Neighborhood factors affecting Fuel Station sales in Tirupati

Nanduri Dinesh Kumar

07-12-2020

1. Introduction

1.1 Business Idea

The idea of this project is to explore the neighborhoods of all fuel filling stations in Tirupati(City in India) using Foursquare to analyse how the presence of different neighborhoods affect the fuel sales of each fuel station. A machine learning model can be created with the help of which any future project proposals to setup a fuel station in Tirupati can have its yearly sales figures predicted based on the neighborhood it is being set up in.

1.2 Who would be interested

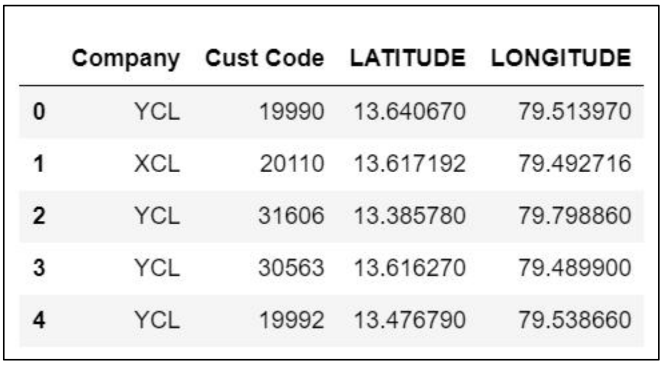
The target audience are the stakeholders, the fuel station dealers and Oil companies who with the help of this model can get an estimate of the return on their investments before starting the project.

2. Data

The data being used is the Sales and Co-ordinates of all the Fuel Stations of 3 Oil Companies - X Corporation Limited(XCL), Y Corporation Limited(YCL), Z Corporation Limited(ZCL). As the data has been taken from a closed source, for confidentiality purpose I have given dummy names to the Companies in question.

There were five ‘csv’ files from which data had been read. ZCL Fuel stations co-ordinates were contained in one file and then XCL,YCL fuel stations co-ordinates in another. These files were filled with a lot of columns like address, district, divisional office being reported to, personal details of the current fuel station owner/dealer, facilities available etc. These files were individually cleaned by removing unnecessary columns, data and missing values to finally result in a precise data required i.e. Company name, Fuel station code to uniquely identify them when plotted on map, and their respective Lat-Long co-ordinates.

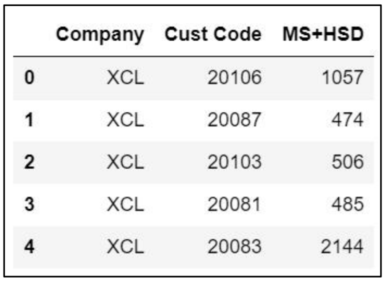
2.1. List of fuel filling stations in Tirupati with their Name, Code, Latitude and Longitude



	Company	Cust Code	LATITUDE	LONGITUDE
0	YCL	19990	13.640670	79.513970
1	XCL	20110	13.617192	79.492716
2	YCL	31606	13.385780	79.798860
3	YCL	30563	13.616270	79.489900
4	YCL	19992	13.476790	79.538660

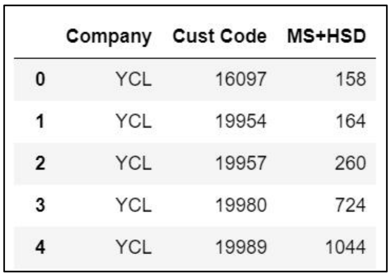
2.2. List of Fuel Filling stations with 2019-2020 Sales Figures in Kilo Liters

XCL:



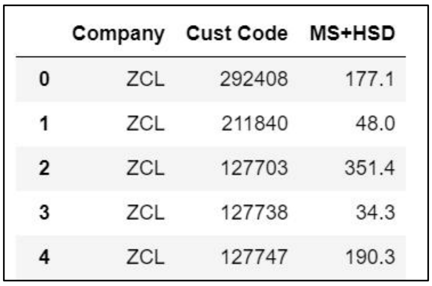
	Company	Cust Code	MS+HSD
0	XCL	20106	1057
1	XCL	20087	474
2	XCL	20103	506
3	XCL	20081	485
4	XCL	20083	2144

