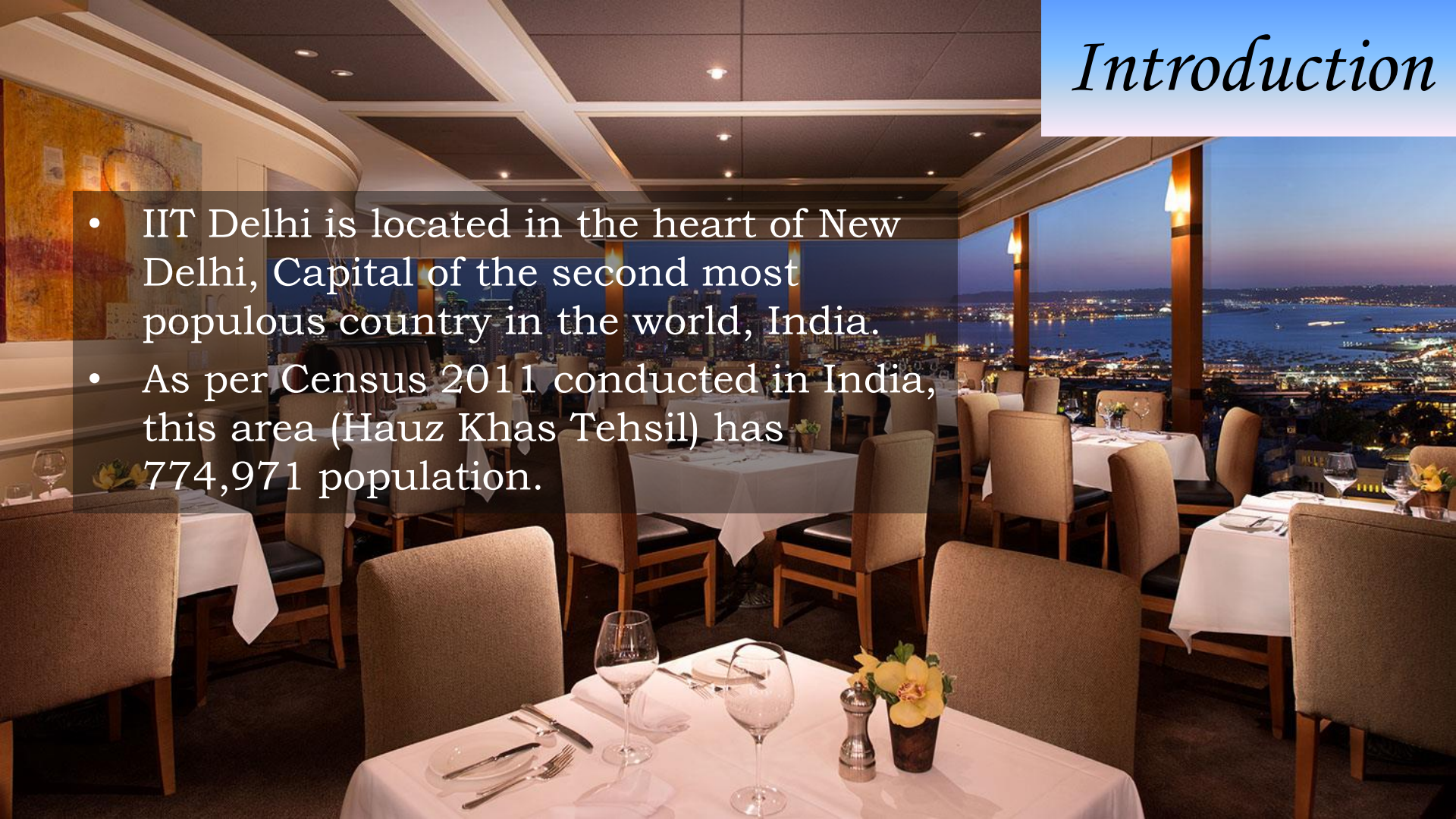
Battle of the Neighbourhoods

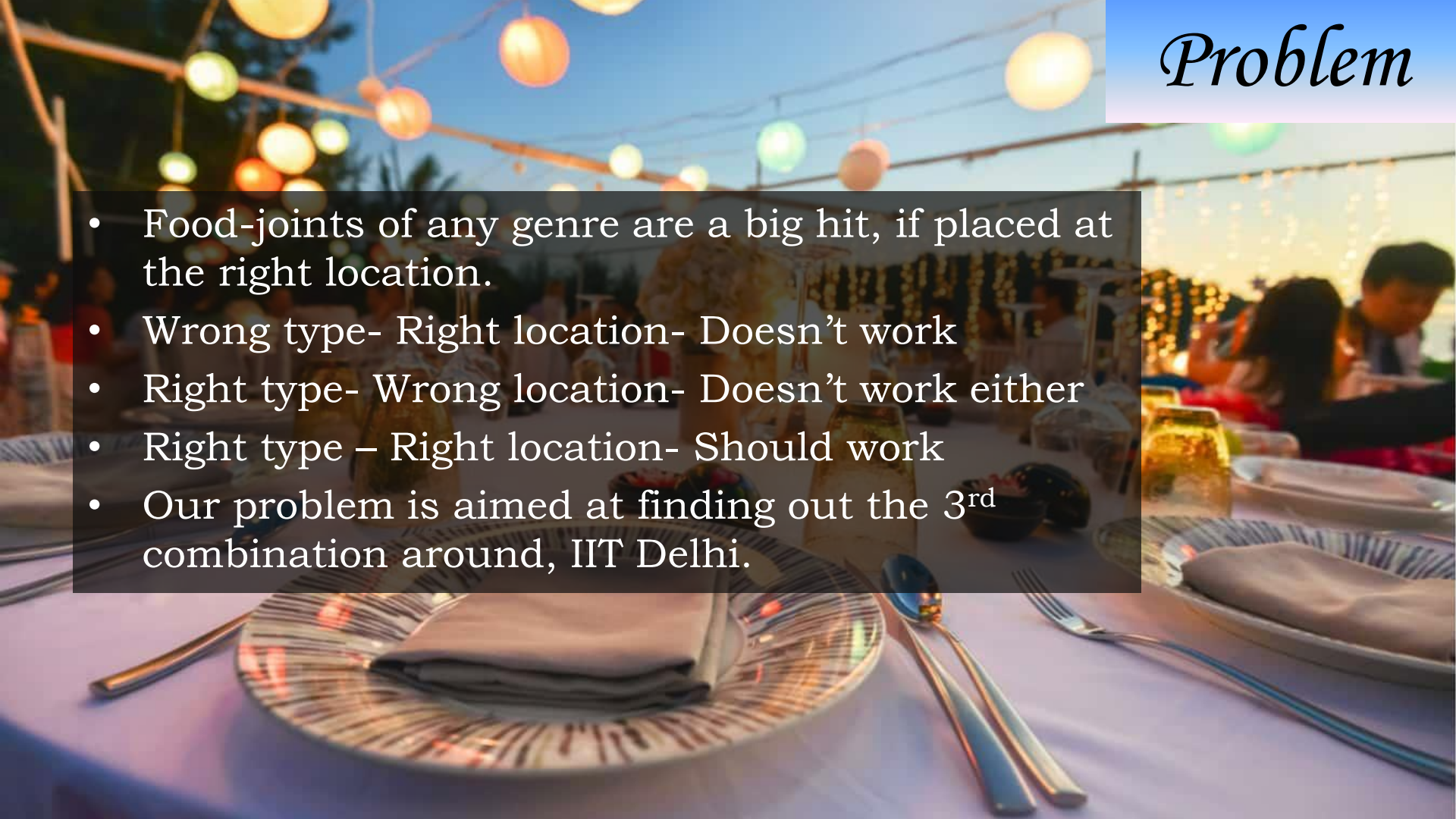
~Sushant

Introduction



- IIT Delhi is located in the heart of New Delhi, Capital of the second most populous country in the world, India.
- As per Census 2011 conducted in India, this area (Hauz Khas Tehsil) has 774,971 population.

Problem



- Food-joints of any genre are a big hit, if placed at the right location.
- Wrong type- Right location- Doesn’t work
- Right type- Wrong location- Doesn’t work either
- Right type – Right location- Should work
- Our problem is aimed at finding out the 3rd combination around, IIT Delhi.

Solution

- Foursquare data will be used to carve out a map of current locations of the restaurants to:- rightfully infer the spatial distribution of these restaurants with their genre to spot distribution density of restaurants type in a particular area.
- Eventually, a location and type of restaurants can be recommended as a prospective future site for a successful restaurant.

DATA ACQUISITION AND CLEANING

The core data source for this project is Foursquare data with “Food” as the top-level category.

Feature Selection

- Name (to identify uniqueness)
- Category
- Location (latitude, longitude)

Tools

Python packages:

- pandas (for dataframes and analysis)
- numpy (to help handle the data)
- scikitlearn (for k-means clustering)
- matplotlib (to create visuals)

Data Cleaning

API calls made to Foursquare return JSON data. It’s necessary to strip out much of what is returned to isolate only the features mentioned in Feature Selection.

