

The problem of choosing the location of a coffee shop in the Toronto Neighborhood

The description of the problem and a discussion of the background

Question: Which place is more appropriate to open a coffee shop in the Toronto Neighborhood? A coffee shop owner wants to open a new coffee shop in the Toronto Neighborhood. He needs to know where to choose and will have a higher probability of getting customers?

Background: First explain that everything is hypothetical. On a train, I had a communication with a coffee shop owner. I learned from the communication that the coffee shop owner wanted to expand his store in the Toronto Neighborhood, and he wanted to know which communities in Toronto are more suitable for opening coffee shops. Because the boss is more anxious and has no cost considerations, professional data such as passenger flow, regional economy, and regional population types cannot be obtained, and can only be analyzed with the help of Internet information.

The description of the data and how it will be used to solve the problem

Data description: Considering the time and cost issues, I chose to obtain data from the Internet. The composition of the data is mainly divided into two parts:

Toronto Neighborhood information, which mainly includes community name and community Geo information.

Toronto Neighborhood venue distribution information, the main information includes: the venue belongs to the Neighborhood information, place type, place name

Problem-solving ideas: Through statistical analysis of the distribution of coffee shops in all communities in Toronto, we look for communities with relatively few coffee shops for selection. While selecting objects, it is also necessary to observe the actual number of shops in the community to avoid the validity of the calculated data in the real environment.

Introduction where you discuss the business problem and who would be interested in this project

Business question: Which place to open a coffee shop in the Toronto Neighborhood is more appropriate. A coffee shop owner wants to open a new coffee shop in the Toronto Neighborhood. He needs to know where to choose and will have a higher probability of getting customers?

Understanding of the problem: According to the conventional understanding, choosing a Neighborhood with a small number of coffee shops to open a store will have a higher probability of acquiring customers. Therefore, the business problem can be simply understood as: looking for a Neighborhood with a relatively small number of coffee shops.

Concerned about the problem: The owner who seeks to open a coffee shop in the Toronto Neighborhood.

Data where you describe the data that will be used to solve the problem and the source of the data

The data used to solve the problem consists of two parts: "Toronto Neighborhood Data" and "Toronto Neighborhood venue Data".

- Toronto Neighborhood Data:

- Data source: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
- Data structure:

Field	Type
PostalCode	object
Borough	object
Neighborhood	object

- Number of data: 103 rows

- Toronto Neighborhood venue Data:

- Data source: <https://api.foursquare.com/v2/venues/explore>
- Data structure:

Field	Type
Neighborhood	object
Neighborhood Latitude	object
Neighborhood Longitude	object
Venue	object
Venue Latitude	float64
Venue Longitude	float64
Venue Category	object