YCL:



	Company	Cust Code	MS+HSD
0	YCL	16097	158
1	YCL	19954	164
2	YCL	19957	260
3	YCL	19980	724
4	YCL	19989	1044

ZCL:

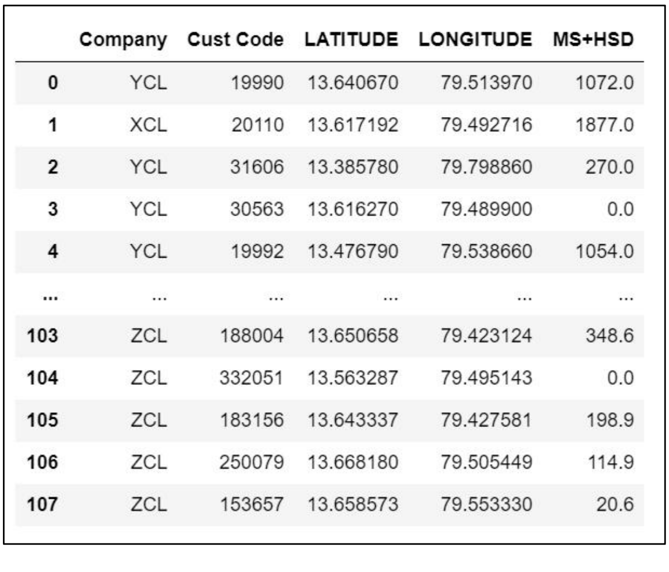


	Company	Cust Code	MS+HSD
0	ZCL	292408	177.1
1	ZCL	211840	48.0
2	ZCL	127703	351.4
3	ZCL	127738	34.3
4	ZCL	127747	190.3

The other 3 tables contained the Sales of each Fuel station of each company in three different files. I had filtered the data from all the files based on the Fuel stations reporting to Tirupati Divisional office. Using this I was able to make a subset from the database results. These files also required a lot a of Data cleaning as the sales data was taken directly from the company database and it contained the sales figures and different types of Sales figures like Cumulative sales of MS+HSD, Annual avg sales of MS, HSD and MS+HSD, Sales figures per year basis and per month basis from 2016-2020. For this, to make things easy, I directly deleted the columns from the main csv files and kept only- Company name, Fuel Station code and MS+HSD figures of 2019-2020, before uploading it to the Jupyter notebook.

This gave me a clear picture of how to create a master table with the data I require. After uploading I also Found out that few of the Fuel station data was missing and hence required some data cleaning before merging the tables together.

All these table are then combined to form a master table containing - Company Name, Code, Latitude, Longitude and MS+HSD(Motor Spirit + High speed Diesel Sales figure in Kiloliter)



	Company	Cust Code	LATITUDE	LONGITUDE	MS+HSD
0	YCL	19990	13.640670	79.513970	1072.0
1	XCL	20110	13.617192	79.492716	1877.0
2	YCL	31606	13.385780	79.798860	270.0
3	YCL	30563	13.616270	79.489900	0.0
4	YCL	19992	13.476790	79.538660	1054.0
...	...	...	...	...	...
103	ZCL	188004	13.650658	79.423124	348.6
104	ZCL	332051	13.563287	79.495143	0.0
105	ZCL	183156	13.643337	79.427581	198.9
106	ZCL	250079	13.668180	79.505449	114.9
107	ZCL	153657	13.658573	79.553330	20.6

Once the Master Data frame was ready, I plotted these on the map to get a visual understanding of the positions of the current Fuel Stations