The Process

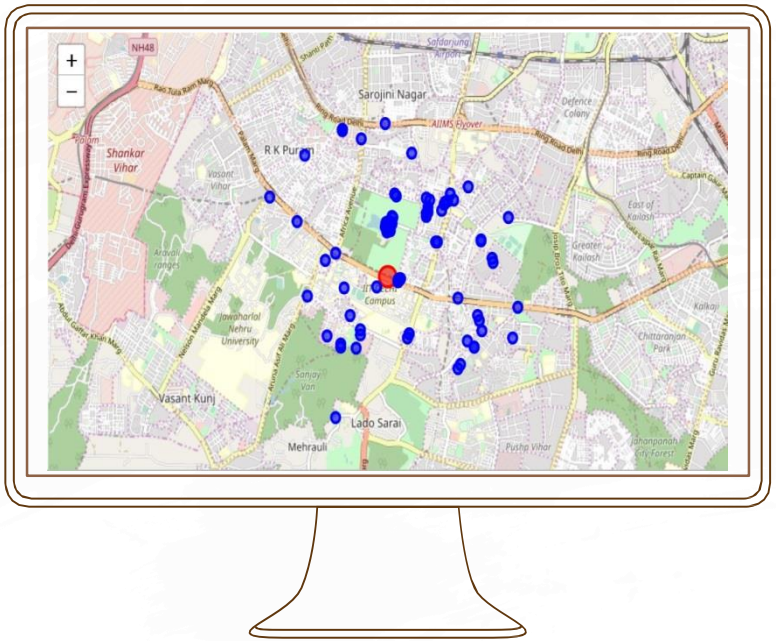
While name and location are easy to isolate in the JSON data, we need to create a function to get an easy to understand category. With this function, we can run through our dataset and clean up category to get a nice dataframe with the four features as columns.

EXPLORATORY DATA ANALYSIS

The Location

Data is extracted using Foursquare, the parameter being radius from a given location.

IIT Delhi in Hauz Khas, New Delhi was chosen as Foursquare data was readily Available for this location.



Primary Inference

With all our data into a simple dataframe, what does a map of the current restaurants around IIT Delhi would look like?

This is a useful step to get an impression if clustering will be the right approach.

k-MEANS CLUSTERING

An Unsupervised Machine Learning Approach

Why Unsupervised?

I want to limit my input and bias when it comes to finding an optimal business type and location.

This unsupervised approach will provide a way to group the current restaurants without my input. This will provide valuable insight compared to me pre-determining groups.

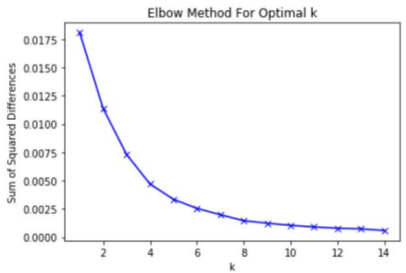
How Many Clusters

I did not want to pick an arbitrary value for *k*, so I utilized the Elbow Method to help determine the optimal *k*.

In this method, I plot the sum of squared differences (ssd) and determine the *k* where there is less value in adding *k* based on the reduction in sum of squared differences.

I determined for my data, the appropriate *k* was four. This means, I have 4 distinct groups of restaurants in the Macon area which I will use to compare.

The Elbow Method



Elbow method for Optimal 'k'

PREDICTIVE MODELING

Setting up *k*-means clustering

4 areas

$k = 4$

Using scikitlearn, we perform Kmeans clustering using 4 to start segmenting the current restaurant locations into 4 distinct clusters.

15 iterations

n_init = 15

Because we randomly drop in centroids and seek the optimal location, we can get different outcomes if we run the process multiple times.

Running this 15 times helps us truly come to an optimal location for a centroid within each cluster.

PREDICTIVE MODELING

Results of *k*-means clustering

Each Venue Labeled

- Performing k-means clustering on our dataset will assign each restaurant a label between 0 and 3.
- The label represents their cluster.
- This information is built back into the original dataframe so we have an easy to read table of the location name and the cluster it belongs to.

The Current Dataframe (head)

[30]:

	name	categories	lat	lng	Cluster
0	Yeti - The Himalayan Kitchen	Tibetan Restaurant	28.553656	77.194261	1
1	Imperfecto	Mediterranean Restaurant	28.554657	77.195092	1
2	Coast Cafe	Café	28.554779	77.195214	1
3	Naivedyam	South Indian Restaurant	28.554987	77.195104	1
4	Smoke House Deli	Deli / Bodega	28.554424	77.193846	1

PREDICTIVE MODELING

Results of *k*-means clustering *(cont.)*

Centroid Locations

- Each cluster has an optimal center, as mentioned previously.
- The center is the mean of all the locations within the cluster.
- I created an additional dataframe with the centroids so they can be easily added to the map later as a visual reference.

