

Coursera Data Science Capstone Project

Open A New Shopping Mall in San Jose, California

Business Problem and Target Audience

- San Jose is booming Economically due to High Technology Expansion
- Where we commend the Developers to open the new Shopping Mall?
- We can apply Data Science Methodology and Machine Learning to solve this problem
- Target Audience of this Project is the Business Developer

Source of Data

- Use of Data of Neighborhood of San Jose California from Wikipedia
- Latitude and longitude of those neighborhoods which is required to plot the map and use to get the venue data
- Venue Data particularly related to shopping malls and it is used to perform clustering of the neighborhood.

Methodology

- Wiki Page:
[https://en.wikipedia.org/wiki/Category:Neighborhoods in San Jose, California](https://en.wikipedia.org/wiki/Category:Neighborhoods_in_San_Jose,_California)
- Web scraping with Python requests and BeautifulSoup to extract list of the neighborhoods data.
- From Neighborhood names get to the geographical coordinates: latitude and longitude use Geocoder package
- After convert address into geographical coordinates populate them in DataFrame
- Visualize the neighborhoods in a map using folium package.
- Sanity check to make sure that the geographical coordinates by Geocoder in plot the city of San Jose.
- Use Foursquare API to obtain Venue Data at each Geographical Location associated with Neighborhoods names with lots of venues and activities in them

Machine learning K-means Clustering

- Analyze each neighborhood and taking the mean of the frequency of occurrence of each venue category.
- Prepare the data for use in clustering.
- Analyzing the “Shopping Mall” data as locus of this project
- “Shopping Mall” is venue category for the neighborhoods.

K-means Clustering leads to Answers

- Cluster the neighborhoods into 3 clusters based on their frequency of occurrence for the “Shopping Malls”.
- Identify neighborhoods higher and fewer numbers of Shopping Malls
- Based on occurrence of shopping malls in different neighborhoods
- Answer question as to which neighborhoods are most suitable to open new shopping malls.

Results of the Study

- Cluster the neighborhoods into 3 clusters based on their frequency of occurrence for the “Shopping Malls
- The results from k-means clustering show that we can categorize the neighborhoods into 3 clusters :
- Cluster 0: 30 Neighborhoods with no existence of shopping malls
- Cluster 1: 7 Neighborhoods with moderate concentration of shopping malls
- Cluster 2: 10 Neighborhoods with high concentrated number of malls

Discussion

- The high concentration Cluster 1 not be recommended to open new shopping malls
- The big opportunity would be in Cluster 0 no competition for foot traffics since there is no shopping malls
- There are other factors such as population density and income levels of interests in future research

Conclusion: Data Science Process

- In this project we have identified the process
 - identifying business problem,
 - specifying the data required
 - extracting and preparing the data
 - performing machine learning by clustering the data into 3 clusters based on their similarities
 - proving and providing recommendations to the relevant stakeholders regarding the best locations to open a new shopping mall.

Conclusion: the Recommendation

- The neighborhoods in cluster 0 are the most preferable locations to open a new shopping mall.
- It capitalizes on opportunities on high potential locations
- Avoids overcrowded areas in their decisions in opening a new shopping mall.