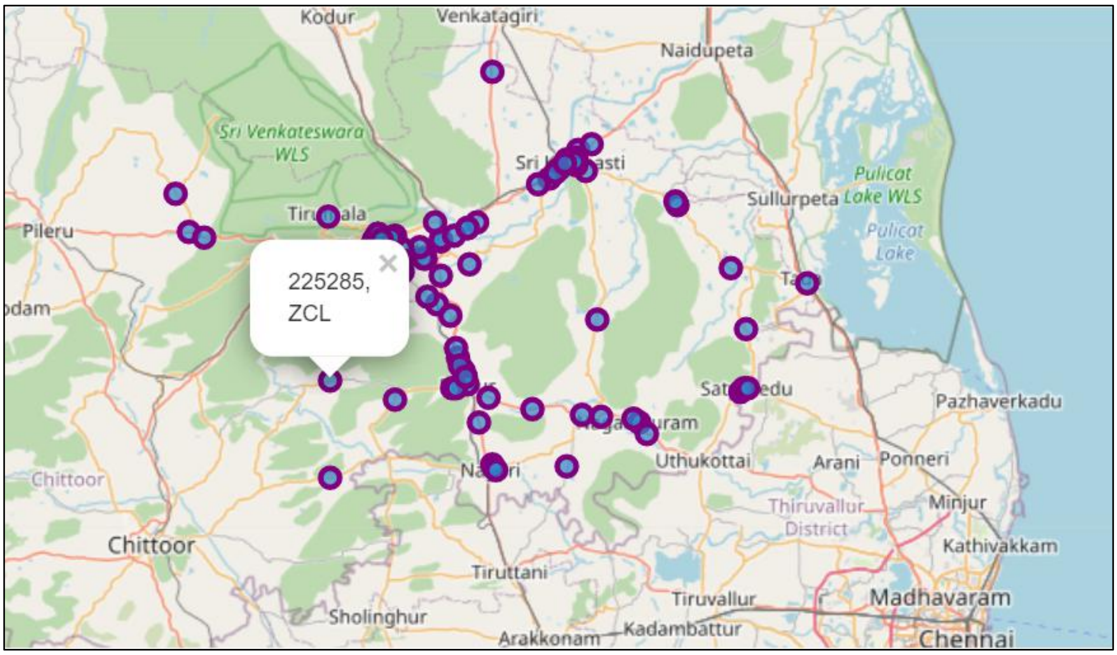
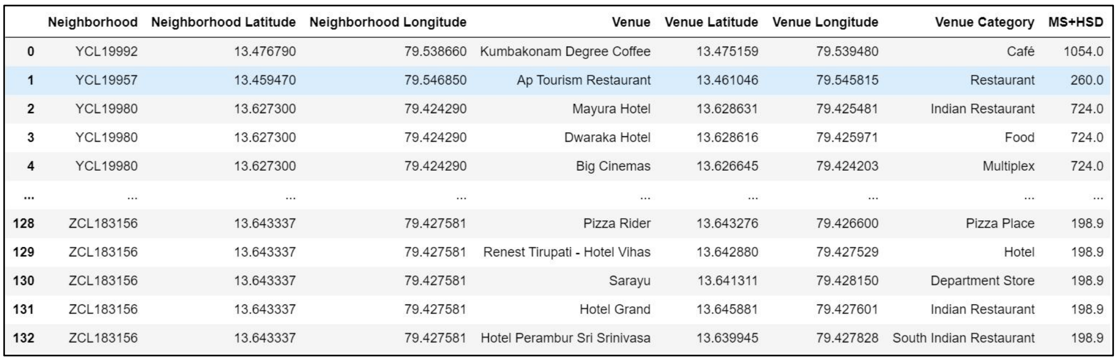


Fig 1. All fuel stations of Tirupati marked on Tirupati city map.

Observing the map I could see few Fuel station which were outside the city were also mapped. But that was fine since on finding the neighborhood venues based on Tirupati Co-ordinates on the map would eliminate them in the future tables.

3. Finding Neighborhoods

The main objective of this project is to find the neighborhoods of all the fuel stations mapped and based on the types of neighborhoods correlate it to the sales made by the fuel stations. Before finding the neighborhoods, I merged the 'Company' and 'Cust Code' columns into one for easy identification of fuel stations from one columns. Now to find the neighborhood venues and their categories for each fuel station, I used Foursquare application. The application returned the neighborhood venues for each fuel stations along with its categories.



