Capstone Project:

Identification of the Best Location for a Medical Practice Based on Professional and Personal Data

Class: Applied Data Science Capstone

William Sanborn

July 3, 2020

Introduction

The purpose of this project is to build a data analysis model that provides prioritized recommendations to an individual as to which city in the United States he or she might want to live. This model considers the individual's personal and professional interests in making its recommendation(s).

For purposes of this project, I am assuming the following:

- The individual is a cardiologist and, in order to ensure a successful medical practice, he/she would like to set up the practice in an area that has a significant need for heart related medical services.
- The cardiologist would also set like to set up his/her practice in an area that features as many areas of personal interest as possible. For purposes of this analysis, I am further assuming that:
    - The doctor enjoys sushi, wine, museums, and live music.
    - The doctor is an avid golfer.
    - The doctor has a dog that requires regular exercise.

The model described herein uses publicly available data sources and data analysis tools from Coursera to obtain and refine data and subsequently provide the desired recommendations.

Data

As described in the introduction, my goal is to build a model that provides recommendations based on personal and professional interests. To address this challenge, it was necessary to design metrics that measure the following:

- The level of need in a city for heart related medical services.
- The attractiveness of the identified city based on the doctor's personal interests

To gauge the need for heart related services, I used information from the Center for Disease Control's "500 City" database (https://chronicdata.cdc.gov/500-Cities/500-Cities-Coronary-heart-disease-among-adults-age/cqcq-r6f8/data). This dataset contains information related to incidences of heart disease (Angina and Coronary Heart Disease) in 500 major us cities. This information was then used to identify data and metrics that identify the 25 cities in the United State with the worst levels of heart health. For purposes of this exercise, these 25 "unhealthy" cities are considered to be the locations with the highest need for heart related services and, therefore, the cities in which the doctor is most likely to enjoy a successful practice.

To measure how well these cities align to my assumed cardiologist personally, I used the Foursquare API to build a database sampling all of the venues in each of the 25 cities that I had previously identified. From this data, I was able to extract information on only those venues that align to the doctor's areas of interest. I was then able design new metrics that reflect the frequency/concentration of type of interesting venues in each of the targeted cities. I then consolidated these calculations and built a final metric that indicates the potential overall interest "density" of a city. In my opinion, this an effective metric to convey the likelihood that the doctor will enjoy living in the city.

Foursquare API relies on GeoLocation coordinates (Latitude, Longitude) to identify nearby venues. As such, I needed to obtain these coordinates for the 25 cities that I intend to explore. The data for this purpose was obtained from a publicly available database: OpenDataSoft: https://developer.foursquare.com/

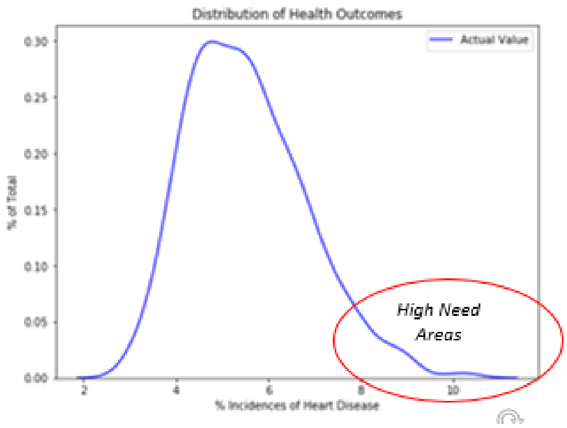
Please note that various data analysis tools were required to generate useable information sets from public data sources. Listings of both data sources and analytical tools are provided as exhibits to this report.

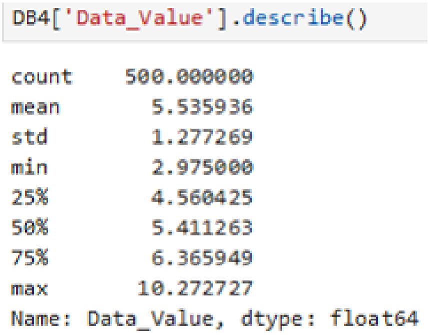
Inferential Statistical Testing

I considered numerous approaches to evaluate whether it would be possible to use inferential statistical methodologies to build a model that would provide the desired recommendations. After evaluating the CDC data, I was unable to identify a series of independent X variables that could be used to logically predict a dependent Y variable. Hence, I determined that a regression model was not suitable for this project. In addition, I did not have any historical data from which I could build training or test sets to design a classification model. This meant that commonly used predictive techniques (such as K Nearest Neighbor, decision trees, or logistic regression) would not feasible for this project. Clearly, a new methodology would be required to solve the problem.

Regardless, I did perform some statistical analysis using the Seaborn library to see if I could glean some insights into the data.

Exhibit: Distribution Curve of Heart Disease by City





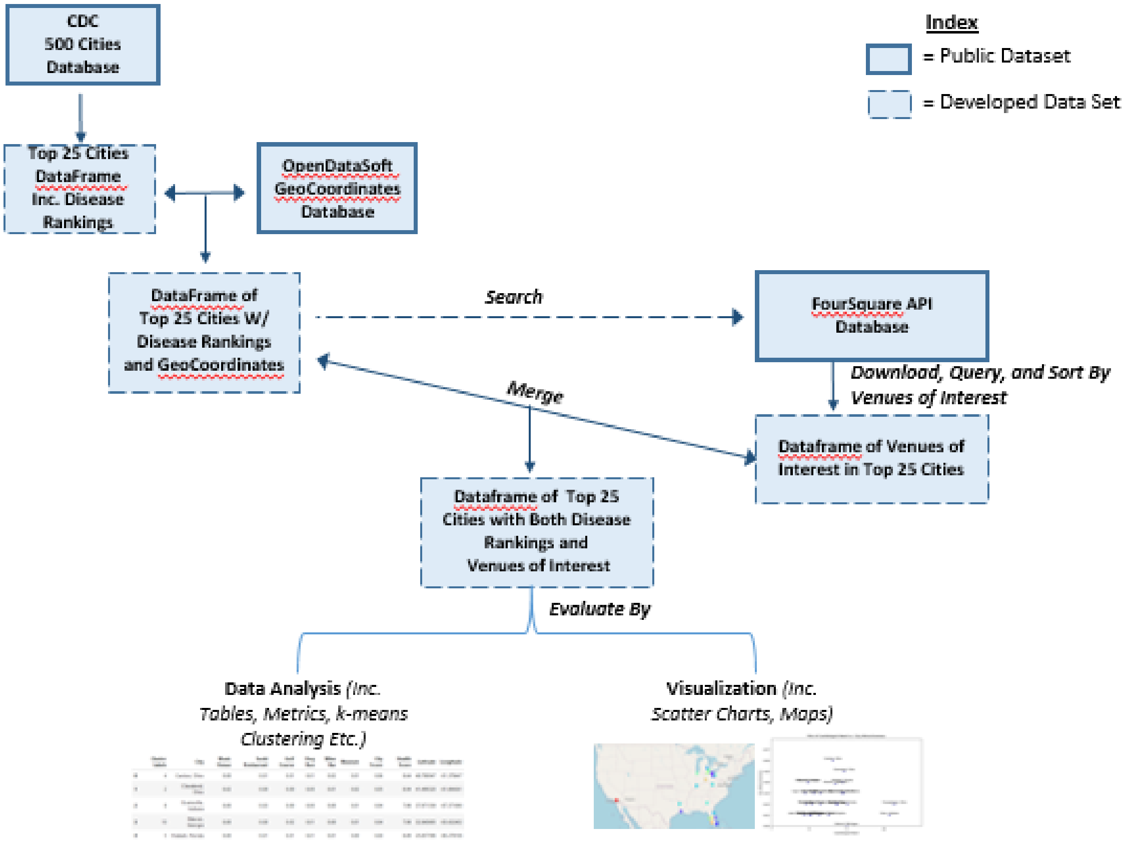
```
DB4['Data_Value'].describe()

count    500.000000
mean       5.535936
std        1.277269
min        2.975000
25%        4.560425
50%        5.411263
75%        6.365949
max       10.272727
Name: Data_Value, dtype: float64
```

This analysis of the CDC data demonstrates that a significant number of cities exist that have a relatively high need for heart related medical services. This understanding helped me design a methodology to identify these cities and subsequently address the question that I have posed.

Methodology

The following exhibit illustrates the methodology that I developed to process and analyze my selected data sources.

Exhibit: Model Methodology and Information Flow


CDC 500 Cities Database
Index
= Public Dataset
= Developed Data Set
Top 25 Cities DataFrame Inc. Disease Rankings
OpenDataSoft GeoCoordinates Database
DataFrame of Top 25 Cities W/ Disease Rankings and GeoCoordinates
Search
FourSquare API Database
Download, Query, and Sort By Venues of Interest
Merge
Dataframe of Venues of Interest in Top 25 Cities
Dataframe of Top 25 Cities with Both Disease Rankings and Venues of Interest
Evaluate By
Data Analysis (Inc. Tables, Metrics, k-means Clustering Etc.)
Visualization (Inc. Scatter Charts, Maps)

As shown in this exhibit, the primary data output of this process is a dataframe that contains both the metrics of heart health as well as the metrics of city attractiveness. This final dataframe was built from a combination of a separate city health dataframe (CDC data) and a separate city attractiveness dataframe (Foursquare data).

To build the city health dataframe, I used the Pandas library to refine the information I downloaded from the CDC database. This required the use of numerous functions from the Pandas library including groupby, sort, merge, join, split and astype. For example, the Latitude and Longitude information was provided as a string. In order for Foursquare and Folium to be able to read this information, I had to convert the data into a “float” number using the “astype” function.

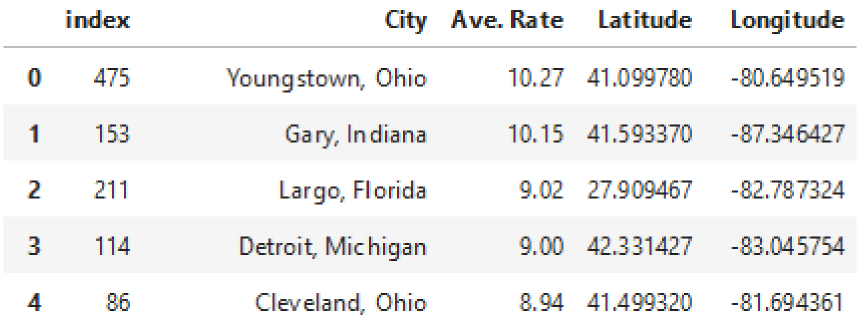
The specific steps I used are as follows:

- Group By City (Average ‘Data_Value’)
    - Average Data_Value=“Ave. Rate”
- Sort By Ave. Rate

    - Ascending = 'False'
- Limit = Top 25
- Merge With GeoLocation Data
    - Coordinates: String➔Float

Once a useable dataframe was created, I needed to figure out how to use this information to identify the cities in which my doctor might thrive professionally. After initial review, I decided that the "Data_Value" field provided in the CDC data might be a good metric of the heart health in a particular city. I subsequently reviewed the CDC web site and learned that this metric shows the instances of certain types of heart disease (angina or coronary heart disease) per 100 citizens in the affected area. Based on this description, I decided that "Data_Value" would be my metric of city heart health going forward.

Here are is the result of the process to create a city health dataframe that includes rankings of cities by incidences of heart disease (df.head() only):



	index	City	Ave. Rate	Latitude	Longitude
0	475	Youngstown, Ohio	10.27	41.099780	-80.649519
1	153	Gary, Indiana	10.15	41.593370	-87.346427
2	211	Largo, Florida	9.02	27.909467	-82.787324
3	114	Detroit, Michigan	9.00	42.331427	-83.045754
4	86	Cleveland, Ohio	8.94	41.499320	-81.694361

The second dataframe that I needed portrayed city attractiveness using data obtained through Foursquare API. This process involved sending a request using the 'Requests' library. The request specified venues within 16,000 meters (~10 miles) of the latitude and longitude coordinates of the city and had a limit of 100 which was based on my subscription. The data file that I received back from Foursquare had several challenges associated with it. First and foremost, the information was provided by Foursquare as a JSON file. As such, several steps had to be taken to turn the information into a useable Pandas dataframe.   These steps incorporated the use of functions from the JSON library.

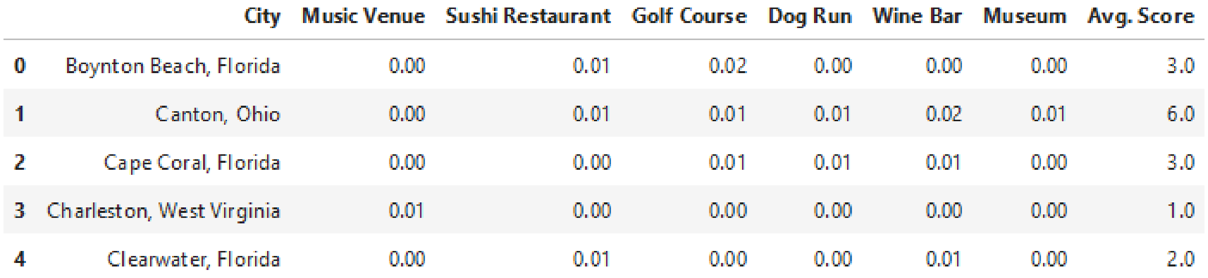
Once the initial dataframe was normalized and useable, I pulled the information/fields related for these types of venues into a new, separate dataframe:

- 'Music Venue'

- 'Sushi Restaurant'
- 'Golf Course'
- 'Dog Run'
- 'Wine Bar'
- 'Museum'

After completion of this step, I now had a new dataframe that included only venues of interest to the cardiologist. Unfortunately, the Foursquare information did not provide a useable metric (like the CDC information did) to measure city attractiveness. Hence, this metric had to be built. To address this challenge, I used "onehot" coding to adjust the dataframe so that I could see the number of times that a certain type of venues appeared in a Foursquare query. I then took the "onehot" data and built a calculation that showed the percent of times that a given type venue appears in a query for a given city. By summing this calculation for all venues of interest in a city, I was able to create a summary metric that conveys the frequency in which all venues of interest appear within a single city query. It is my opinion that this metric effectively measures the level of personal interest for a given city. As such, I included this metric, "City Score", in my city attractiveness dataframe going forward.

Here is the result of the process to rank cities by levels of personal interest (df.head() only):



	City	Music Venue	Sushi Restaurant	Golf Course	Dog Run	Wine Bar	Museum	Avg. Score
0	Boynton Beach, Florida	0.00	0.01	0.02	0.00	0.00	0.00	3.0
1	Canton, Ohio	0.00	0.01	0.01	0.01	0.02	0.01	6.0
2	Cape Coral, Florida	0.00	0.00	0.01	0.01	0.01	0.00	3.0
3	Charleston, West Virginia	0.01	0.00	0.00	0.00	0.00	0.00	1.0
4	Clearwater, Florida	0.00	0.01	0.00	0.00	0.01	0.00	2.0

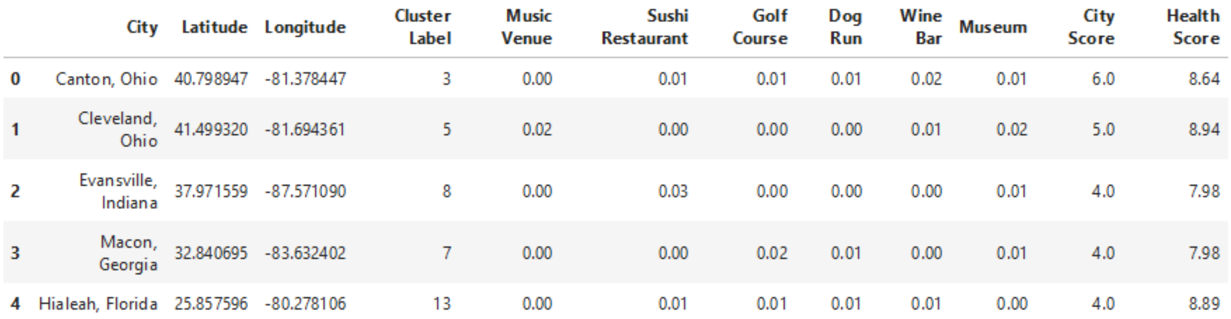
In addition to the aforementioned actions, I used Kmeans clustering to assess whether there were potential clusters that might provide insight into where the doctor might want to live. For this purpose, I used the Kmeans library from sklearn to cluster the data and generate cluster labels. After this analysis was complete, I added the clustering information to my database.

Once I was satisfied with each of these dataframes, I used the "join" function to create a consolidated dataframe. Now it was time to analyze the data and develop a recommendation.

Results

The following shows a sample of my consolidated dataframe including cluster labels:

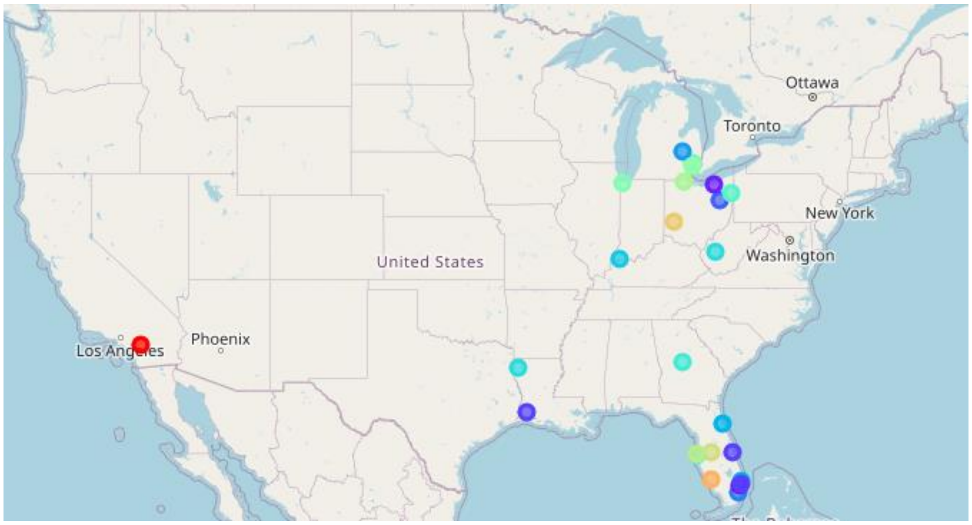
Exhibit: Final DataFrame



	City	Latitude	Longitude	Cluster Label	Music Venue	Sushi Restaurant	Golf Course	Dog Run	Wine Bar	Museum	City Score	Health Score
0	Canton, Ohio	40.798947	-81.378447	3	0.00	0.01	0.01	0.01	0.02	0.01	6.0	8.64
1	Cleveland, Ohio	41.499320	-81.694361	5	0.02	0.00	0.00	0.00	0.01	0.02	5.0	8.94
2	Evansville, Indiana	37.971559	-87.571090	8	0.00	0.03	0.00	0.00	0.00	0.01	4.0	7.98
3	Macon, Georgia	32.840695	-83.632402	7	0.00	0.00	0.02	0.01	0.00	0.01	4.0	7.98
4	Hialeah, Florida	25.857596	-80.278106	13	0.00	0.01	0.01	0.01	0.01	0.00	4.0	8.89

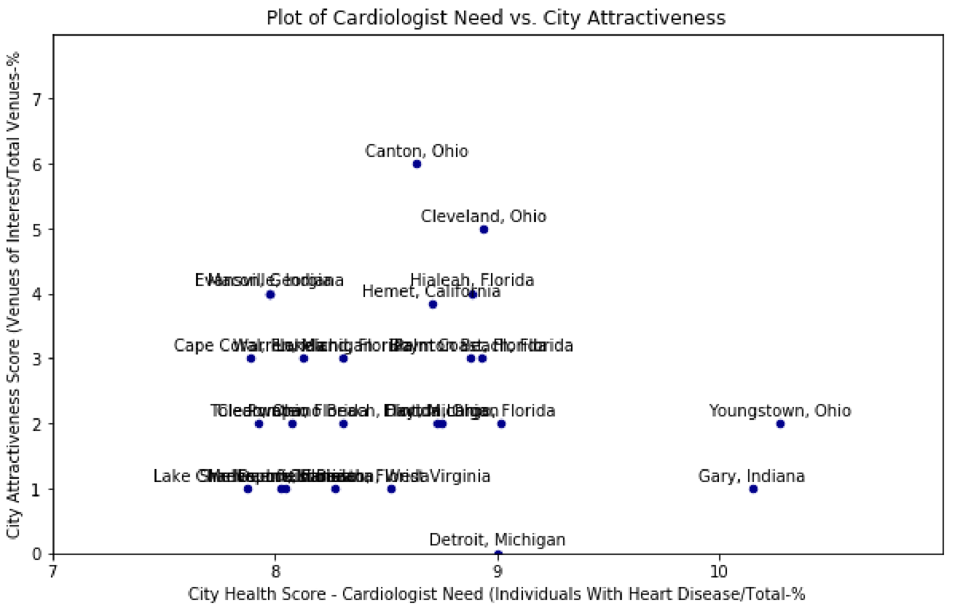
While this table provides useful information, I decided that it did not convey the information in such a way so as to identify a final recommendation. Hence, I decided to use additional visualization tools to see if they could provide an answer. First, I generated a map using Folium to see if it might provide any insights as to where the cardiologist might want to live.

Exhibit: Folium Map With Kmeans Clustering Labeling



The answer was still not evident to me so I decided to also run a scatter chart that conveyed Heart Health per city (”City Health Score”) on the X axis versus city attractiveness (”City Attractiveness Score”) on the Y axis.   This scatter chart is shown in the following exhibit:

Exhibit: Scatter Chart of City Heart Health vs. City Person Interests



After reviewing all these results, I decided that I had now generated adequate information to make a final recommendation to the cardiologist.

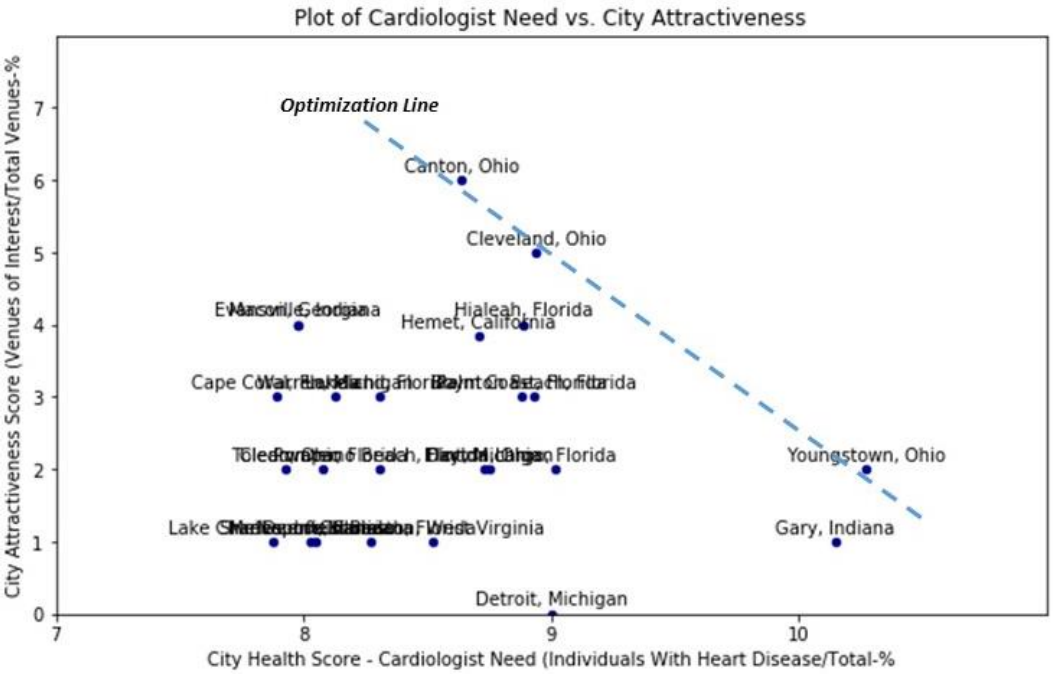
Discussion

The various results and outputs of the model convey varying levels of insight into the targeted objective of identifying where the assumed cardiologist might want to establish a practice and live. For example, the data table summarizes the key metrics but does not provided a framework to provide an optimal decision. Similarly, the Folium map provides a helpful visualization of where heart disease is concentrated in the US (Midwest US and southern Florida) but it does not succinctly show the cities that have highest correlation to the cardiologists personal interests and preferences.

In my opinion, the scatter chart *does* effectively portray how various cities align to both the Health Score and City Score metrics. For example, the chart shows that Canton, Ohio has the highest amount of desired venues among the top 25 cities. Clearly, Canton has the most to offer the doctor in terms of personal interests. On the other hand, Youngstown, Ohio has the highest level of heart disease so the doctor can assume that his or her services would be in significant demand there. Unfortunately, Youngstown provides a very low "City Score" so it seems unlikely that doctor would be happy living there on a personal level.

To gain additional insight, I also decided to draw an optimization line to see if there might be any other candidates. This step showed that Cleveland, Ohio might a good third option because it provides an improved balance of personal and professional interest.

Scatter Chart of City Heart Health vs. City Personal Interests with Optimization Line



Based on this visualization, I believe that either of Canton, OH, Cleveland, OH, or Youngstown, OH would be a viable candidate city for the doctor.

Obviously, this methodology provided a set of recommendations rather than a single option. To identify a single choice, I would want to speak with the doctor and get a better sense of how he or she might prioritize the personal or professional attributes of city. If I knew this, I could apply weights to these choices and determine a single output. Please note that my model has other items worthy of further refinement. One major issue is that the number of venues returned by the FourSquare API is limited to 100 based on my current subscription. There is some risk that this limited data set might not accurately portray the frequency of interesting venues within a city. As such, increasing the number of venues could potentially increase the accuracy of the personal attributes of the model.

Conclusion

In summary, the model described herein effectively provides a framework for identifying and evaluating potential places for an individual to live based on that individual's professional and personal preferences. The model achieves this goal through the use of publicly available data (such as FourSquare API) and Python based data analysis and visualization tools.

After building the model, the author set out to validate the results against the baseline assumptions of a hypothetical cardiologist with a defined set of personal preferences. Once run, the model outputs suggested that Canton, OH, Youngstown, OH, and Cleveland OH are cities that the cardiologist should target. In the author's opinion, these recommendations effectively align with both the data obtained from public sources as well as with the assumptions used to build the model. It should be noted that the model could be further refined through additional clarification of the doctor's personal and professional objectives as well as through increasing the limit of venue information received from Foursquare. Regardless, it is the author's opinion that the model successfully provides useful, supportable information that an individual can use to select a city in which to live and work.

EXHIBITS

Data Sources

1) **Information on Health Issues By City:** https://chronicdata.cdc.gov/500-Cities/500-Cities-Coronary-heart-disease-among-adults-age/cqcq-r6f8/data
Note: Measure: % Respondents aged ≥18 years who report ever having been told by a doctor, nurse, or other health professional that they had angina or coronary heart disease.
https://www.cdc.gov/500cities/definitions/health-outcomes.htm

2) **Database of Geographic (Latitude/Longitude) