Applied Data Science Capstone

Gyan Prakash Singh

February 2020

Contents

1 Introduction **3**
1.1 Background 3
1.2 Problem 3

2 Data acquisition and cleaning **3**
2.1 Data sources 3
2.2 Data acquisition 4
2.3 Data Processing 4

3 Methodology **6**
3.1 Data exploration 6

4 Results **7**

5 Discussion **8**

6 Conclusion **9**
6.1 Three neighbourhoods in Delhi for establishing a fine dining restaurant 9

1 Introduction

1.1 Background

Using data to solve is a business problem is one of the things which makes data science useful. Deciding the location of a new business venture is one of the crucial decisions which can have critical consequences for the future of a business. However, the factors which affect the suitabIlity of a location for a new business venture vary and as businesses have unique requirements to ensure success.

In this report, we will explore various data sets to help potential entrepreneurs who want to establish a fine dining restaurant in Delhi, India. Delhi is the capital of India and a very populous city. People of Delhi are known for their love of food which makes Restaurants flourish in general. However, with growth in income, tastes of people are getting more refined and it is making a new restaurant venture a challenging task.The task is particularly more difficult for fine dining restaurants. With high income inequality combined with high average income compared to rest of the country, finding a suitable location for a fine dining restaurant in delhi is a challenging but most crucial task in planning a new restaurant business.

1.2 Problem

Core problem is use of data science tools to find 3 candidate locations for setting up a fine dining restaurant in delhi. Factors which may lead to higher footfall such as number of high quality fine dining restaurants in the locality, other landmarks bazaars, museums, art galleries, shopping centers, parking lots etc. need to be identified and explored to make a suitable model which may solve this problem.

2 Data acquisition and cleaning

2.1 Data sources

Foursquare location data will be utilized as core of our analysis. It will be used to extract information such as nearby venues of interest. Based on the definition of the problem, data points which I would like to explore are:

- number of existing restaurants in the neighborhood (any type of restaurant)

- number of and distance to fine dining restaurants in the neighborhood, if any
- distance of neighborhood from major landmarks such as museums, shopping centers, art galleries etc.
- distance from coparatively affluent residential localities of the city.
- tranport facilities nearby.

We will also use data published by Office of the Registrar General Census Commissioner, India Ministry of Home Affairs, Government of India under national census. Census data contains extensive information ranging from population to households amenities and assets. This data is avaliable at Tehsil (roughy translated to locality) level. For example we can use census data to find out the average car ownership in a locality. It may be noted that car ownership is considered a sort of luxary and not everyone owns a car as is normal in developed countries. It may be assumed that potential patrons of a fine dining restaurant will be car owners. We can first select a suitable number of districts/localities in delhi based on relative affluence of the neighbourhood and limit further analysis to these districts/localities only.

2.2 Data acquisition

Data was downloaded from censusindia.gov.in. Raw data is in separate excel sheets for each of 9 districts of National Capital Territory of Delhi. The sheets contained almost 140 columns. The sheets were combined into one sheet and all unnecessary dolumns were dropped. The final excel sheet was uploaded on my google drive for excess through Jupyter notebook.

Latitude and longitude data was extracted through Nominatim of geopy.gecoder. Foursquare credentials were earlier set in previous courses of the IBM Data Science Specialization Pragramme on Coursera which were utilized in this project. The venue data was collected from foursquare.

2.3 Data Processing

As discussed in the Introduction, only names of tehsils(from now own mentioned as neighbourhoods) and percentage car ownership columns were selected for further analysis. For each neighbourhood, data was available local-

	tehsil	car	latitude	longitude
0	Rajouri Garden	37.241176	28.642152	77.116060
1	Defence Colony	36.568000	28.571791	77.232010
2	Vivek Vihar	34.555556	28.669164	77.312267
3	Karol Bagh	34.187500	28.652998	77.189023
4	Connaught Place	30.000000	28.631383	77.219792
5	Preet Vihar	27.551613	28.641441	77.295259
6	Model Town	27.073333	28.702714	77.193991
7	Pahar Ganj	25.330000	28.639852	77.213031
8	Vasant Vihar	24.756667	28.560691	77.160791
9	Patel Nagar	23.971429	28.465564	77.039294
10	Hauz Khas	23.185246	28.544256	77.206707
11	Parliament Street	22.783333	28.617188	77.207808
12	Chanakya Puri	20.400000	28.613896	77.207592
13	Saraswati Vihar	20.376389	28.477224	77.083276
14	Delhi Cantonment	20.225000	28.593833	77.134979

