Applied Data Science Capstone

April 8, 2020

Introduction

This section provides a description of the problem and a discussion of the project background.

Background

There are many situations where knowing the distance from a specific location is important. For example:

- Distance from a chemical spill
- Distance from an earthquake epicenter

In addition, knowing the distance to the closest occurrence of a specific capability is also an important factor in deciding which instance of that capability to utilize, specifically when there is a choice to be made. For example:

- Distance from the nearest fire station
- Distance to the nearest hospital, when there are multiple facilities to choose from within a defined radius.

Having access to accurate distance information (or distances, if there are multiple locations) relative to a known object or location can aide in decision making. For example, in the event of a medical emergency while in an unfamiliar city, it would be useful to know how many hospitals are nearby and which one is the closest.

Problem

This project will create a process for calculating distance from a specific location to a set of facilities. The process will be illustrated using an example based in Seattle, Washington, which is a densely populated urban area in the northwestern United States. By the way, Seattle is also the home of Amazon and nearby Microsoft (headquartered in Bellevue).

LYNNWOOD
SHORELINE
WOODINVILLE
KIRKLAND
REDMOND
SEATTLE
BURIEN
RENTON

Seattle, Washington

The specific location chosen for this illustration is Seattle's Space Needle. It is a city landmark and is considered an icon of Seattle. It was built in 1962 for the World's Fair and stands 605 feet tall.

Space Needle

This process will calculate the straight-line distance from the Space Needle to surrounding hospitals, where the primary objective is to locate the closest hospital.

Data

This section provides a description of the data and how it will be used to solve the problem.

Data Sources

The Foursquare API will be used to obtain location data for acute care hospitals, using search_query = 'Hospital', that are within 20 miles (32286.9 meters) from the Space Needle. One of the anticipated challenges with this search is that Foursquare will likely return a variety of facilities where the facility name includes 'Hospital', but the facility does not treat humans, e.g. a veterinarian hospital facility. See the Appendix for an example of the Foursquare data that will be used for this project.

The Nominatim package from GeoPy will be used to obtain the latitude and longitude for the Space Needle.

In addition, hospital bed information will be obtained from two sources:
https://en.wikipedia.org/wiki/List_of_hospitals_in_Washington_(state)

https://www.wsha.org/our-members/member-listing/

See the Appendix for an example of the hospital bed data that will be used for this project.

Since the number of beds varies across hospitals, an attempt will be made to see the extent to which the hospitals returned by the API search will cluster based on number of beds. For some people, deciding which nearby hospital to use may depend on the size of the hospital, where larger hospitals tend to provide more advanced services.

Data Preparation

The Foursquare venue search will focus on three of the four response fields (name, location, and categories):

Field	**Relevant Content**
id	n/a
name	Hospital name
location	Address, latitude, longitude
categories	Type of facility

The results returned from the Foursquare search will be transformed into Pandas data frame and filtered for relevant columns. It is anticipated that some data cleanup will be required to address issues such as missing values, duplicates, etc. Furthermore, the 'categories' response field will be parsed to extract the value for 'name', which indicates the type of facility, e.g., 'Hospital', 'Veterinarian', 'Doctor's Office', etc. This should not be confused with the 'name' response field in the above table.

The hospital bed data will be merged with the Foursquare data frame, after which data analysis can proceed.

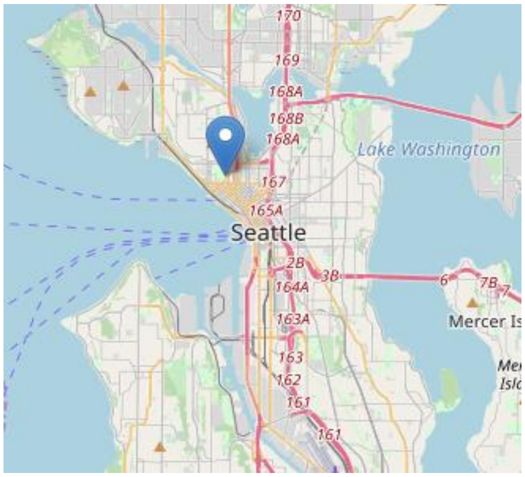
These data will be used to:

- Calculate the straight-line distance from the Space Needle to each hospital using the geodesic function.
- Perform a cluster analysis using the k-means unsupervised machine learning algorithm to see if the selected hospitals cluster based on number of beds.