- Number of data: 1096 rows

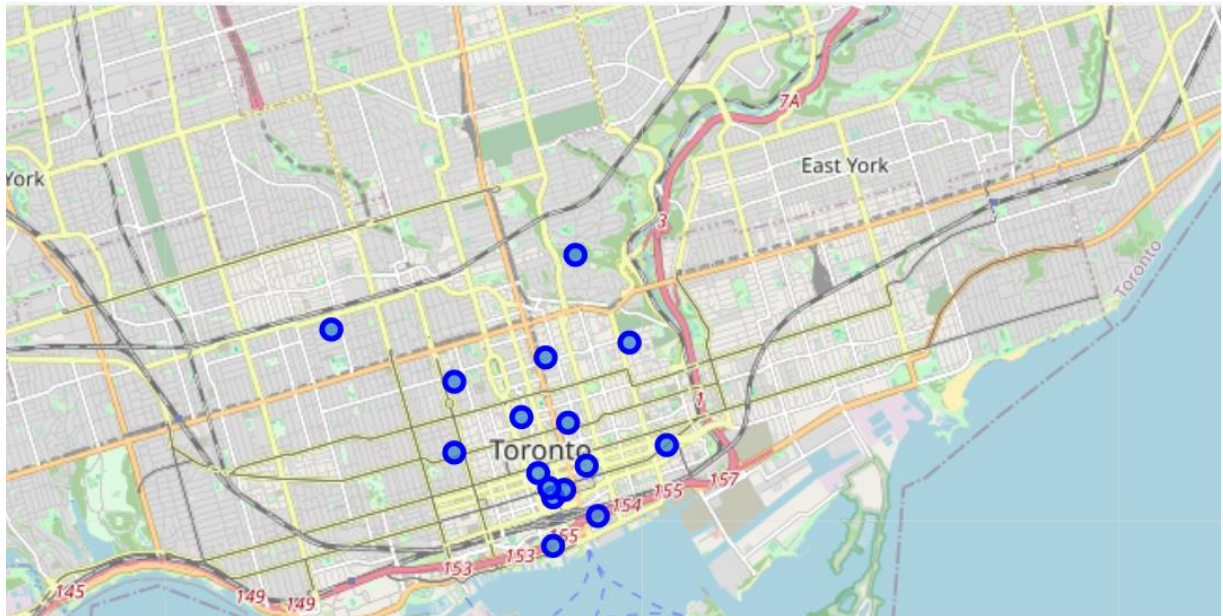
Methodology section

The First: on the basis of business understanding, determine the use of descriptive statistical analysis as a method to solve the problem, and then build the model.

The second step: complete the preparation of the data, clarify and reconstruct the content of the acquired data, the goal is to facilitate the establishment of the model.

- The initial data obtained is the Toronto Neighborhood information obtained through WIKI. This group of data is obtained from the web page, so the Soup module is used to capture the data, and the names and coordinate distributions of all Toronto communities are obtained by cleaning the data. It can also be seen that some communities are treated as Neighborhood groups in terms of data processing because of their relatively close

geographical locations.



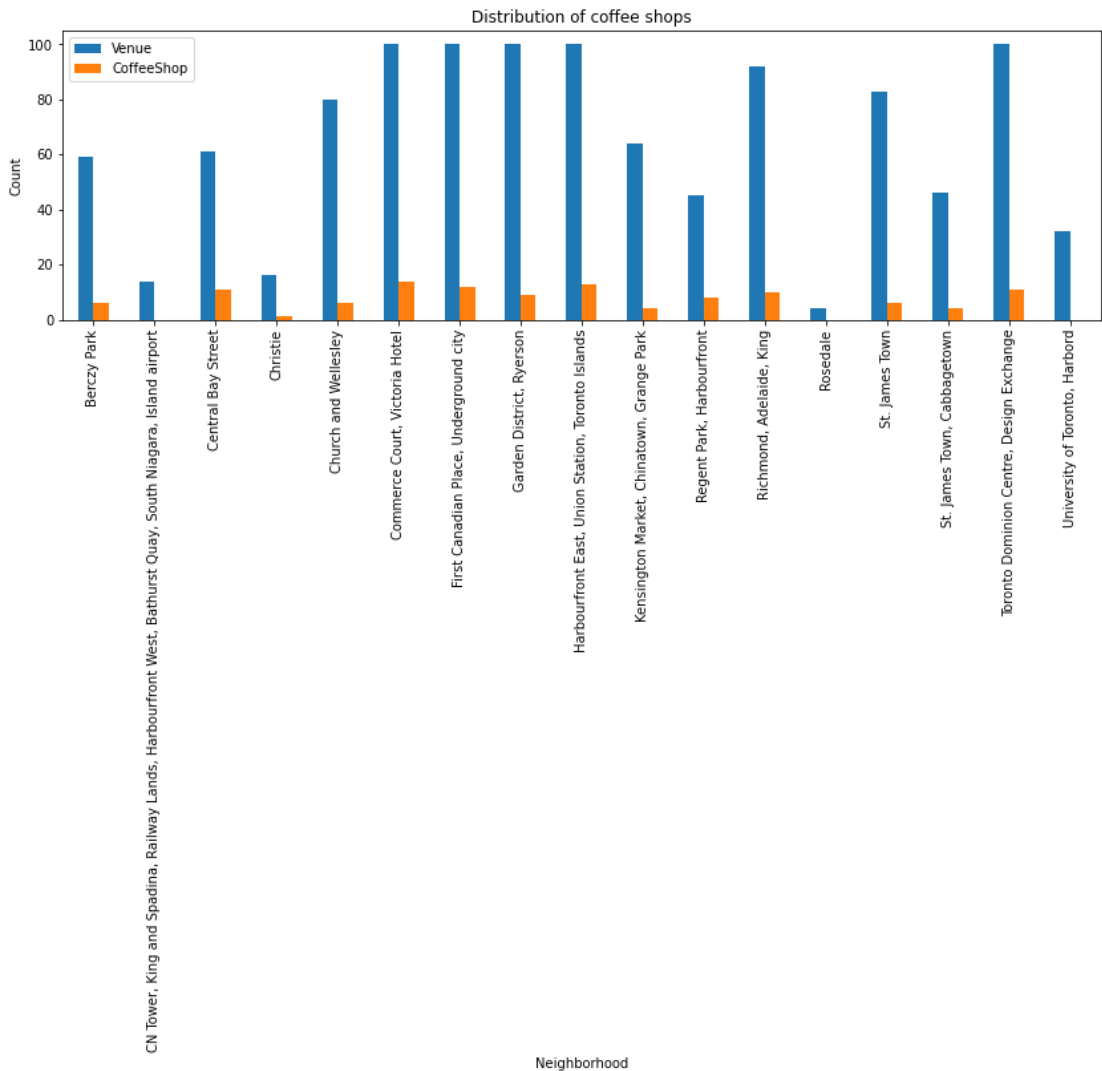
- Using Toronto Neighborhood data, you can get the location data within the coordinate range from the API interface of four. The venue data includes: the name of the venue and the type of venue. Associate the name and type data of the venue with the Neighborhood data to complete the data preparation

