Hospital Neighborhood selection for Traveling Nurse Use Case

Katherine Terlesky, Ph.D.

Introduction – Business Problem

Traveling nurses are healthcare professionals who choose to work in cities with short term nurse staffing needs. The benefits of being a traveling nurse can be higher pay and the ability to travel around the country exploring cities. Challenges with this traveling profession are identification of cities that meet your interests and professional needs.

This tool will help prospective travelling nurses identify hospital neighborhoods with appealing venues within a radius of the hospital. This data science project will explore integration of hospital rating data available from the Center for Medicare and Medicaid services with Foursquare venue data.

In the United States, the Medicare program is a government sponsored health insurance program for adults age 65 and older and some people with disabilities. Medicaid is a government-sponsored health program for low-income people. The U.S. Gov has developed metrics and requirements hospitals must follow to continue to receive Federal aid.

The analysis will be performed just for hospitals in the five boroughs of New York City. In addition, the Foursquare data will be optimized to provide venues focusing on one of the parameters of a Foursquare query (section). Parameters from Foursquare are food, drinks, coffee, shops, arts, outdoors, sights, trending, next Venues, or top Picks. Choosing one of these limits results to venues with the specified category or property.

The results will provide neighborhood clusters centered around the hospitals in New York and show the top venues for each neighborhood. This data will be displayed on a map and help the prospective nurses better select their desirable hospital neighborhoods.

Data

Data Sources

1. The Foursquare API will be used to retrieve venue information around the hospitals in New York City. https://foursquare.com/
2. The Hospital Ratings data will be used to evaluate hospitals in the city. The complete data set of all Hospital Measures is available at https://data.medicare.gov/data/hospital-compare. A complete Data Dictionary is available as the same location describing the data fields and methodology of hospital rating assessment.
    - Hospitals in the United States are rated based on 7 areas of Quality including Mortality, Safety of Care, Readmission, Patient Experience, Effectiveness of Care, Timeliness of Care, and Efficient Use of Medical Imaging.
    - The ratings in these seven areas are combined into a star rating for each hospital (range of 1-5 starts).

- There are up to 51 measures across the 7 categories, but hospital ratings are only calculated using those measures for which data are available. The average hospital rating is based on about 37 measures. A hospital summary score is then calculated by taking the weighted average of these group scores.
- Hospitals in the United States (excluding Veterans Affairs and Dept of Defense Hospitals) report data to the Centers for Medicare & Medicaid Services. This includes about 4,000 Medicare-certified hospitals across the country.

The following table shows the data fields and description and type of data in each field

Data Field	**Description and Type**
Facility ID	6 characters
Facility Name	72 characters
Address	51 characters -Street address of hospital
City	20 characters City of hospital
State	2 characters
ZIP Code	8 digit number
County Name	25 character county name
Phone Number	14 character phone number
Hospital Type	34 character type of hospital
Hospital Ownership	43 character type of hospital ownership
Emergency Services	3 character (Yes/No) for Emergency Services
Meets criteria for promoting interoperability of EHRs	1 character (Y/N) for meets interoperability
Hospital overall rating	13 character 1-5 or not available
Hospital overall rating footnote	8 digit number
Mortality national comparison	28 character rating Below, At or Above National average or Not Available
Mortality national comparison footnote	8 digit number
Safety of care national comparison	28 character rating Below, At or Above National average or Not Available
Safety of care national comparison footnote	8 digit number
Readmission national comparison	28 character rating Below, At or Above National average or Not Available
Readmission national comparison footnote	8 digit number
Patient experience national comparison	28 character rating Below, At or Above National average or Not Available
Patient experience national comparison footnote	8 digit number
Effectiveness of care national comparison	28 character rating Below, At or Above National average or Not Available
Effectiveness of care national comparison footnote	8 digit number
Timeliness of care national comparison	28 character rating Below, At or Above National average or Not Available
Timeliness of care national comparison footnote	8 digit number

Efficient use of medical imaging national comparison	28 character rating Below, At or Above National average or Not Available
Efficient use of medical imaging national comparison footnote	8 digit number

Methodology

Use Case Development – A fictional nurse use case was developed to focus the business case.