Methodology

Data Extraction

The Nominatim package from GeoPy was used to obtain the latitude and longitude for the Space Needle as well as the geographic boundaries of Seattle.



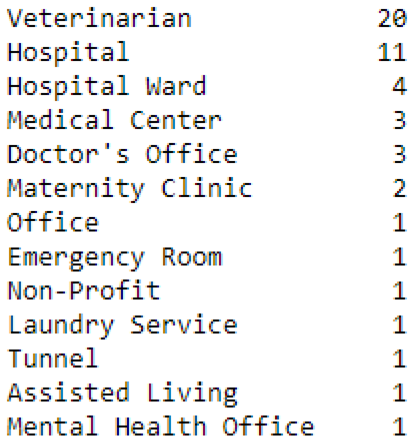
Space Needle location in Seattle, WA

A Foursquare search was conducted using the following parameters:

- Search_query = 'Hospital'
- Radius = 32286.9 (20 miles)

The search results obtained from Foursquare were transformed into a data frame consisting of 50 rows.

A review of the search results indicated that most of the facilities were not in scope for this analysis, i.e., they were not acute care hospitals. Rather, they were Veterinarian hospitals, Doctor's offices, Clinics, etc. Also, while it may appear that the three Medical Centers should be in scope, based on additional investigation it was determined that all three were either psychiatric hospitals or outpatient facilities.



```
Veterinarian            20
Hospital                11
Hospital Ward            4
Medical Center           3
Doctor's Office          3
Maternity Clinic         2
Office                   1
Emergency Room           1
Non-Profit               1
Laundry Service          1
Tunnel                   1
Assisted Living          1
Mental Health Office     1
```

Frequency Distribution

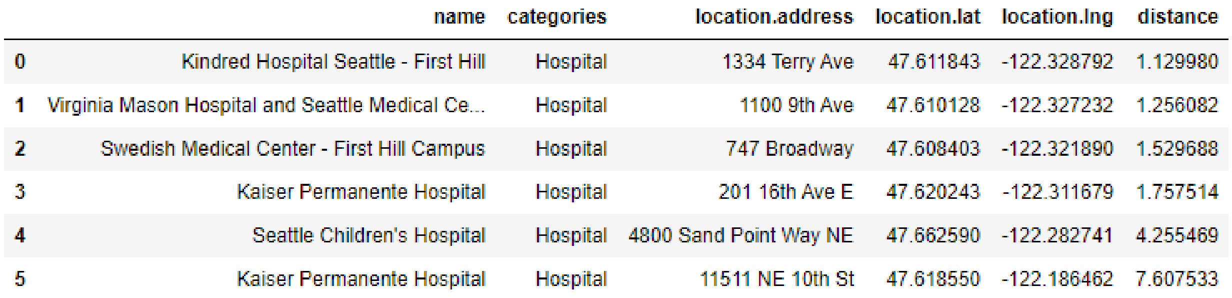
Data Clean-up

Further review revealed the following data issues:

- NaN values for address
- Duplicate addresses
- Hospitals that had closed

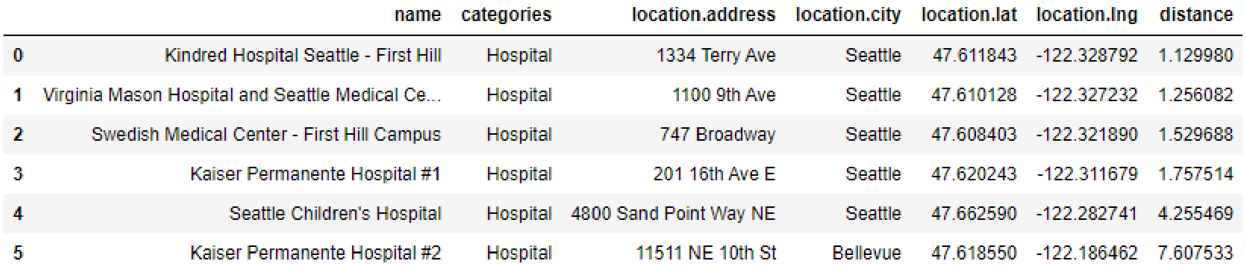
Based on this review, rows were dropped for all the above conditions, resulting in a new data frame consisting of six rows. The straight-line distance from the Space Needle to each hospital was calculated using the geodesic function.

However, additional review revealed that two hospitals at different locations had the same name. It was determined that these two hospitals were part of a large managed care organization and were in fact unique. Therefore, the names were modified to reflect this fact.



	name	categories	location.address	location.lat	location.lng	distance
0	Kindred Hospital Seattle - First Hill	Hospital	1334 Terry Ave	47.611843	-122.328792	1.129980
1	Virginia Mason Hospital and Seattle Medical Ce...	Hospital	1100 9th Ave	47.610128	-122.327232	1.256082
2	Swedish Medical Center - First Hill Campus	Hospital	747 Broadway	47.608403	-122.321890	1.529688
3	Kaiser Permanente Hospital	Hospital	201 16th Ave E	47.620243	-122.311679	1.757514
4	Seattle Children's Hospital	Hospital	4800 Sand Point Way NE	47.662590	-122.282741	4.255469
5	Kaiser Permanente Hospital	Hospital	11511 NE 10th St	47.618550	-122.186462	7.607533

Before Renaming



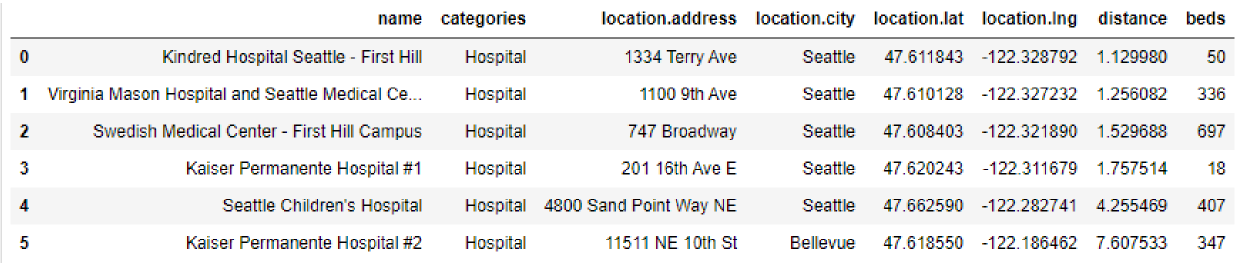
	name	categories	location.address	location.city	location.lat	location.lng	distance
0	Kindred Hospital Seattle - First Hill	Hospital	1334 Terry Ave	Seattle	47.611843	-122.328792	1.129980
1	Virginia Mason Hospital and Seattle Medical Ce...	Hospital	1100 9th Ave	Seattle	47.610128	-122.327232	1.256082
2	Swedish Medical Center - First Hill Campus	Hospital	747 Broadway	Seattle	47.608403	-122.321890	1.529688
3	Kaiser Permanente Hospital #1	Hospital	201 16th Ave E	Seattle	47.620243	-122.311679	1.757514
4	Seattle Children's Hospital	Hospital	4800 Sand Point Way NE	Seattle	47.662590	-122.282741	4.255469
5	Kaiser Permanente Hospital #2	Hospital	11511 NE 10th St	Bellevue	47.618550	-122.186462	7.607533

After Renaming

Given the small size of the data frame, the number of hospitals beds for each facility was manually determined from the web pages listed above. Bed count was merged to create the final data frame.