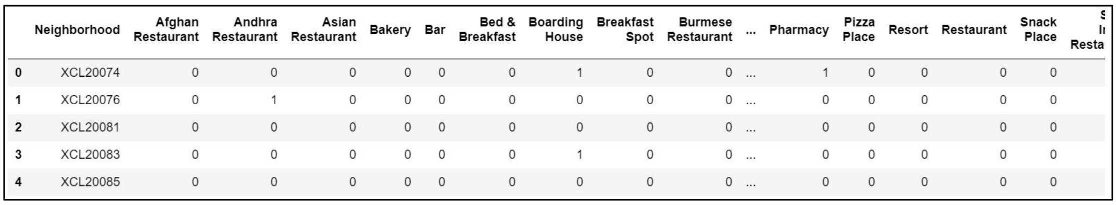
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	MS+HSD
0	YCL19992	13.476790	79.538660	Kumbakonam Degree Coffee	13.475159	79.539480	Café	1054.0
1	YCL19957	13.459470	79.546850	Ap Tourism Restaurant	13.461046	79.545815	Restaurant	260.0
2	YCL19980	13.627300	79.424290	Mayura Hotel	13.628631	79.425481	Indian Restaurant	724.0
3	YCL19980	13.627300	79.424290	Dwaraka Hotel	13.628616	79.425971	Food	724.0
4	YCL19980	13.627300	79.424290	Big Cinemas	13.626645	79.424203	Multiplex	724.0
...	...	...	...	...	...	...	...	...
128	ZCL183156	13.643337	79.427581	Pizza Rider	13.643276	79.426600	Pizza Place	198.9
129	ZCL183156	13.643337	79.427581	Renest Tirupati - Hotel Vihas	13.642880	79.427529	Hotel	198.9
130	ZCL183156	13.643337	79.427581	Sarayu	13.641311	79.428150	Department Store	198.9
131	ZCL183156	13.643337	79.427581	Hotel Grand	13.645881	79.427601	Indian Restaurant	198.9
132	ZCL183156	13.643337	79.427581	Hotel Perambur Sri Srinivasa	13.639945	79.427828	South Indian Restaurant	198.9

I then found the number of Unique categories which was equal to 45.

4. One-hot encoding and frequency mapping

Using the 45 venue categories, I performed one-hot encoding for each fuel station. This resulted in many rows for the same Fuel station due to the presence of 2 or more venues of the same venue category around a Fuel station. For example, there were more 4 different Indian restaurants present around one YCL Fuel station.

Now, in order to get distinct columns, I made the table to return the frequency of occurrence of each unique category for each fuel station from the one-hot encoding dataframe. I then merged the Fuel sales figures with the resultant Fuel stations.



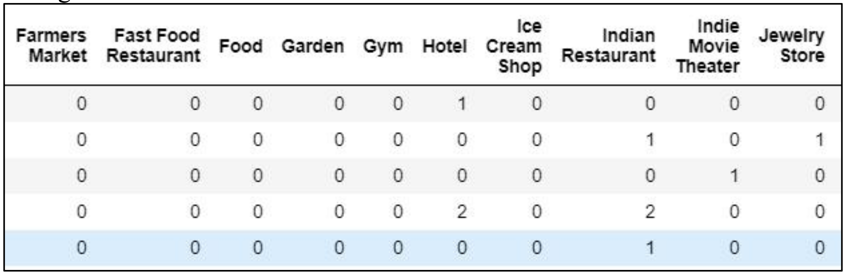
	Neighborhood	Afghan Restaurant	Andhra Restaurant	Asian Restaurant	Bakery	Bar	Bed & Breakfast	Boarding House	Breakfast Spot	Burmese Restaurant	...	Pharmacy	Pizza Place	Resort	Restaurant	Snack Place	S I Resta
0	XCL20074	0	0	0	0	0	0	1	0	0	...	1	0	0	0	0	
1	XCL20076	0	1	0	0	0	0	0	0	0	...	0	0	0	0	0	
2	XCL20081	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
3	XCL20083	0	0	0	0	0	0	1	0	0	...	0	0	0	0	0	
4	XCL20085	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	

From a quick view at the table and finding the sum of frequencies for each unique venue category we can see that the the most popular venues near Fuel stations are - Boarding house, Cafe, Department Store, Hotel and Indian restaurant and the Sales of Fuel stations near these venues are High.

5. Machine Learning Model

5.1 Processing Data

To create a model we need the X and Y values - the independent and dependent variables respectively. I assigned all the 45 features to X.