Development environment - A Jupyter notebook development environment in the IBM Cloud was utilized to write and execute the necessary code. A Github repository was utilized for version control. The following libraries were utilized in the execution of this project: NumPy, pandas, json, matplotlib, geopy geocoders, folium, seaborn, and sklearn.

Data Preprocessing

The following data cleaning actions were taken on the Hospital data set.

1. Removed Columns that reference the footnotes and other data not important to this analysis
2. Dropped the ratings other than overall or safety ratings
3. Removed psychiatric hospitals from the data set
4. Converted Safety ratings to numbers
    a. 0= Not Available
    b. 1=Below national average
    c. 2=At national average
    d. 3=Above national average

Exploratory data analysis

The entire hospital data set (5319 hospitals) was analyzed for the distribution of Overall Rating and Safety Rating in the data set.

Seaborn was used to visualize the overlap of safety rating and hospital ratings.

Defining the hospital neighborhood

The hospital data was narrowed down to hospitals just in state of New York by setting the index to ‘State” and filtering for ‘NY’.

The index was reset and then set to ‘County’ and filtered on just the five boroughs of New York City. This includes Bronx, New York, Queens, Kings, and Richmond.

The index was then reset and set ‘Safety’ rating and all hospitals with a Safety rating of Not Available or At or Above the national average (0,2,3) were removed.

Adding geolocation information

Multiple attempts using geopy Nominatum resulted in the necessity for additional clean up. Numeric street names spelled out like "SIXTH" were converted to numeric representation (6^{th})

To obtain the full address, the columns of address (street address), city, state and Zip code were concatenated into one column 'Full Address'.

Geopy Nominatum was used to look up the latitude and longitude of each hospital address in the narrowed down DataFrame.

Folium was used to create a map of NYC with the hospitals depicted on the map.

Collection of venue data

Information about venues around the hospitals were collected using data available in Foursquare from the Foursquare API. https://foursquare.com/

The Foursquare data tended to return mostly restaurant data which was not of interest to our Use Case. To increase the impact of venues in Arts and & Entertainment and Outdoors, the 'section' constraint was used in the Foursquare call

Venues were first retrieved using the 'Outdoors' section for the Foursquare call. The frequency of 'Outdoors' venues by neighborhood was evaluated. Venues were then retrieved using the 'Arts' section for the Foursquare call. The frequency of venues by neighborhood was performed. The two data sets were combined, and the top 10 venues of the combined set were converted to a Pandas DataFrame.

Machine Learning - K-means clustering

Unsupervised machine learning was selected as an approach to cluster like neighborhoods. Unsupervised machine learning was selected because there is no training set available. The objective of clustering is to group like objects together.

K-means clustering was chosen as the method of unsupervised machine learning. The clustering was performed using clusters of 3-7 and the final analysis used 4 clusters.

Other types of machine learning algorithms that were considered were hierarchical clustering. This was not selected because the data used to generate clusters in the set are very close and not easily distinguishable from each other.

The K-means clustering methodology was imported from sci kit learn.

Geo-visualization of clusters

Folium was used to create a map of NYC with the hospitals depicted as colored clusters on the map.

Selection of ideal cluster

The characteristics of each cluster were summarized.

The ideal cluster for the use case was determined by present the nurse in Use Case 08302020 with the information from the analysis. This guided them to make an informed decision about the best hospital neighborhood for their traveling nurse experience.

Results

Travelling Nurse Use Case 8302020 - A Use Case is defined for each traveling nurse profile. This then allows the search to be tailored.

1. The nurse in Use Case 08302020 has experience in developing processes to improve the Safety of Care measure. They are interested in working in hospitals with the Safety of Care measure below the national average.
2. Preferred hospital type = Acute Care hospital.
3. Preferred venue type in neighborhood = Arts and Outdoors

Preferred hospital rating focus area = Prefers hospitals with low Safety of Care The entire hospital data set (5319 hospitals) was analyzed for the distribution of Overall Rating in the data set.