Figure 1: Processed dataframe

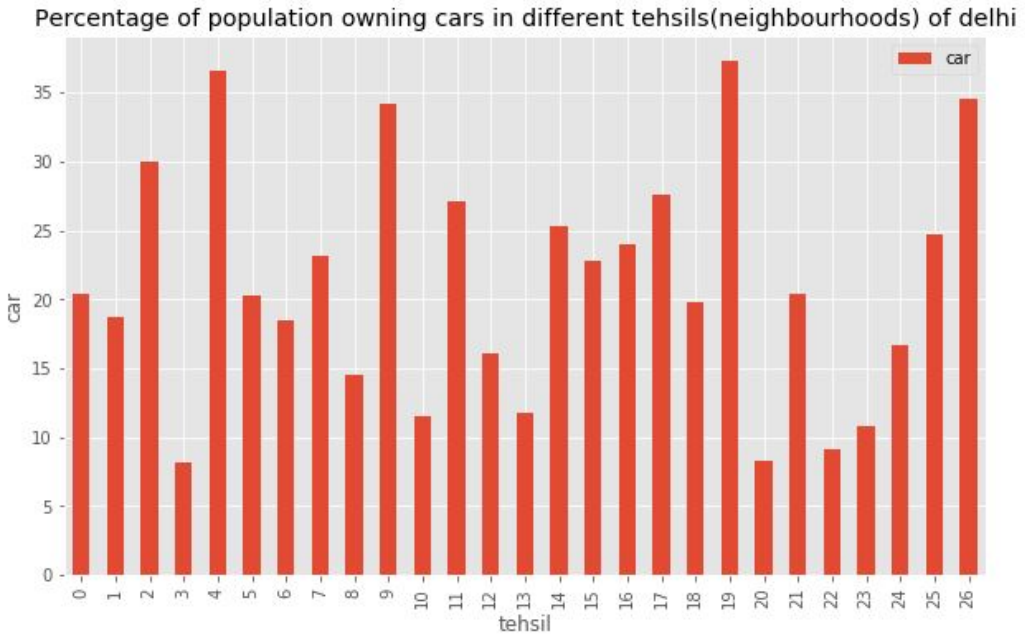

Figure 2: Bar-chart of percentage car ownership

ity wise. So data was grouped neighbourhood wise and sorted in descending order of percentage car ownership.

In the next data processing step, latitude and longitude data for each neighbourhood, which was collected through nominatim, was added to the dataframe.

3 Methodology

3.1 Data exploration

The locality wise percentage car data was plotted in a bar plot. In next step, the processed data frame was containing percentage car ownership and latitude/longitude information was used to plot a map of delhi with the neighbourhoods imposed thereon. Further, for each neighbourhood, nearby venues in the radius of 1500 were obtained through foursquare. The data was put in a separate dataframe delhi-venue alongwith the names of the neighbourhoods. In the next step, one-hot encoding was done and put into a new dataframe. This dataframe was transposed and all the values were added to get score for each neighbourhood. This score represented the number

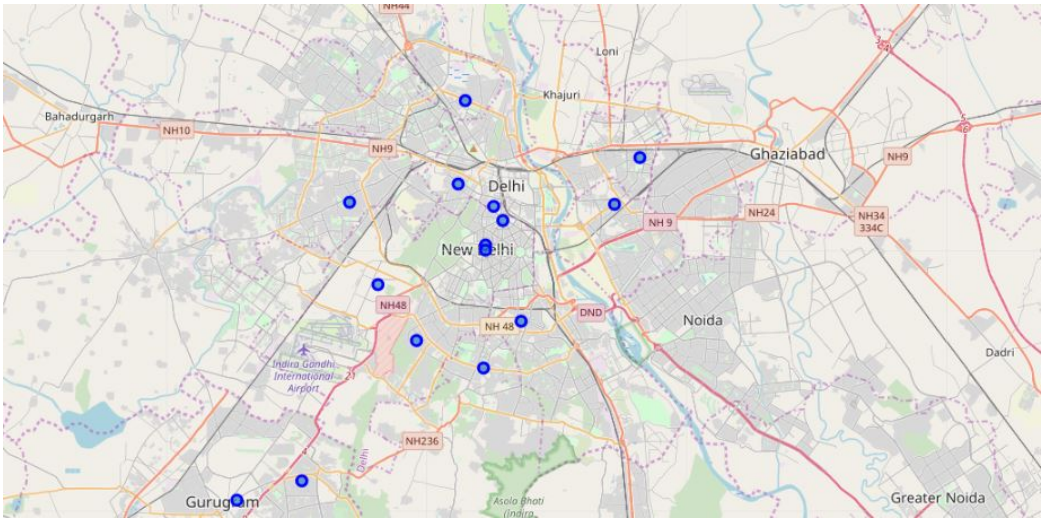

Figure 3: Neighbourhoods superimposed on map of Delhi

of venues around each neighbourhood. The venue data was grouped neighbourhood wise and means of the one-hot encoding was taken. A dataframe containing top 10 venues around each neighbourhood was also taken

This dataframe was used to perform k-means clustering and resulting neighbourhoods were marked cluster-wise. This data was put on map of delhi.

4 Results

Out of total 26, neighbourhood, top 15 were selected for further analysis as rest of the neighbourhoods were having very low density of car ownership. From personal knowledge, the discarded neighbourhoods were mostly rural ones and not suitable for a fine dining restaurant business.

As detailed in previuos section, a neighbourhood score was calculated for all the neighbourhood. This score represents number of interesting venues around each neighbouhood. Since it is one of the criteria to decide the top three neighbourhood to set up the fine dining restaurant, a list of top five neighbourhood in terms of the score are listed below:

- Connaught Place
- Pahar Ganj
- Defence Colony

- Hauz Khas
- Rajauri Garden

When the locations on the projected map were observed, it was seen that Saraswati Vihar has been placed in neighbouring town of gurgaon, although it is part of National Capital Region. So it may be an acceptable situation but due to abundant caution, I decided to drop the Saraswati Vihar from the top five neighbourhoods. It may be recalled that we have to suggest top three places to set-up the fine dining restaurants. It was felt that there is probability that all suggestions could be very similar in terms of venues near-by. To solve this problem, three clusters were prepared by k-means clustering containing all the fifteen neighbourhood. Now cluster labels are added to the top five neighbourhoods.

- Connaught Place, (Cluster: 0)
- Pahar Ganj, (Cluster: 0)
- Defence Colony, (Cluster: 1)
- Hauz Khas, (Cluster: 1)
- Rajauri Garden, (Cluster: 1)

It may be noted that top three neighbourhood contain two from first cluster and one from second cluster which is a desirable situation. It was also observed that there is only one neighbourhood in the third cluster with very low score. Therefore, we need not worry about including a member from third cluster.

Therefore, finally the three top locations selected for establishing a fine dining restaurant are given next.

5 Discussion

Based on personal knowledge of delhi, the recommendations provided by the above model are totally satisfactory. However, it was noted that there were many other data-points available in census data which were discarded for the sake of simplicity of modelling and analysis. However, those data points may be explored and analysed to give better and more refined results. It was also felt that the scoring of neighbourhoods can be differentially weighted for different categories of venue with the view to give more weight to those venues which have direct correlation with the fine-dining retaurant business. It is left as future work.

6 Conclusion

The problem was to give three location in delhi to open a new fine dining restaurants. Based on our analysis, we give our recommendation:

6.1 Three neighbourhoods in Delhi for establishing a fine dining restaurant

- Connaught Place
- Pahar Ganj
- Defence Colony