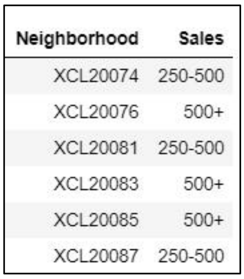


Farmers Market	Fast Food Restaurant	Food	Garden	Gym	Hotel	Ice Cream Shop	Indian Restaurant	Indie Movie Theater	Jewelry Store
0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	1	0	1
0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	2	0	2	0	0
0	0	0	0	0	0	0	1	0	0

For Y, I classified the Sales figures into 4 categories to get a better estimate. The four categories are:

a. 0-100 kl
b. 100-250 kl
c. 250-500 kl
d. 500+ kl

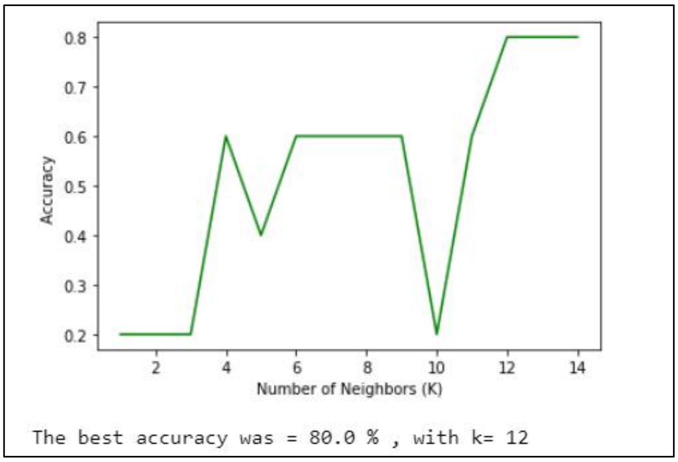
These were replaced in the Sales(MS+HSD) column for each fuel station and assigned to Y, The dependent attribute.



Neighborhood	Sales
XCL20074	250-500
XCL20076	500+
XCL20081	250-500
XCL20083	500+
XCL20085	500+
XCL20087	250-500

5.2 K-Nearest Neighbor Algorithm

Since the entire project is based on neighborhoods and the predicted value is calculated based on neighborhoods, I chose the K- nearest neighbor algorithm to train and test the model. I obtained an accuracy of 80% which I felt was satisfactory to give an estimate for a future fuel station based on its neighborhood data. The K with the highest accuracy was K= 12.



```
The best accuracy was = 80.0 % , with k= 12
```

Fig 2. Finding the K value which returns the best accuracy for the model

The model was trained with K=12 and was then used on a completely new Data set to observe the results.

5.3 New Dataset and Results

I chose 4 different locations: FS-A, FS-B, FS-C, FS-D, on Tirupati map for which I wanted the model to predict the Yearly Sale of MS+HSD based on its neighborhoods.

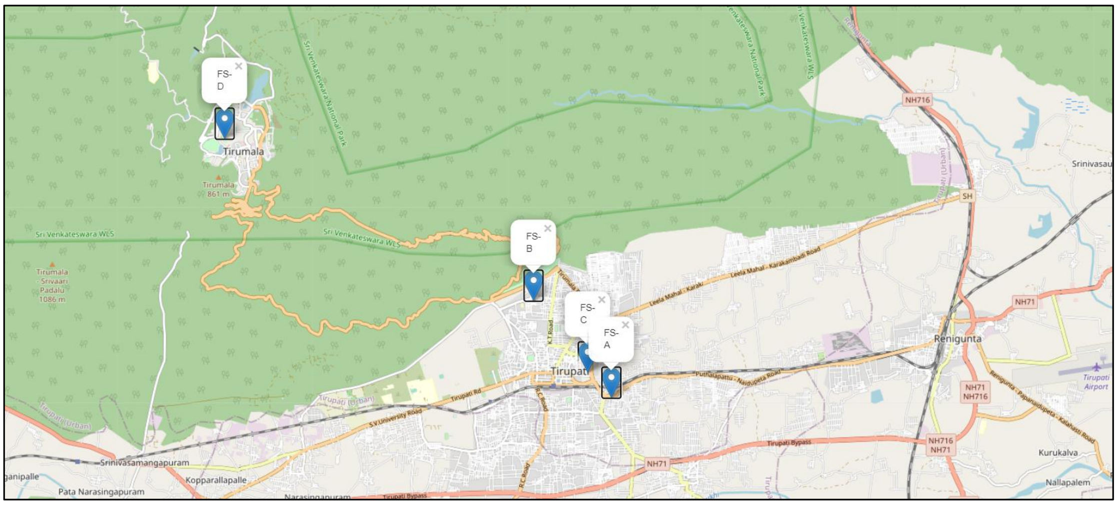
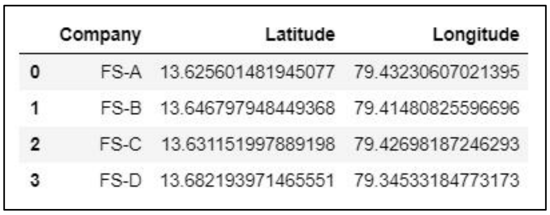


