Final Report:Coursera Capstone Project

Introduction/Business Problem:

In the fast moving, effort-intense environment that a student inhabits,It's a frequent occurrence that one is too tired to fix oneself a home-cooked meal. And of course, even if one gets home-cooked meals every day, it is not unusual to want to go out for a good meal every once in a while for social/recreational purposes. Either way, it's a commonly understood idea that regardless of where one lives, the food one eats is an important aspect of the lifestyle one leads.

Now, imagine a scenario where a person has newly moved into a new location. They already have certain preferences, certain tastes. It would save both the student and the food providers a lot of hassle if the student lived close to their preferred outlets. Convenience means better sales, and saved time for the customer.

Food delivery apps aside, managers of restaurant chains and hotels can also leverage this information. For example, if a manager of a restaurant already knows the demographic of his current customers, they'd ideally want to open at a location where this demographic is at its highest concentration, ensuring short commute times to the location and more customers served.If potential hotel locations are being evaluated, a site that caters to a wide variety of tastes would be ideal, since one would want every guest to have something to their liking.

Since the data is for non-Indian students, Indian universities can easily leverage this information to help better accommodate international students of varied tastes as well.

Essentially, the problem statement is this:

"How can one identify trends in an individual's* daily routine and leverage them?"

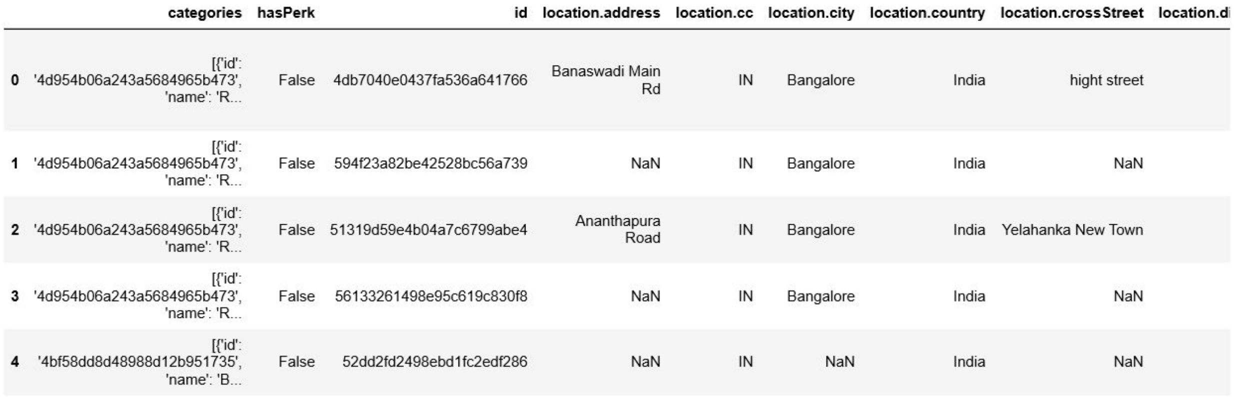
*Here, a college student

Data

I used a variety of sources, primary among them, the foursquare Database via the API:

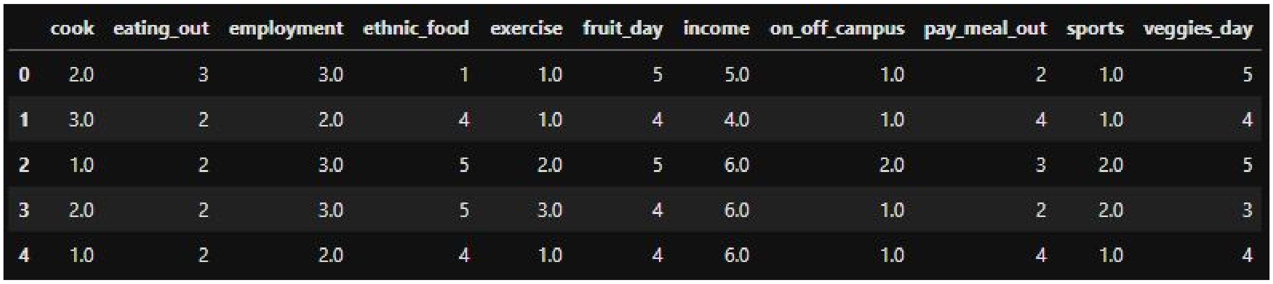
As can be seen on the next page, it was very messy, and not usable out of the box. A lot of cleaning needed to be done before it was in a usable state.