There were 5319 hospitals in the dataset in all 50 states plus U.S. territories. This was narrowed down to 3262 hospitals after preprocessing and data clean up.

Summary statistics

The entire hospital data set (5319 hospitals) was analyzed for the distribution the Overall Rating in the data set.

Not Available	1749
Overall Rating = 1	223
Overall Rating = 2	702
Overall Rating = 3	1103
Overall Rating = 4	1135
Overall Rating = 5	407
Total	5319

The entire hospital data set (5319 hospitals) was analyzed for the distribution the Safety Rating in the data set.

Not Available	2711
Below National average	847
Same as national average	543
Above national average	1218
Total	5319

A seaborn plot was used to visualize how the safety ratings are distributed in the data set of the overall ratings. The following plot shows this representation. Each plot represents an Overall Rating (0-5). Where 0 is Not Available. The Safety rating is shown as a colored bar.

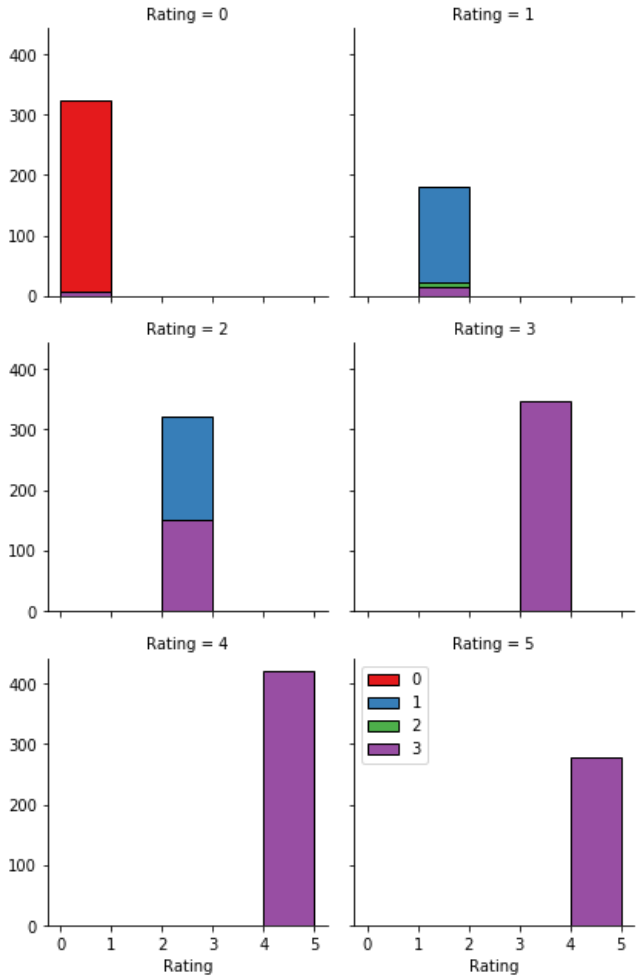

Figure 1 Distribution of Safety Ratings as a function of Overall Rating.

To optimize the experience for the Nurse in Use Case 08302019, only the hospitals with a safety rating of 1 were selected. These are hospitals below the national average in safety.

The initial process of narrowing down the number of hospitals in the State of New York resulted in 143 hospitals in the New York DataFrame. The second step of narrowing to the 5 boroughs of New York City resulted in 37 hospitals. And selection of hospitals with a Safety Rating of 1 resulted in a final hospital set of 26 hospitals.

The following 26 hospitals were used to develop the "Hospital Neighborhoods" to compare the cultural experience of the neighborhood around the Hospital with Foursquare data.