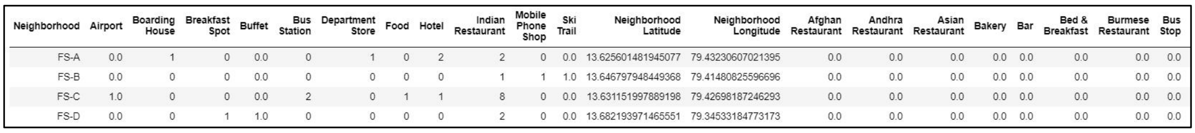
Fig3. New Locations in Tirupati where Fuel stations are to be set up

Co-ordinates of new Fuel Station Locations



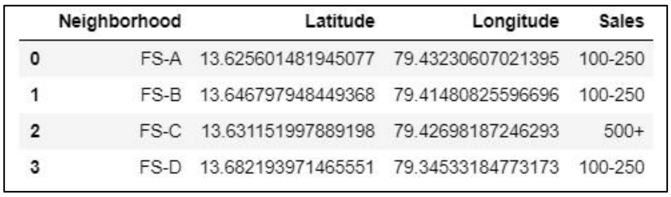
	Company	Latitude	Longitude
0	FS-A	13.625601481945077	79.43230607021395
1	FS-B	13.646797948449368	79.41480825596696
2	FS-C	13.631151997889198	79.42698187246293
3	FS-D	13.682193971465551	79.34533184773173

Using the above process, I found the neighborhoods of the 4 new locations and one hot coded them. I then applied frequency mapping on the same. I had prepared a default data set of all 45 venue categories which we had used before and gave them a default one-hot encoding value = 0. With 2 encoded data frames now(one from the new locations and one from the default data set) I merged both of them retaining the values from the new locations and appending the rest of the venue categories with zero value.



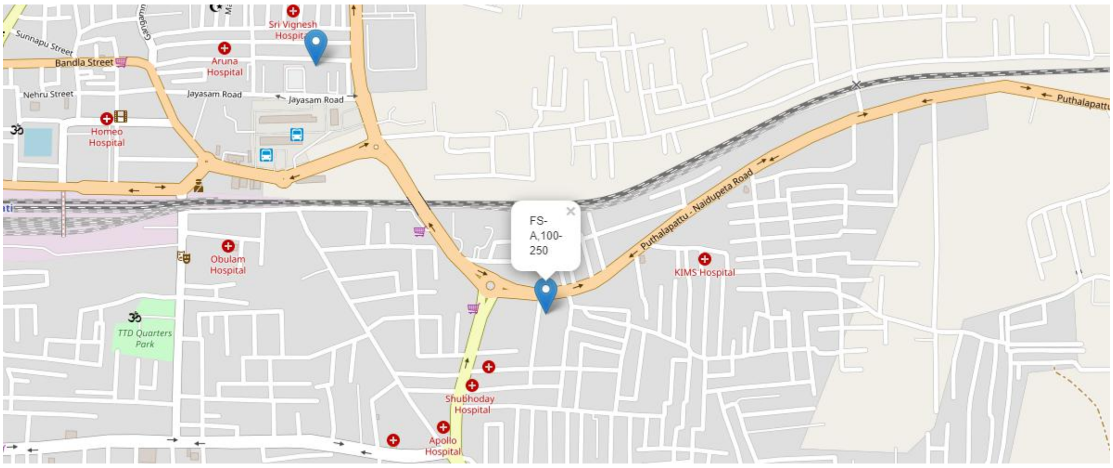
Neighborhood	Airport	Boarding House	Breakfast Spot	Buffet	Bus Station	Department Store	Food	Hotel	Indian Restaurant	Mobile Phone Shop	Ski Trail	Neighborhood Latitude	Neighborhood Longitude	Afghan Restaurant	Andhra Restaurant	Asian Restaurant	Bakery	Bar	Bed & Breakfast	Burmese Restaurant	Bus Stop
FS-A	0.0	1	0	0.0	0	1	0	2	2	0	0.0	13.625601481945077	79.43230607021395	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
FS-B	0.0	0	0	0.0	0	0	0	0	1	1	1.0	13.646797948449368	79.41480825596696	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
FS-C	1.0	0	0	0.0	2	0	1	1	8	0	0.0	13.631151997889198	79.42698187246293	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
FS-D	0.0	0	1	1.0	0	0	0	0	2	0	0.0	13.682193971465551	79.34533184773173	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Upon selecting the features set as X, I set it as input to the KNN model previously developed and Voila! We got the Sales Estimates for each new Fuel Station.

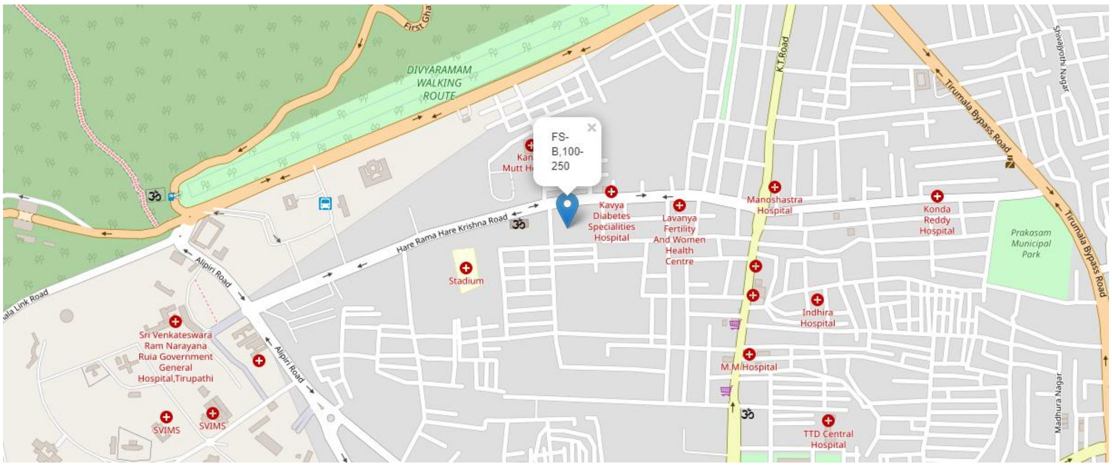


	Neighborhood	Latitude	Longitude	Sales
0	FS-A	13.625601481945077	79.43230607021395	100-250
1	FS-B	13.646797948449368	79.41480825596696	100-250
2	FS-C	13.631151997889198	79.42698187246293	500+
3	FS-D	13.682193971465551	79.34533184773173	100-250

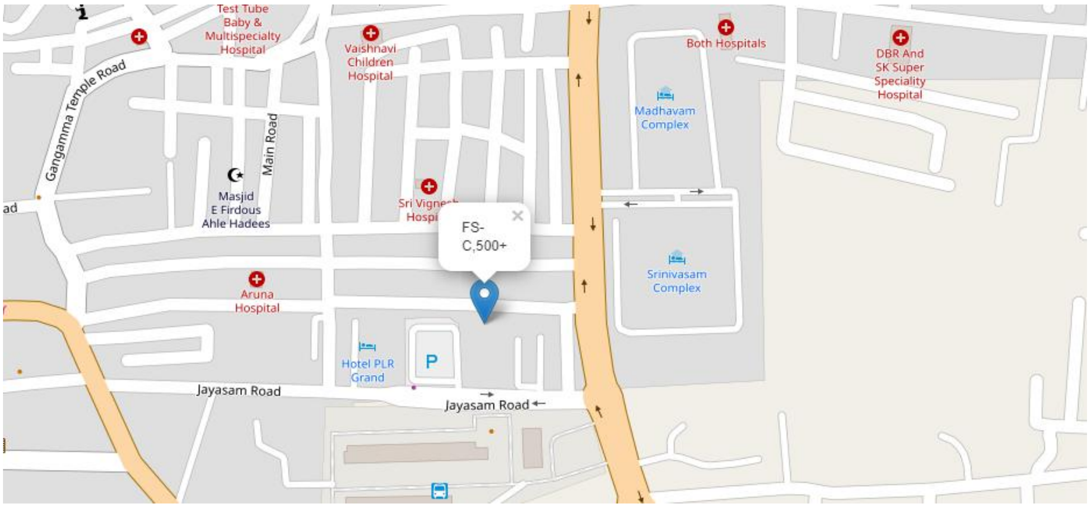
Fuel Station - A (FS-A)