The Centroid Dataframe

[28]:

	Cluster	lat	lng
0	0	28.540192	77.187898
1	1	28.554448	77.197763
2	2	28.565913	77.183400
3	3	28.542641	77.213371

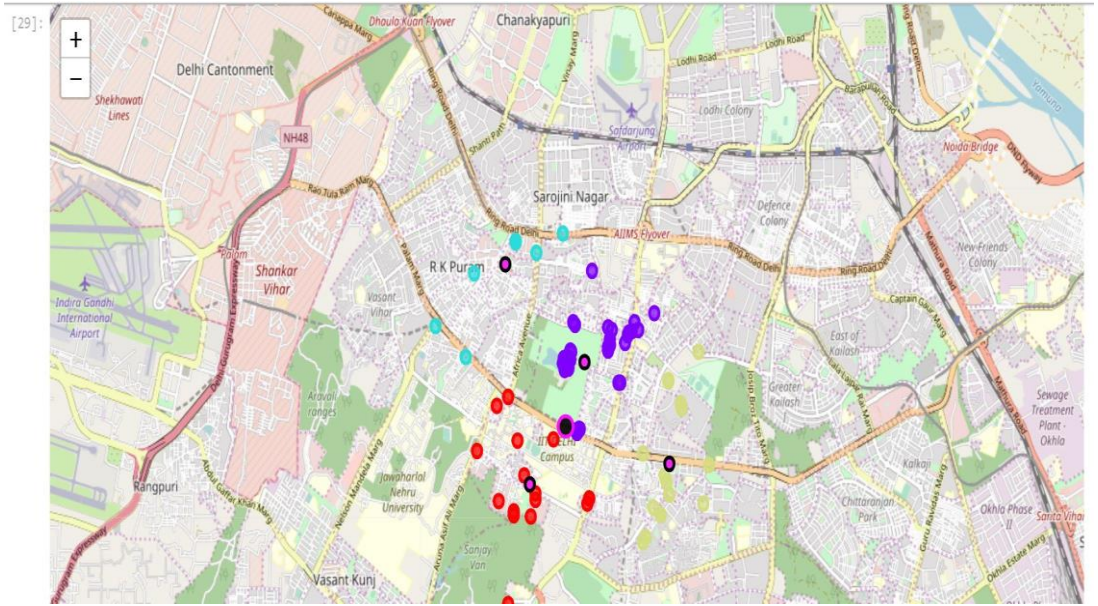
VISUALIZE THE CLUSTERS

Results of *k*-means clustering *(cont.)*

Cluster Information

- **IIT Delhi:** pink circle, black fill
- **Cluster Centroids:** black circle, pink fill

- Clusters are color coded dynamically, location relative to IIT Delhi
  - **Cluster 0** – Red (north-northwest)
  - **Cluster 1** – Purple (west)
  - **Cluster 2** – Teal (south)
  - **Cluster 3** – Olive (northeast)



EXAMINING THE CLUSTERS

FREQUENCY

Rank the categories

- Need to determine where a category is under-represented.
- Ranking the categories by frequency within each cluster offers a way to see this.
- My approach looks at the top 5 most common types of restaurants within a given cluster.

ONE-HOT ENCODE

Categories to Dummies

- Need dummy variables instead of categorical variables to determine frequency.
- One-hot encoding of the variables effectively translates the categorical variables into dummy variables.
- From here we can look at the mean frequency of a category across the cluster.

GOAL

Under-Represented

- In this situation, under-represented will be determined by the following criteria:

1. Category is represented in the top 5 for at least 2 other clusters.
2. Category is not in the top 5 for a selected cluster.

- If those criteria are met, I will select the most represented category based on criteria 1 and the cluster which applies to criteria 2.

THE RESULTS

Café

Café is the 2nd,2nd & 1st most common venue in cluster 0, cluster 1 & cluster 2 respectively

Cluster 3

1.Café is the 2nd,2nd & 1st most common venue in cluster 0, cluster 1 & cluster 3 respectively.
2. & It doesn't even make the top 5 in Cluster 3.

Conclusion

Cluster 3 appears to be a very good option for a fast food restaurant, especially given the proximity to a university and the lack of competition for that type of restaurant.

POSSIBLE FURTHER REFINEMENTS

Where to go from here?

Repeatable Process

You can easily replicate this process for any place with Foursquare data.

Improvement

This approach only capitalizes on location.

Factors such as migration, demographic profile, crime rate etc. may be incorporated

Improvement

The Elbow Process described doesn't always provide a clear answer to the optimal *k*. Other methods could be explored to determine an optimal value.

Disclaimer

This is a work for the Applied Data Science Capstone towards IBM Data Science Professional Certificate

The Fact

This project is a part of IBM Data Science Professional Certificate.