Index	Neighborhood	Venue	CoffeeShop	Percent
0	BerczyPark	59	6	10%
1	CNTower,KingandSpadina, RailwayLands, Har...	14	0	0%
2	CentralBayStreet	61	11	18%
3	Christie	16	1	6%
4	ChurchandWellesley	80	6	8%
5	CommerceCourt,VictoriaHotel	100	14	14%
6	FirstCanadianPlace, Undergroundcity	100	12	12%
7	GardenDistrict, Ryerson	100	9	9%
8	HarbourfrontEast, UnionStation, TorontoIslands	100	13	13%
9	KensingtonMarket, Chinatown, GrangePark	64	4	6%
10	RegentPark, Harbourfront	45	8	18%
11	Richmond, Adelaide, King	92	10	11%
12	Rosedale	4	0	0%
13	St. JamesTown	83	6	7%
14	St. JamesTown, Cabbagetown	46	4	9%
15	TorontoDominionCentre, DesignExchange	100	11	11%
16	UniversityofToronto, Harbord	32	0	0%

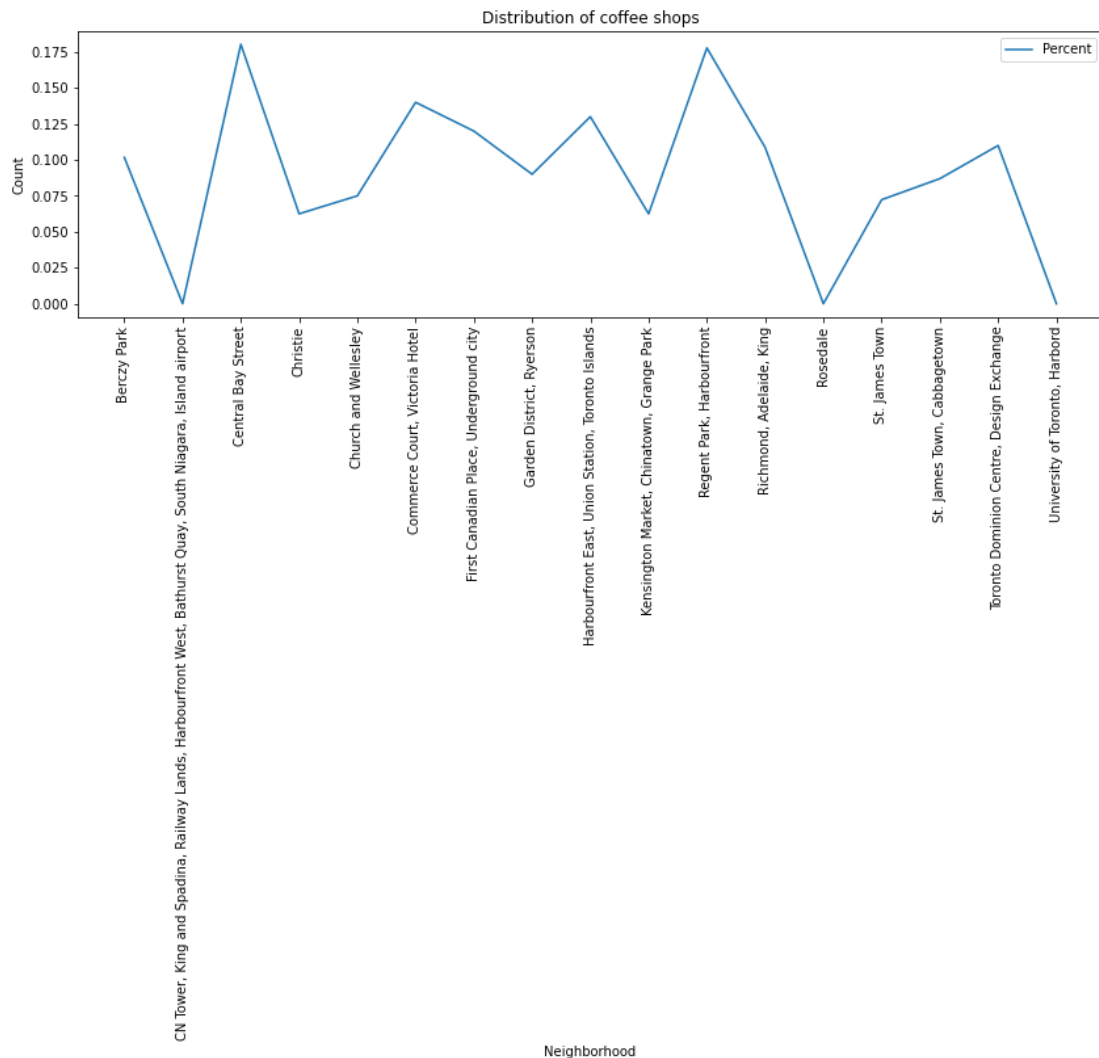
The third step: find out the algorithm that can solve the problem and complete the model establishment. The key to the problem is to find out which Neighborhood has fewer coffee shops. There are two levels of understanding of less: one is that the number of specific coffee shops is small, and the other is that the proportion of coffee shops to the total number of venue in the area is small. . Both levels have practical significance, so it is necessary to find algorithms to solve them separately.