Another Dataset I used was “Food choices” by BoraPajo. Since this dataset was assembled by someone already involved in Data Science (being hosted on Kaggle after all), this dataset was much easier to use, and I simply needed to slice the data relevant to me.



	categories	hasPerk	id	location.address	location.cc	location.city	location.country	location.crossStreet	location.di
0	[{'id': '4d954b06a243a5684965b473', 'name': 'R...	False	4db7040e0437fa536a641766	Banaswadi Main Rd	IN	Bangalore	India	hight street	
1	[{'id': '4d954b06a243a5684965b473', 'name': 'R...	False	594f23a82be42528bc56a739	NaN	IN	Bangalore	India	NaN	
2	[{'id': '4d954b06a243a5684965b473', 'name': 'R...	False	51319d59e4b04a7c6799abe4	Ananthapura Road	IN	Bangalore	India	Yelahanka New Town	
3	[{'id': '4d954b06a243a5684965b473', 'name': 'R...	False	56133261498e95c619c830f8	NaN	IN	Bangalore	India	NaN	
4	[{'id': '4bf58dd8d48988d12b951735', 'name': 'B...	False	52dd2fd2498ebd1fc2edf286	NaN	IN	NaN	India	NaN	

Pictured above: A snapshot of the raw foursquare data



	cook	eating_out	employment	ethnic_food	exercise	fruit_day	income	on_off_campus	pay_meal_out	sports	veggies_day
0	2.0	3	3.0	1	1.0	5	5.0	1.0	2	1.0	5
1	3.0	2	2.0	4	1.0	4	4.0	1.0	4	1.0	4
2	1.0	2	3.0	5	2.0	5	6.0	2.0	3	2.0	5
3	2.0	2	3.0	5	3.0	4	6.0	1.0	2	2.0	3
4	1.0	2	2.0	4	1.0	4	6.0	1.0	4	1.0	4

BoraPajo’s Food Choices Dataset, after trimming

To clean the Foursquare data to a usable state, I dropped all irrelevant entries like categories, hasPerk, etc. and retained only information like Location Name, Address, and its latitude and longitude. The problem with Foursquare data is that it is very finicky to call the API: sometimes the queries would return 2, sometimes 50 locations depending on the search word. The results varied wildly with minute changes.

I had to identify the top categories that represented the Food and Grocery categories, as well as restaurants which were amusingly overshadowed by bus stops in the initial queries.

The main challenge with BoraPajo’s dataset was initially making sense of the coded values, but the attached codebook was a lifesaver, and eventually the values became intuitive: Values were always graded, it was simply a matter of understanding whether they were sorted low-high or high-low.

To get a general idea of the dataset, a brief explanation of the most important parameters follows:

employment – do you work?
1 - yes full time
2 - yes part time
3 – no
4 - other

eating_out - frequency of eating out in a typical week
1 - Never
2 - 1-2 times
3 - 2-3 times
4 - 3-5 times
5 - every day

income
1 - less than $15,000
2 - $15,001 to $30,000
3 - $30,001 to $50,000
4 - $50,001 to $70,000
5 - $70,001 to $100,000
6 - higher than $100,000

