Customized recommendation engine for identification of residential communities in Toronto

1. Introduction

Choosing a location when looking for a home is very important! After all, you can always update or fix your house, but you can't easily change its location and the vibe of the community! In this project, I would like to create a hypothetical business scenario with an aim of finding a compatible location for one of my clients.

Ideally, when we decide to purchase a new house/condo, we should get a feel for it by walking the neighborhood's blocks and spending some quality time in local restaurants and parks. Perhaps we could also chat with neighbors on that street to talk about crime, traffic, noise or any issues. However, what if the customer doesn't have time to go through such a thorough evaluation? What if the customer lives in a different location and they have no clue about the local neighborhood? Or, if they do have a general idea of the local neighborhood but they cannot make up their mind about multiple choices? Or, in the middle of the process, the customer thought of new important metrics that they or their significant others care about, how to make a decision effectively and efficiently in such a scenario? In this project, I aim to create a solution to address these issues.

2. Problem Statement

Which neighborhood in Toronto should be picked for one of my clients for purchasing a new house or condo?

3. Potential audience and market

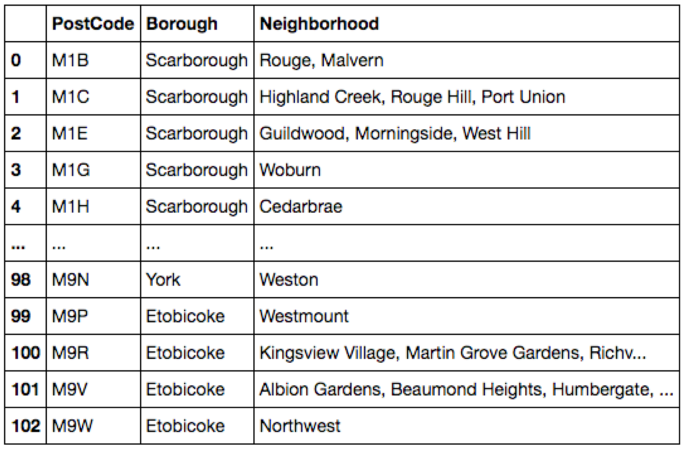
According to National Association of realtors, there is a significant home search process nearly among all generations of home buyers, ranging from buyers 39 years and younger (Millennials/Gen Yers), buyers 40 to 54 (Gen Xers), buyers 55 to 64 (Younger Baby Boomers) and buyers 65 to 73 (Older Baby Boomers), except for the buyers 74 years older (The Silent Generation). In fact, buyers typically searched for 10 weeks and looked for at a median of 10 homes. This project aims to provide customized recommendation engines for home buyers during their home search process. This solution will be particularly beneficial for buyers who have a variety of purchase metrics, which makes the normal home search process more complicated than average. These buyers can leverage the customized solution provided by this project to make their home search process faster and better.

4. Data

In this project, I will use four sets of raw data: wikipedia, geographical coordinates, data retrieved from FourSquare and a dataset provided by the customer as his own expectation/rating. The description of each dataset is provided as follows. The details can be further obtained from this link: https://github.com/yanruoyu2005/Coursera_Capstone/blob/master/Capstone%20project-week4-no_s.ipynb

4.1 Data from Wikipedia contains a list of postal codes of Canada. It will be retrieved by scraping a table from the website. This data is used to create a geographical segmentation of Toronto based on postal code, and it will be linked with the actual coordinates. Table 1 shows an overview of the data content.

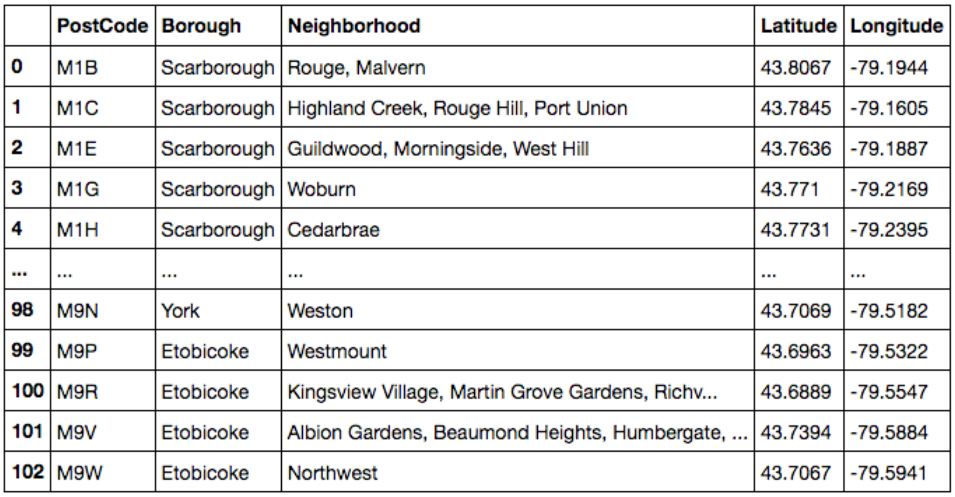
Table 1. A list of postal codes of Canada



	PostCode	Borough	Neighborhood
0	M1B	Scarborough	Rouge, Malvern
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union
2	M1E	Scarborough	Guildwood, Morningside, West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae
...	...	...	...
98	M9N	York	Weston
99	M9P	Etobicoke	Westmount
100	M9R	Etobicoke	Kingsview Village, Martin Grove Gardens, Richv...
101	M9V	Etobicoke	Albion Gardens, Beaumond Heights, Humbergate, ...
102	M9W	Etobicoke	Northwest

4.2 Geographical coordinates. This dataset was provided by the previous project. It contains the coordinates for various postal code. The aim of this dataset is to link postal code with coordinates. Table 2 provides an overview of the content of this dataset. Figure 1 provides an illustration of how this dataset can be used to generate maps.

Table 2. A list of Toronto neighborhoods with coordinates



	PostCode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Rouge, Malvern	43.8067	-79.1944
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.7845	-79.1605
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.7636	-79.1887
3	M1G	Scarborough	Woburn	43.771	-79.2169
4	M1H	Scarborough	Cedarbrae	43.7731	-79.2395
...	...	...	...	...	...
98	M9N	York	Weston	43.7069	-79.5182
99	M9P	Etobicoke	Westmount	43.6963	-79.5322
100	M9R	Etobicoke	Kingsview Village, Martin Grove Gardens, Richv...	43.6889	-79.5547
101	M9V	Etobicoke	Albion Gardens, Beaumond Heights, Humbergate, ...	43.7394	-79.5884
102	M9W	Etobicoke	Northwest	43.7067	-79.5941

Please see the full table in Notebook. This is a incomplete screenshot.

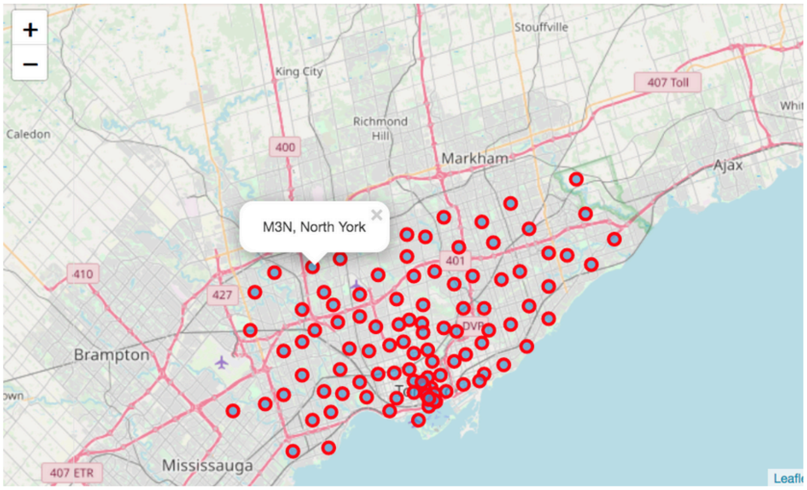
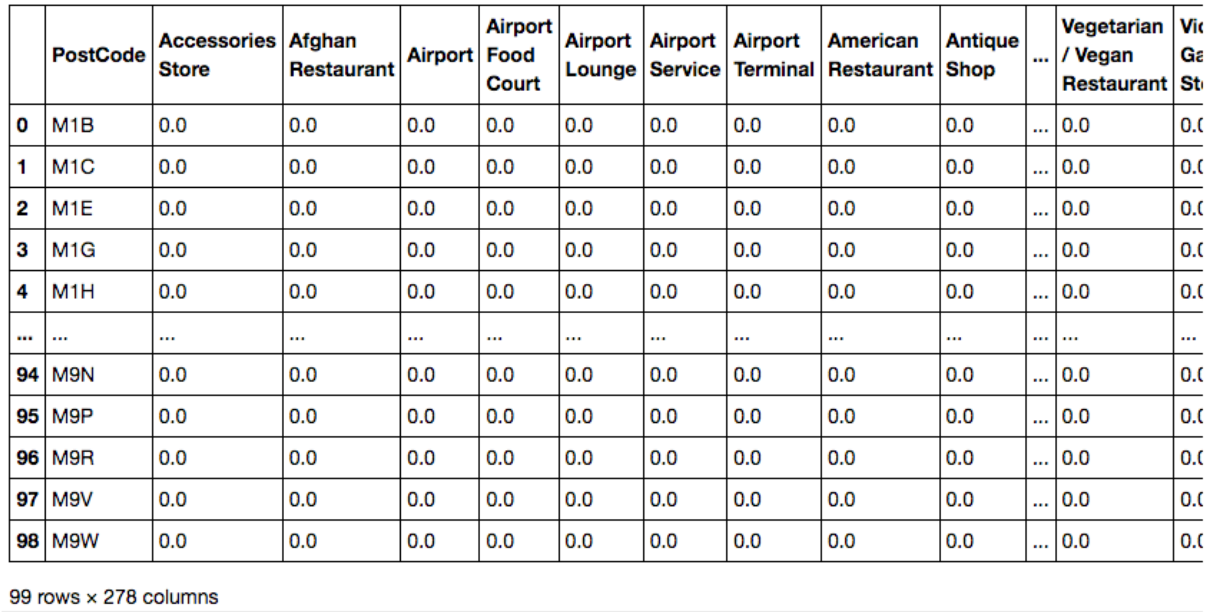


Figure 1. Mapping a list of Toronto neighborhoods.

4.3 Data retrieved from FourSquare will be used to evaluate each zone of interest. For each postal code area, obtain the frequency of occurrence of interesting venues by further data wrangling and preparation. Table 3 shows the retrieved results further processed to contain the frequency of occurrence of interesting venues of each zone. This dataset has the information to be further combined with customers' rating to create a customized recommendation.

Table 3. Frequency of occurrence of venues in various postal code areas in Toronto



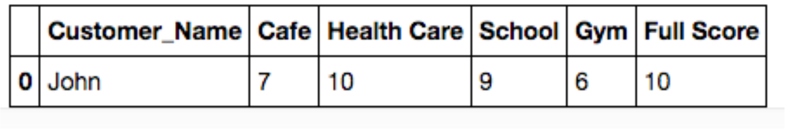
	PostCode	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	...	Vegetarian / Vegan Restaurant	Vi Ga St
0	M1B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.(
1	M1C	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.(
2	M1E	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.(
3	M1G	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.(
4	M1H	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.(
...	...	...	...	...	...	...	...	...	...	...	...	...	...
94	M9N	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.(
95	M9P	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.(
96	M9R	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.(
97	M9V	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.(
98	M9W	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.(

99 rows × 278 columns

Please see the full table in Notebook. This is a incomplete screenshot.

4.4 Dataset provided by the customer as his own expectation/rating will be used to create a customized recommendation engine. Detailed data processing will be provided in the notebook. Note that the metrics that are important to this customer have a broad and vague definition, so it will be carefully mapped to corresponding categories retrieved from FourSquare. Table 4 shows a rating result by a hypothetical customer "John". Note that he only provides metrics that he cares about, and he gave scores to each of these metrics as shown in Table 4. This information will be further used to convert Table 3 into a customized profile for John.

Table 4. Personalized rating



	Customer_Name	Cafe	Health Care	School	Gym	Full Score
0	John	7	10	9	6	10

5. Methodology

5.1 Retrieve data and perform data format cleaning and preprocessing. Details are explained in the data section.

5.2 Interpret customers’ expectation with venues more precisely. For example, café could be café and coffee shop, so both of them should be taken into consideration. Or, school could be college, high school, university, etc. These alternative key words should be identified and considered.

5.3 Filter the dataset based on customers’ expectation to reduce computational complexity.

5.4 Weighting the venues based on the actual frequency of occurrence of each venue as well as the customers’ personal preference (the rating dataset). This will be performed by doing linear algebra operations of the venue matrix.

5.5 Generate recommended candidate venues for the customer.

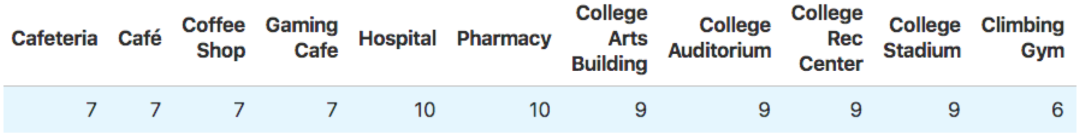
6. Results

6.1 Details of data preparation have been described as above.

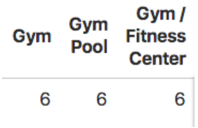
6.2 Interpretation of customer’s expectation

Originally, the customer rated four categories, they are Café, School, Gym and Health Care. It is important to consider other words that have the same or similar meaning these exact words in the retrieved results. Table 5 shows a more comprehensive list of the metrics.