- Count the number of coffee shops and venue and present them through visualization. The final data obtained in the second step is filtered twice: the first time is grouped by Neighborhood name, and the total number of venue in the Neighborhood is calculated; the second time is grouped by Neighborhood name and type of venue, and the filter venue type is equal to " The data of "coffee shop" is then counted to obtain the number of coffee shops in the Neighborhood; then the two data are correlated according to the name of the

Neighborhood, and the correlated data is visualized. According to the characteristics of the data, choose the visualization of the column chart, the effect is more obvious.



- On the basis of the "number of coffee shops and venue" data, the number of coffee shops is divided by the number of venue to calculate the proportion of coffee shops, and the proportion of coffee shops is used as a new data column-"Percent". And display the "Percent" data in a line graph.



Results section where you discuss the results

Through the establishment and calculation of the model, the following four results can be obtained:

1. There are Neighborhoods without coffee shops: "Rosedale", "CN Tower, King and Spadina, Railway Lands, AirPort", "University of Toronto, Harbord"
2. Except for the Neighborhoods without coffee shops, the Neighborhoods with the fewest coffee shops is: "Christie", the number is: 1.
3. Except for the Neighborhoods without coffee shops, the Neighborhoods with the least proportion of coffee shops are: "Christie" and "Kensington Market, Chinatown, Grange Park", and the proportion is: 6.25%.
4. Except for the Neighborhoods without coffee shops, the Neighborhoods where the number and proportion of coffee shops are relatively small are: "Christie", "Church and Wellesley", "Kensington Market, Chinatown, Grange Park", "St. James Town", "St. James Town, Cabbagetown", number: ≤ 6 & proportion: $< 10\%$.

Discussion and Recommendations section

Based on the above results, the following suggestions can be drawn:

1. For Neighborhoods without coffee shops, it is necessary to consider whether there is any inaccurate data acquisition. If data issues are excluded, it is recommended to choose to open coffee shops in Neighborhoods without coffee shops, unless the Neighborhood does not allow them to be established.
2. Comparing the analysis of "small number and small proportion", The small proportion can more accurately describe the unsaturation of the coffee shop in the Neighborhood, so priority is given to opening a coffee shop in "Christie" and "Kensington Market, Chinatown, Grange Park" Neighborhoods
3. Neighborhoods with relatively small numbers and proportions can also be used as candidate Neighborhoods to focus on.

Conclusion section

Through the study of the course, I learned about the meaning and methods of data science. With regard to the location of the Toronto community coffee shop, I used the knowledge of data science to try to solve the problem in accordance with the steps of business understanding, data understanding, data preparation, modeling, evaluation and description. Although many unfamiliar places were encountered during the homework, they all tried to solve them. In the follow-up, I will try to continue to learn and use data science methods to solve problems at work. I am very happy to learn this course and I have learned a lot.