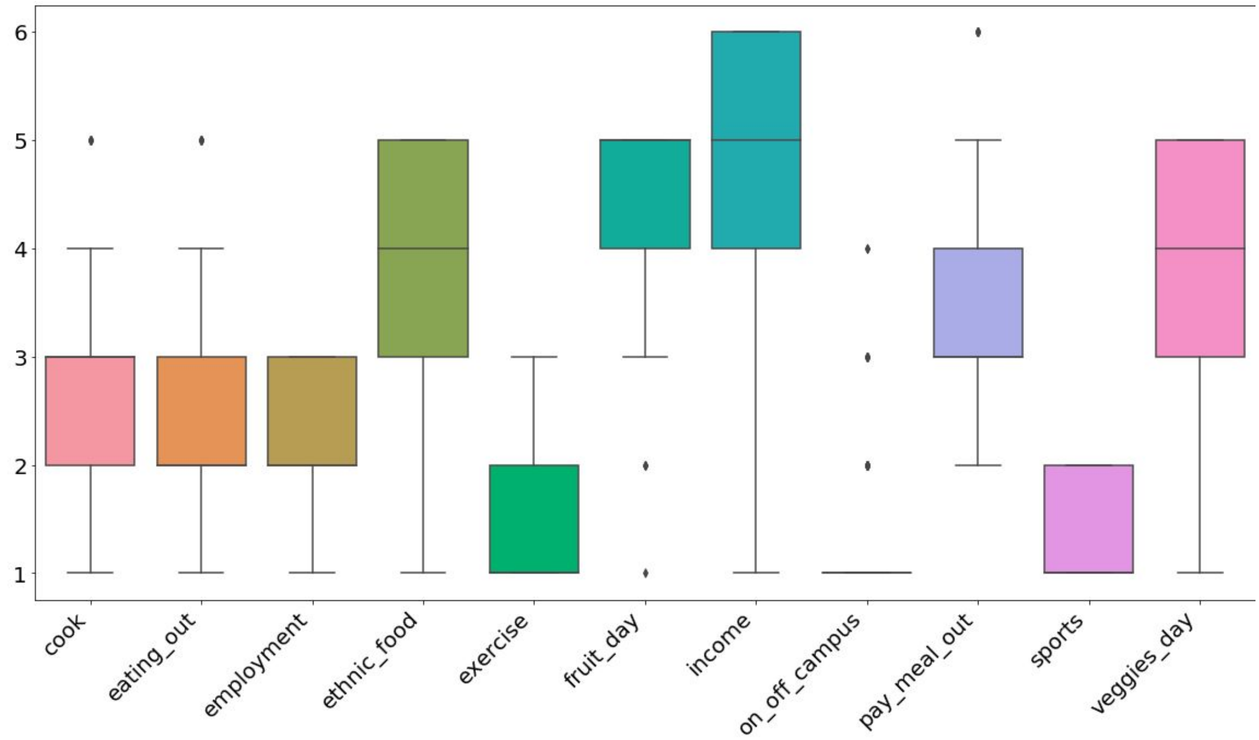
cook – how often do you cook?
1 - Every day
2 - A couple of times a week
3 - Whenever I can, but that is not very often
4 - I only help a little during holidays
5 - Never, I really do not know my way around a kitchen

how often do you exercise in a regular week?
1 - Everyday
2 - Twice or three times per week
3 - Once a week
4 - Sometimes
5 – Never

I worked with a relatively small set, as after cleaning the NaN values I was left with around 100 students to work with. In essence, I worked with about 50 locations in the foursquare database, but queried for around 2000 while checking the fit of accommodation for the students.

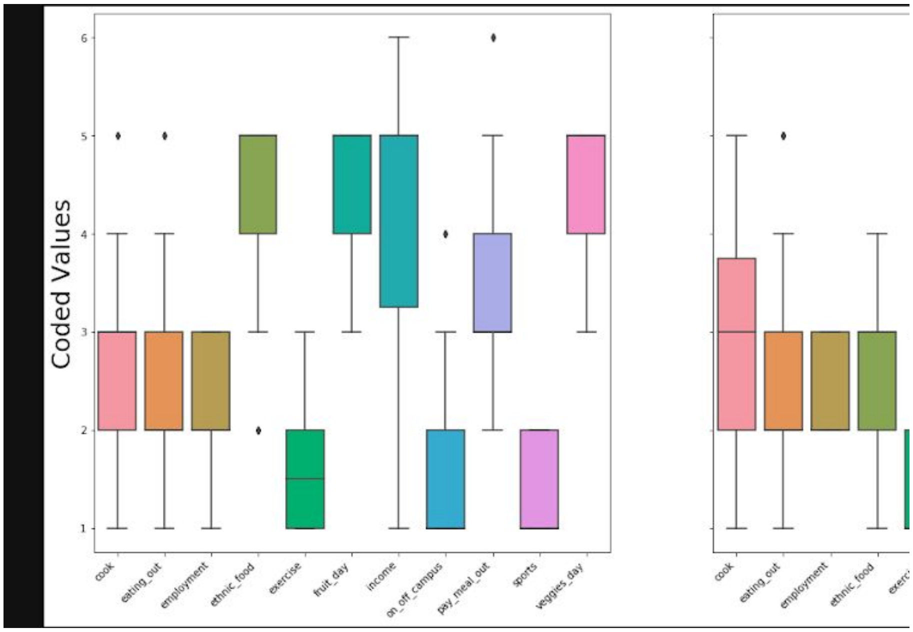
Methodology

K-means clustering is an ideal algorithm to identify unknown trends, so it is best suited to our problem. After all, by just this box-plot, we can't decide on a factor which would help discern students:



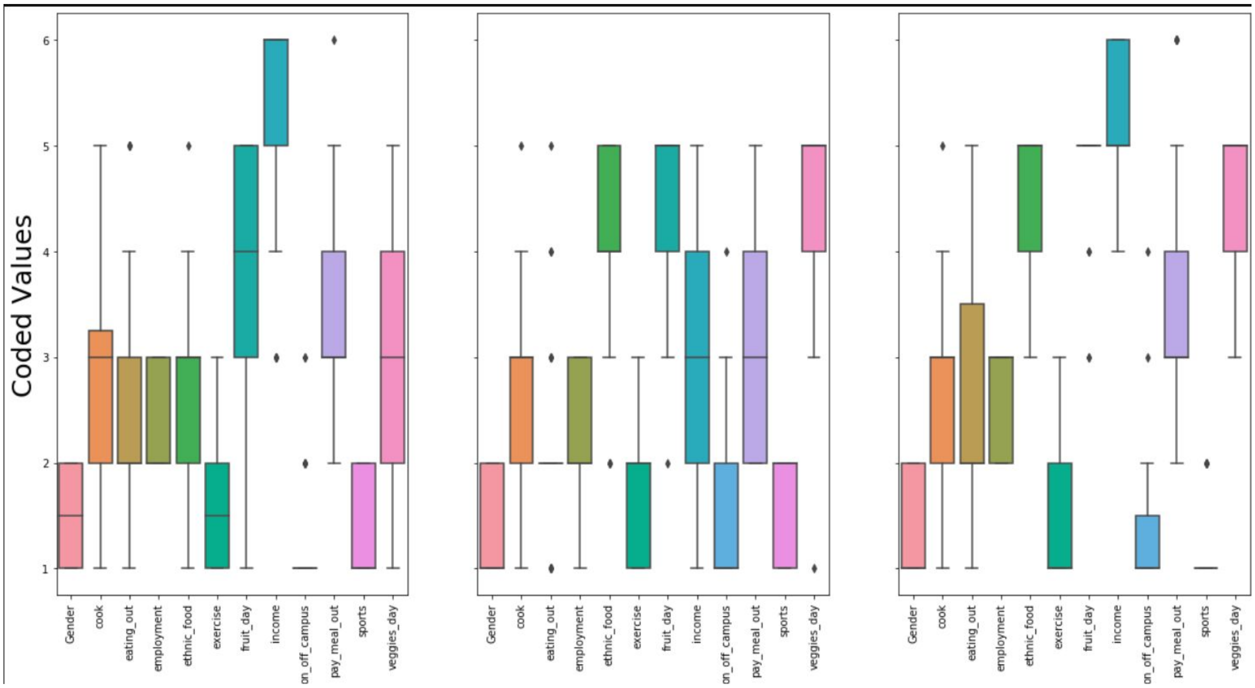
I initially also considered gender, but ended up dropping it for the final analysis as gender turned out to be a non-factor: The findings were equally true for both males and females. From the above plot, we can infer that in general, students tend to cook now and then, and eat out the days they don't. Ethnic food is enjoyed by a wide host of students, and nearly everyone exercises daily. Fruits and vegetables figure pretty high on the food list for students, and income is high for most. A very high number of people stay on campus, and on average, students are willing to pay roughly $10 for a meal.