	Facility	latitude	longitude
0	BRONX-LEBANON HOSPITAL CENTER - CONCOURSE DIVI...	40.83	-73.90
1	MONTEFIORE MEDICAL CENTER	40.88	-73.88
2	LINCOLN MEDICAL & MENTAL HEALTH CENTER	40.82	-73.92
3	JACOBI MEDICAL CENTER	40.85	-73.85
4	ST BARNABAS HOSPITAL	40.82	-73.92

5	NEW YORK-PRESBYTERIAN HOSPITAL	40.76	-73.95
6	LENOX HILL HOSPITAL	40.77	-73.96
7	MOUNT SINAI BETH ISRAEL	40.73	-73.98
8	BELLEVUE HOSPITAL CENTER	40.74	-73.98
9	NEW YORK UNIVERSITY LANGONE MEDICAL CENTER	40.74	-73.97
10	JAMAICA HOSPITAL MEDICAL CENTER	40.69	-73.81
11	ELMHURST HOSPITAL CENTER	40.74	-73.88
12	FLUSHING HOSPITAL MEDICAL CENTER	40.76	-73.82
13	ST JOHN'S EPISCOPAL HOSPITAL AT SOUTH SHORE	40.60	-73.75
14	BROOKLYN HOSPITAL CENTER - DOWNTOWN CAMPUS	40.69	-73.98
15	MAIMONIDES MEDICAL CENTER	40.64	-74.00
16	NYC HEALTH + HOSPITALS/CONEY ISLAND	40.59	-73.96
17	KINGSBROOK JEWISH MEDICAL CENTER	40.66	-73.93
18	KINGS COUNTY HOSPITAL CENTER	40.66	-73.94
19	WYCKOFF HEIGHTS MEDICAL CENTER	40.70	-73.92
20	BROOKDALE HOSPITAL MEDICAL CENTER	40.66	-73.91
21	NEW YORK-PRESBYTERIAN/BROOKLYN METHODIST HOSPITAL	40.67	-73.98
22	SUNY/DOWNSTATE UNIVERSITY HOSPITAL OF BROOKLYN	40.65	-73.94
23	INTERFAITH MEDICAL CENTER	40.68	-73.94
24	RICHMOND UNIVERSITY MEDICAL CENTER	40.64	-74.11
25	STATEN ISLAND UNIVERSITY HOSPITAL	40.59	-74.09

The following map shows the distribution of the hospitals across the five boroughs of New York City.

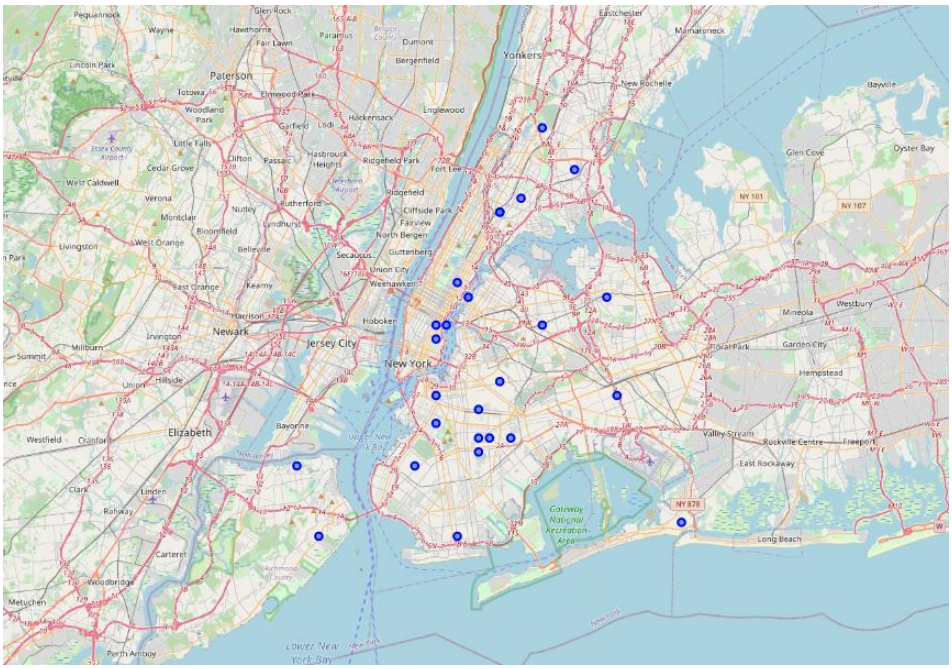

Figure 2 Distribution of the hospitals in the data set on the map of New York City

The Arts and Outdoors venues were obtained from Foursquare and clustered using K-means. The clusters were visualized on a map of New York City using Folium. The following descriptions of each cluster summarize the characteristics of each cluster.

