000	0.650	0.416	0.015
med_income	0.104	0.931	0.119	0.099	0.143	0.000	1.000	0.023	0.016	0.529
employment_rate	0.038	0.004	0.000	0.000	0.064	0.650	0.023	1.000	0.001	0.006
unemployment_rate	0.481	0.098	0.007	0.003	0.910	0.416	0.016	0.001	1.000	0.091
commuters	0.869	0.000	0.009	0.001	0.903	0.015	0.529	0.006	0.091	1.000

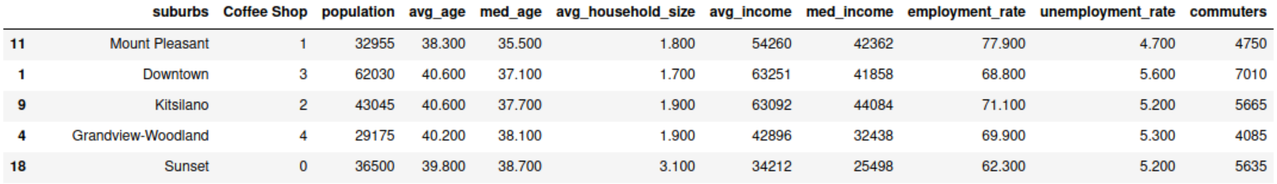
Table 6: p-values to analyse statistical significance

Based on the above table we can see that the p-values for median age ('med_age') and employment rate ('employment_rate') are statistically significant (p<=0.05). Therefore, the lower the median age in a suburb and the more people are employed, the more number of coffee shops can be found in a suburb!

Results and Discussion

Now that we know that median age and the employment rate are key factor influencing the number of coffee shops in a suburb, we need to use these results and identify which suburb we should investigate further.

Sorting our merged table based on median age, we find that Mount Pleasant has the lowest median age of all suburbs and the employment rate is higher than in 'South Cambie' and 'Grandview-Woodland'. Please refer to Table 7, which shows an extract.



	suburbs	Coffee Shop	population	avg_age	med_age	avg_household_size	avg_income	med_income	employment_rate	unemployment_rate	commuters
11	Mount Pleasant	1	32955	38.300	35.500	1.800	54260	42362	77.900	4.700	4750
1	Downtown	3	62030	40.600	37.100	1.700	63251	41858	68.800	5.600	7010
9	Kitsilano	2	43045	40.600	37.700	1.900	63092	44084	71.100	5.200	5665
4	Grandview-Woodland	4	29175	40.200	38.100	1.900	42896	32438	69.900	5.300	4085
18	Sunset	0	36500	39.800	38.700	3.100	34212	25498	62.300	5.200	5635

Table 7: Sorting our data by 'Median Age'

Hence, I would choose 'Mount Pleasant' to investigate further with regards to the suitability of establishing a new coffee shop.

Next steps would be to look at available properties and overall costs. This is required as the above analysis did not cover all variables that play a role in deciding if establishing a new coffee shop is a profitable endeavour in a certain suburb or not. I can, however, indicatively point to factors that have a statistically significant impact on the number of coffee shops in a suburb, which can and should be used as a starting point for further analysis.

Conclusion

This project set out to apply lessons learned throughout all stages of the data science methodology from defining the business problem, understanding the data that will be used, the methodology, analysing the results and discussing the practical application and next step.

In this particular project, I wanted to know if combining census data and location data can provide any meaningful insights into how population characteristics in Vancouver's suburbs impact on the number of coffee shops. This has been successfully achieved.