I proceeded with K-means clustering by initially employing only 2 clusters:



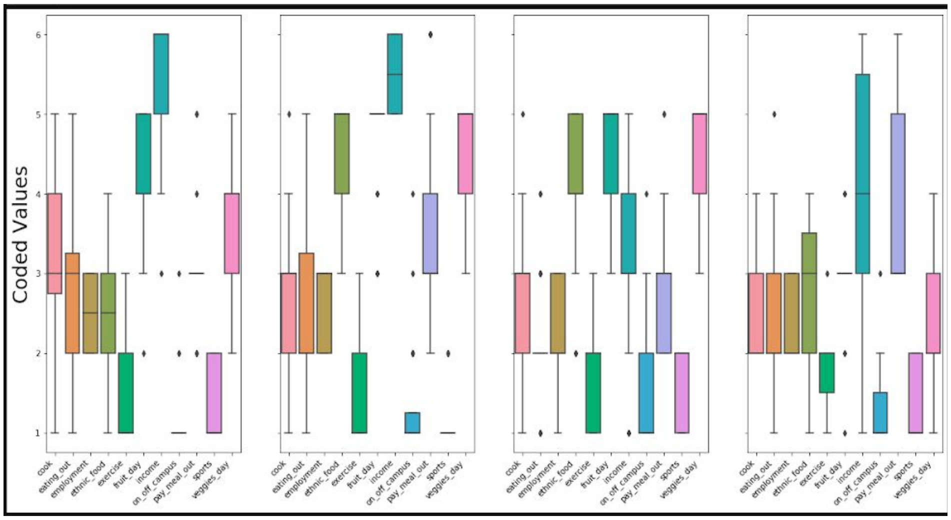
But this clustering didn’t yield ideal results; There wasn’t any clear split among the two groups.

Next, with 3 clusters:



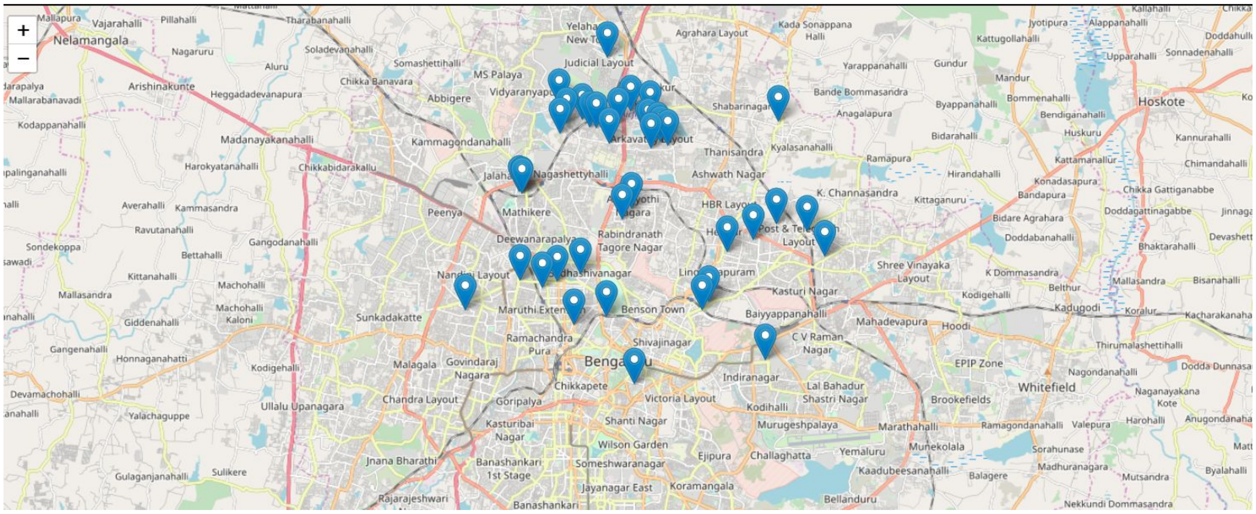
Now we begin to see a clear split: Students can be split on income!

Now, with 4 clusters:

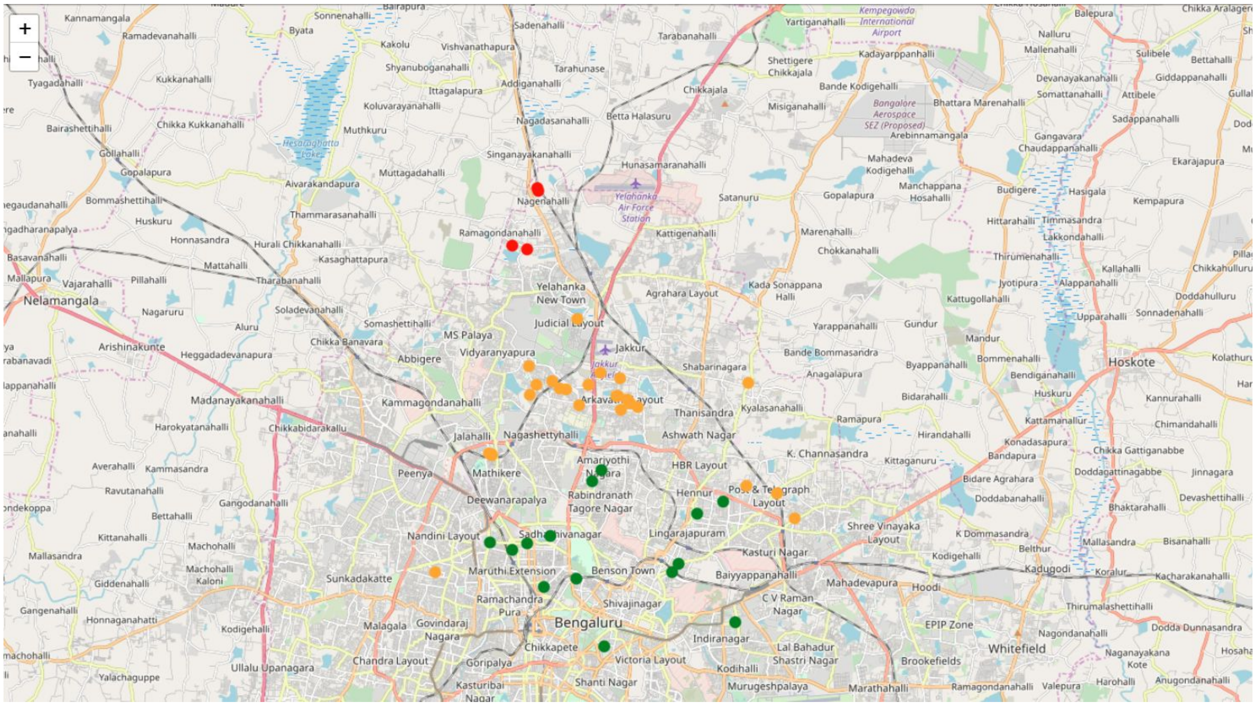


From here on out, the clusters again became very hard to analyse and find clear boundaries to differentiate on.

After employing K-means on the students, I found out all the potential locations students could stay (that were reasonably close to the college of study)



Next, I applied K-means to cluster these locations into 3 clusters:



Results:

K-means Clustering on Student Data

- One cluster of high income students seem to eat out more often and spend more per meal, and care less about fruits, vegetables and ethnic food, and stay almost exclusively on campus.

- The second cluster of high income students seem to eat out less, pay less per meal, and are more likely to stay off campus.

- The cluster of low income students eat out less, and cook more often than the high income group. They eat as much vegetables as the second cluster but eat less fruits (perhaps because fruits are more expensive than vegetables?) and are most likely to stay off campus.

K-means Clustering on Location Data

Three prominent clusters emerged after applying the method on the data:

- Cluster 0(Green) Where both (fruits and vegetables) and (restaurants) are abundant
- Cluster 1(Yellow): Restaurants are plentiful, but groceries less so.
- Cluster 2(Red): Restaurants and groceries are relatively hard to find.

Discussion

Ideally, students should be maximised at the Green(Cluster 0) locations since both kinds of students can be catered to there, and obviously, unless renting their own house, it's very difficult to open a new housing for just a few students!

Another aspect to think of is cost. One can easily notice, the further away from the college and the closer to the city centre one gets, the more options one finds for food.The same can be said about other amenities as well.The closer to the city centre, the more expensive property gets, as well as the cost of living. Therefore, in reality, Cluster 1 locations might be better value for money.

Finally, Cluster 2 locations, while not ideal, offer the shortest travel times to college, and may be viable for students willing to compromise on food or making alternative arrangements. With the advent of food delivery apps, it is quite easy to get both groceries and prepared meals both, so there might be a few locations which could be classified as Yellow or Green depending on coverage.

One thing I would like to note is that the foursquare data seems incomplete; Many locations seem to be missing or ill-classified. India definitely needs better locational data sets!

Conclusion

K-means clustering can help find patterns where none are apparent. Income is a very useful factor to classify students' behaviour and spending patterns.Diet is another important factor to account for when looking at accommodation for students. Foursquare data is limited but can provide insights into a city's infrastructure. This data could be supplemented with other sources to provide better results.