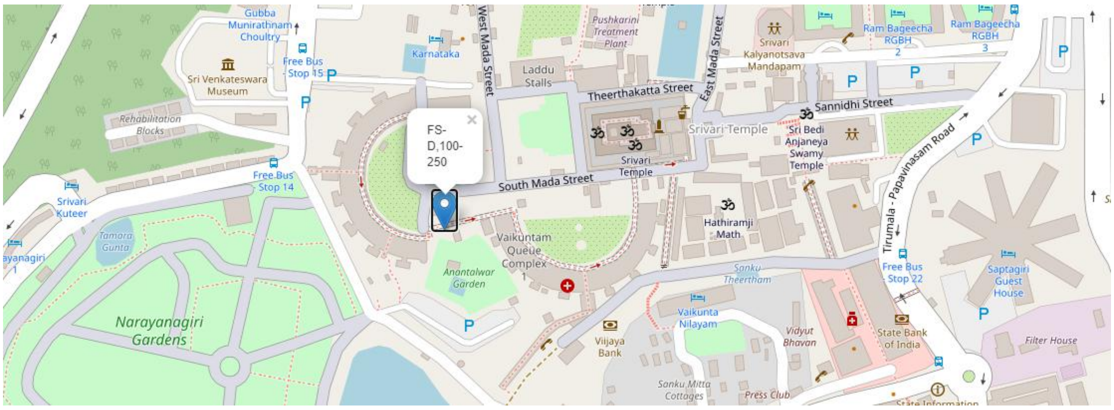
Fuel Station - B (FS-B)



Fuel Station - C (FS-C)



Fuel Station - D (FS-D)



6. Discussion

We can observe that for FS-C, the neighborhood contains many Indian Restaurants which played a major factor in predicting higher sales for the Fuel station in Tirupati. From observing the Sales figures and the venues extracted for each neighborhood, I can say that the sales tend to be higher when a Fuel station is placed near Hotels, Boarding houses since they invite more tourists and vehicles- Car, jeep, van, bus - arranged for tourist trips incline to fill fuel near the place of pickup before starting the journey and in many cases the payment is also made by the tourist instead of the vehicle driver. This can be one of the reason why sales are higher. So for future location scouts, it is beneficiary if these neighborhoods can be considered.

For Low sale Fuel Stations, the neighborhoods are venues with low footprint like- Jewelry store, Small cafes/ice cream shops, Gym, Local Business services, where people usually prefer to walk rather than travelling in a vehicle. These localities are common around many neighborhoods which eliminates the need to travel in a vehicle to reach these venues. Hence these for future location scouts these venues are a red flag.

7. Conclusion

The model was based on the neighborhood factors of each Fuel station. I was able to achieve an 80% accurate model which thus can be very helpful in selecting land sites/plots for Fuel stations. However, this maybe applies to a city. This can be further improved and refined with more factors like the traffic congestion of roads near the Fuel stations and facilities available at each fuel station like - Car cleaning, Washrooms, Air pumps etc. Another factor are types of vehicles around the area, For example - if a locality is in the outskirts of the city where mostly only large vehicles like trucks and buses visit the fuel station, the sale of diesel may be very high compared to the sale of petrol which might be bringing down the overall value of yearly sales compared to other Fuel stations.

Overall I felt this is a good use of data science and predictive analysis which can play a major role in decision making on an investment.