Table 5. Corresponding metrics based on customer’s expectation



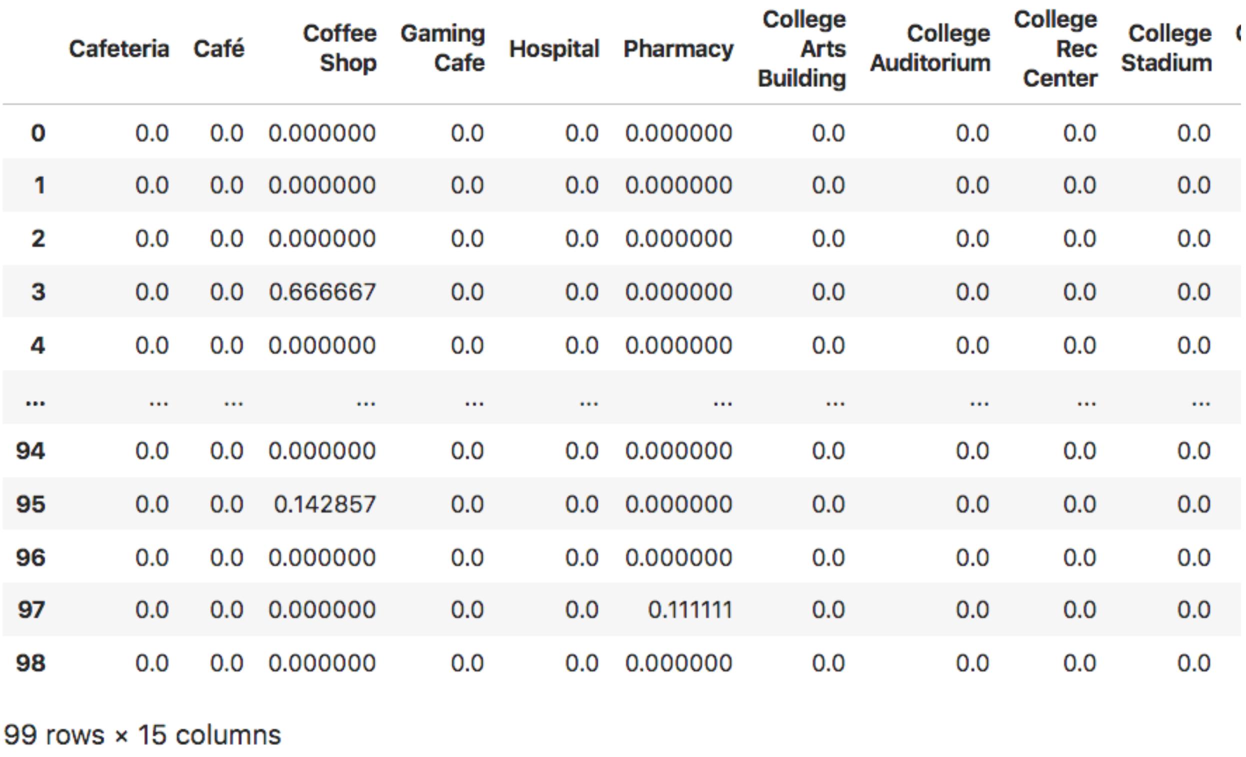
Cafeteria	Café	Coffee Shop	Gaming Cafe	Hospital	Pharmacy	College Arts Building	College Auditorium	College Rec Center	College Stadium	Climbing Gym
7	7	7	7	10	10	9	9	9	9	6



Gym	Gym Pool	Gym / Fitness Center
6	6	6

6.3 Filter dataset to reduce memory consumption. The metrics of the retrieved results from FourSquare were reduced (dimensions were reduced) based on customer's expectation, which became a smaller dataset as shown in Table 6.

Table 6. Filtered dataset.



	Cafeteria	Café	Coffee Shop	Gaming Cafe	Hospital	Pharmacy	College Arts Building	College Auditorium	College Rec Center	College Stadium
0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
1	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
2	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
3	0.0	0.0	0.666667	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
4	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...
94	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
95	0.0	0.0	0.142857	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
96	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
97	0.0	0.0	0.000000	0.0	0.0	0.111111	0.0	0.0	0.0	0.0
98	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.0

99 rows × 15 columns

Please see the full table in Notebook. This is a incomplete sreenshot.

6.4 Customized weighting of the venues

First, customer rating score was transformed to a relative rating score by taking the relative significance of each metric to the customer. Then the actual frequency of occurrence of the venues retrieved from FourSquare for each postal code area was scaled using the customized weighting scheme. Then the sum of the weighted frequency of each metric was calculated for each postal code area. The sum is used as the score to rank the venues. Please see the notebook for details.

6.5 Then a recommendation is generated for the customer. The top 10 areas were visualized on the map as shown in Figure 2. In this figure, the size of the circle represents the significance/rank of the area. The area with the highest score has the biggest size to guide the eyes.