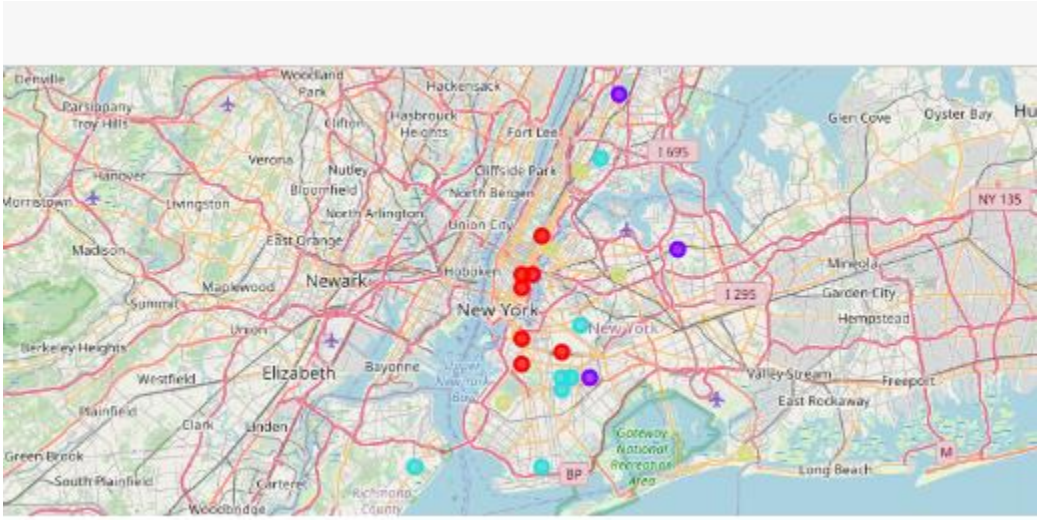

Figure 3 Visualization of the Clusters of hospitals on the map of New York City

Cluster Characteristics

Results-Cluster 0 (Red) is the Downtown living Cluster with high density of Arts & Entertainment. The following table shows the hospitals and most common 5 venues in Cluster 0.

Cluster 0 = Downtown living Cluster with high density of Arts & Entertainment						
Number	Hospital Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
6	LENOX HILL HOSPITAL	Art Gallery	Gym / Fitness Center	Theater	Gym	Art Museum
7	MOUNT SINAI BETH ISRAEL	Art Gallery	Theater	Playground	Indie Theater	Gym / Fitness Center
8	BELLEVUE HOSPITAL CENTER	Theater	Art Gallery	Gym / Fitness Center	Gym	Comedy Club
9	NEW YORK UNIVERSITY LANGONE MEDICAL CENTER	Arts & Entertainment	Art Gallery	Outdoor Sculpture	Scenic Lookout	Garden
14	BROOKLYN HOSPITAL CENTER - DOWNTOWN CAMPUS	Gym / Fitness Center	Theater	Performing Arts Venue	Dance Studio	Opera House
21	NEW YORK-PRESBYTERIAN/BROOKLYN METHODIST HOSPITAL	Art Gallery	Gym	Theater	Yoga Studio	Gym / Fitness Center
23	INTERFAITH MEDICAL CENTER	Art Gallery	Museum	Theater	Playground	Park

Results -Cluster 1 (purple) is High Park density over Arts & Entertainment

Cluster 1 = High Park density over Arts & Entertainment						
Number	Hospital Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	MONTEFIORE MEDICAL CENTER	Park	Dance Studio	History Museum	Performing Arts Venue	Museum
12	FLUSHING HOSPITAL MEDICAL CENTER	History Museum	Plaza	Park	Public Art	Yoga Studio
20	BROOKDALE HOSPITAL MEDICAL CENTER	Art Gallery	Performing Arts Venue	Park	Pool	Music Venue

Results - Cluster 2 (light blue) is the Music Cluster with balanced access to Outdoors venues

Cluster 2 = Music Cluster with balanced access to Outdoors venues						
Number	Hospital Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	BRONX-LEBANON HOSPITAL CENTER - CONCOURSE DIVI...	Park	Movie Theater	Music Venue	Playground	Gym
16	NYC HEALTH + HOSPITALS/CONEY ISLAND	Concert Hall	Performing Arts Venue	Music Venue	Playground	Gym / Fitness Center
17	KINGSBROOK JEWISH MEDICAL CENTER	Music Venue	Yoga Studio	Gym / Fitness Center	Dance Studio	Beach
18	KINGS COUNTY HOSPITAL CENTER	Music Venue	Dance Studio	Gym	Playground	Park
19	WYCKOFF HEIGHTS MEDICAL CENTER	Gym	Music Venue	Art Gallery	Park	Circus
22	SUNY/DOWNSTATE UNIVERSITY HOSPITAL OF BROOKLYN	Music Venue	Museum	Gym	Dog Run	Lake
25	STATEN ISLAND UNIVERSITY HOSPITAL	Dance Studio	Yoga Studio	Park	Gym	Scenic Lookout

Results - Cluster 3 (light yellow) is the Art Gallery Cluster with balanced access to outdoors recreation

Cluster 3 = Art Gallery Cluster with balanced access to outdoors recreation						
Number	Hospital Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
2	LINCOLN MEDICAL & MENTAL HEALTH CENTER	Art Gallery	Movie Theater	Gym	Recreation Center	Baseball Field
4	ST BARNABAS HOSPITAL	Art Gallery	Movie Theater	Gym	Recreation Center	Baseball Field
5	NEW YORK-PRESBYTERIAN HOSPITAL	Art Gallery	Indie Theater	Gym	Baseball Field	Park
11	ELMHURST HOSPITAL CENTER	Art Gallery	Music Venue	Martial Arts School	Park	Playground
13	ST JOHN'S EPISCOPAL HOSPITAL AT SOUTH SHORE	Art Gallery	Martial Arts School	Park	Cycle Studio	Beach
15	MAIMONIDES MEDICAL CENTER	Art Gallery	Dance Studio	Martial Arts School	Park	Skate Park

Discussion

Use Case 08302020 Resolution

The traveling nurse for this use case selected to investigate open travelling nurse positions for hospitals in Cluster 3. The high density of Art Galleries with good access to recreational amenities was most appealing. Cluster 3 had hospitals in the boroughs of New York, Brooklyn, Queens, and Long Island.

This analysis and visualization allowed this nurse to narrow down their decision.

The second choice cluster was Cluster 0.

Utilization of the Medicare and Medicaid data set on hospital ratings was informative in the evaluation of hospitals for a traveling nurse use case. The data set provided a challenge to integrate it into this scenario. The street addresses used in the original data set presented some challenges and a few had to be verified independently because the geopy Nominatum could not retrieve a latitude and longitude for them.

The Foursquare data is very heavy in restaurant information. For this use case of a person desiring Arts and Entertainment and Outdoors activities, there may have been a better dataset to use. Such as a list of Art Galleries and list of parks across NYC. That was not investigated for this analysis.

Conclusion

This project demonstrated a way to combine data for a very specific use (Hospital Ratings) with data with a general use (venue ratings) to deliver a framework for a decision process at the intersection of the two data sets. The most challenging aspect was getting the geolocation to work for the hospital data set. The second most challenging was figuring out how to get specific value out of the Foursquare data that wasn't restaurants.

Additional data sources that may be valuable in further analysis would be an evaluation of short term rental options (location and pricing) in each hospital neighborhood.

About the Author

The conclusion of the Capstone report for the Coursera IBM Data Science series was a very rewarding and challenging adventure. I am an executive in Life Science company with a Ph.D. in Microbiology. I have no background in software development or anything IT. I recognized that data science is a necessary skill set in the business and science and this course has opened my eyes to new ways to think about and tackle problems.

I plan on continuing my learning by tackling problems specific to our business and working with a collaborator who specializes in Data Science. I believe I took "long cuts" to get some functions to work in Pandas and have a lot of room for optimization and learning.