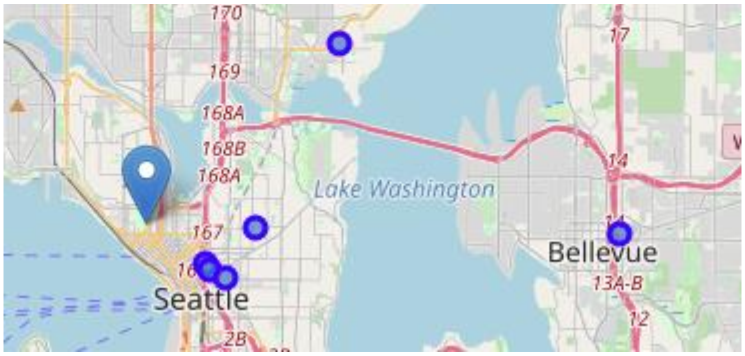
Results

Based on the selection criteria, six acute care hospitals were identified within a twenty-mile radius of the Space Needle, as displayed below.



	name	categories	location.address	location.city	location.lat	location.lng	distance	beds
0	Kindred Hospital Seattle - First Hill	Hospital	1334 Terry Ave	Seattle	47.611843	-122.328792	1.129980	50
1	Virginia Mason Hospital and Seattle Medical Ce...	Hospital	1100 9th Ave	Seattle	47.610128	-122.327232	1.256082	336
2	Swedish Medical Center - First Hill Campus	Hospital	747 Broadway	Seattle	47.608403	-122.321890	1.529688	697
3	Kaiser Permanente Hospital #1	Hospital	201 16th Ave E	Seattle	47.620243	-122.311679	1.757514	18
4	Seattle Children's Hospital	Hospital	4800 Sand Point Way NE	Seattle	47.662590	-122.282741	4.255469	407
5	Kaiser Permanente Hospital #2	Hospital	11511 NE 10th St	Bellevue	47.618550	-122.186462	7.607533	347

Final Data Frame

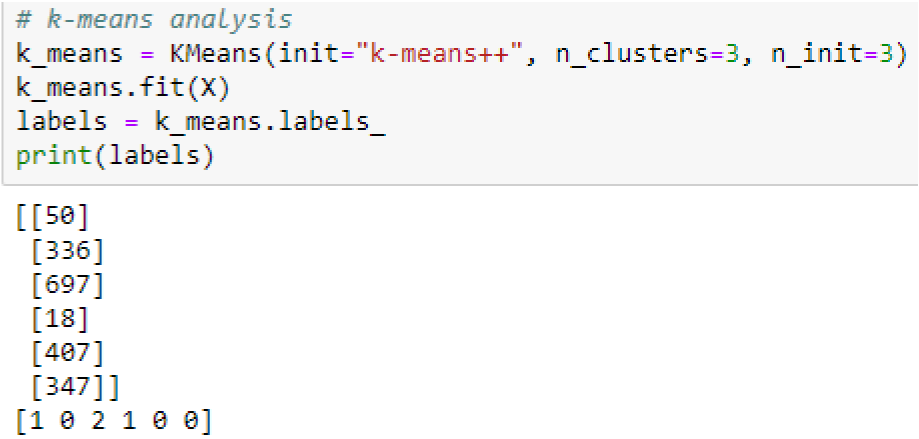


Hospital Plot

The second objective of this project was to determine whether hospitals would cluster based on number of beds. Since the analysis data set is so small, visual examination suggests that hospitals appear to cluster into three groups.

- Small (2): Kindred Hospital; Kaiser Permanente Hospital #1
- Medium (3): Virginia Mason; Seattle Children’s; Kaiser Permanente Hospital #2
- Large (1): Swedish Medical Center

To confirm this observation, and to illustrate the use of an unsupervised machine learning algorithm, a k-means cluster analysis was performed using n_clusters=3 and n_init=3 (followed by n_int=4 and n_int=5). All three analyses produced identical results, as displayed below.



```
# k-means analysis
k_means = KMeans(init="k-means++", n_clusters=3, n_init=3)
k_means.fit(X)
labels = k_means.labels_
print(labels)
```

```
[[50]
 [336]
 [697]
 [18]
 [407]
 [347]]
[1 0 2 1 0 0]
```