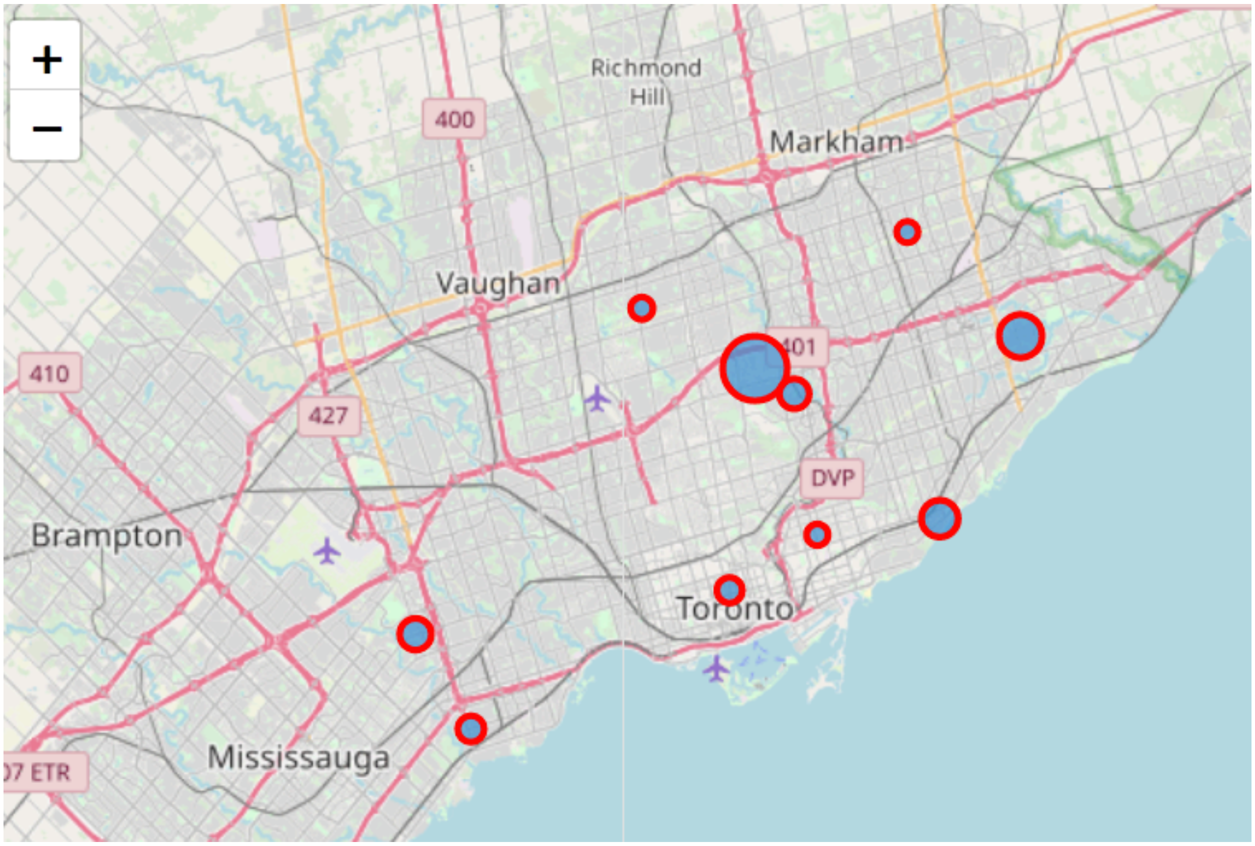


Figure 2. Mapping the top 10 recommended areas in Toronto neighborhoods

Conclusion and Discussion

The top 10 recommended areas for the customer have the postal code M2L, M1G, M1N, M9C, M3B, M8W, M7A, M2R, M1V and M4J. When using texts as the information to build recommendation, it is important to clarity ambiguity. For example, in this project, Café was further expanded to Coffee Shop, Gaming Café and Cafeteria; health care was expanded to Pharmacy and Hospital; gym was expanded to Climbing Gym, Gym Pool, Fitness Center. If these alternative words were not considered, we could potentially underestimate the significance of some neighborhoods. Also, we need to adjust the weight of each venue based on the preference of the customer. This process helps create a more customized recommendation. There are also limitations of this project. For example, the customer did not provide more details of ratings of other important metrics such as price or house type. Also, if the customer could be more specific about the expectation of each metric, we could make the recommendation engine more precise.