K-Means Analysis

Hospital	Beds	Manually Determined Cluster	K-Means Cluster
Kindred	50	Small	1
Virginia Mason	336	Medium	0
Swedish	697	Large	2
Kaiser #1	18	Small	1
Children’s	407	Medium	0
Kaiser #2	347	Medium	0

Comparison of Results

While the use of k-means was clearly not required to cluster these six hospitals, it was employed in the context of this project to illustrate an example of unsupervised machine learning and to corroborate the manual cluster determinations.

Discussion

As anticipated, working with address data can be inherently “messy”, and that was certainly the case with this project. Care must be taken to closely scrutinize the data at each step of the process, and respond as appropriate, whether that be data deletion, reformatting, or otherwise.

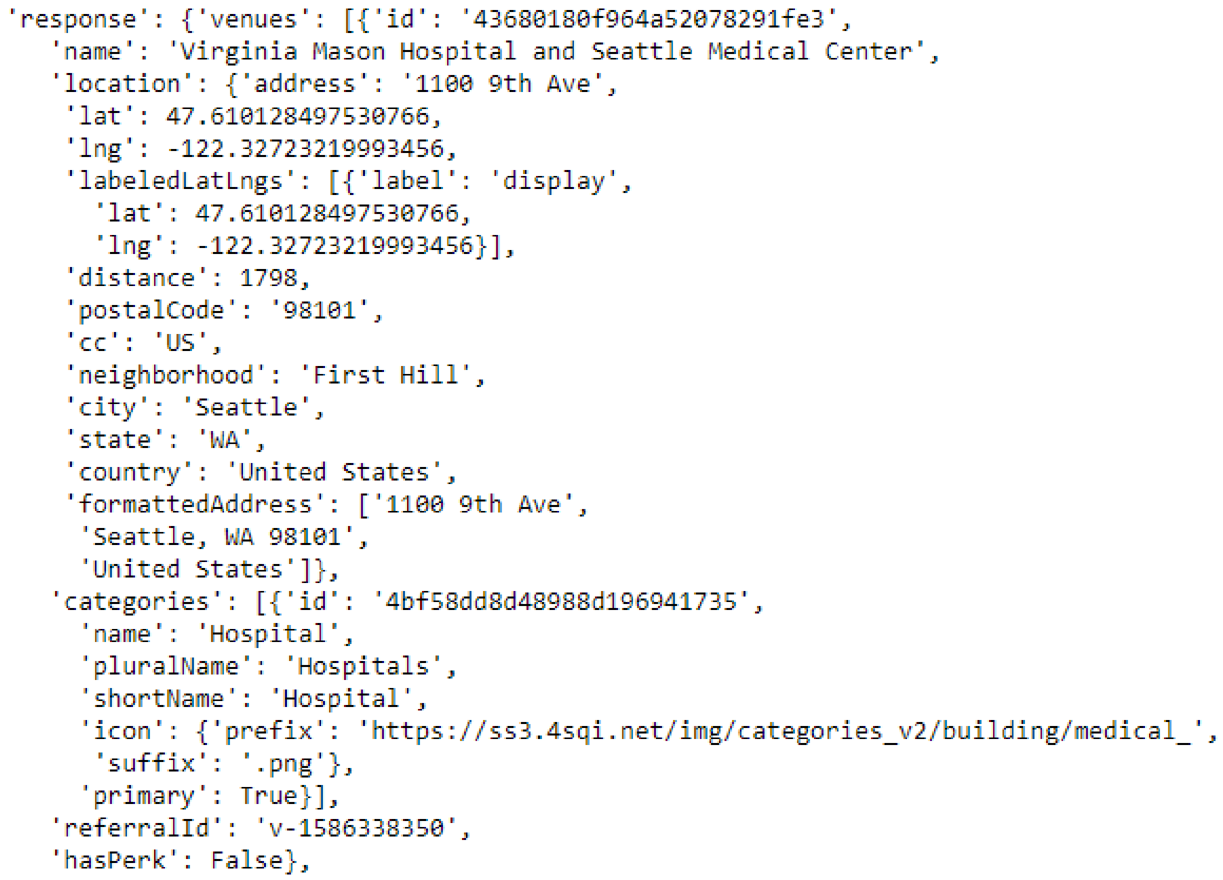
Conclusion

The two objectives of this project were completed, with the following results:

- Six acute hospitals were identified within a twenty-mile radius of the Space Needle.
- The six hospitals clustered into three groups based on bed count.

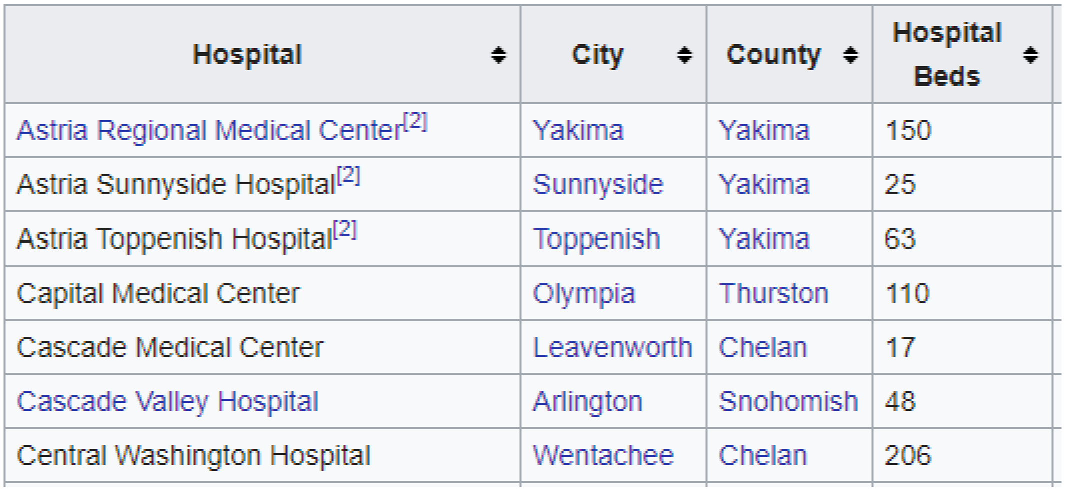
Appendix

Example of Foursquare data:



```
'response': {'venues': [{'id': '43680180f964a52078291fe3',
   'name': 'Virginia Mason Hospital and Seattle Medical Center',
   'location': {'address': '1100 9th Ave',
    'lat': 47.610128497530766,
    'lng': -122.32723219993456,
    'labeledLatLngs': [{'label': 'display',
      'lat': 47.610128497530766,
      'lng': -122.32723219993456}],
    'distance': 1798,
    'postalCode': '98101',
    'cc': 'US',
    'neighborhood': 'First Hill',
    'city': 'Seattle',
    'state': 'WA',
    'country': 'United States',
    'formattedAddress': ['1100 9th Ave',
     'Seattle, WA 98101',
     'United States']},
   'categories': [{'id': '4bf58dd8d48988d196941735',
     'name': 'Hospital',
     'pluralName': 'Hospitals',
     'shortName': 'Hospital',
     'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/building/medical_',
      'suffix': '.png'},
     'primary': True}],
   'referralId': 'v-1586338350',
   'hasPerk': False},
```

Example of hospital bed data:



Hospital ⬍	City ⬍	County ⬍	Hospital Beds ⬍
Astria Regional Medical Center[2]	Yakima	Yakima	150
Astria Sunnyside Hospital[2]	Sunnyside	Yakima	25
Astria Toppenish Hospital[2]	Toppenish	Yakima	63
Capital Medical Center	Olympia	Thurston	110
Cascade Medical Center	Leavenworth	Chelan	17
Cascade Valley Hospital	Arlington	Snohomish	48
Central Washington Hospital	Wentachee	Chelan	206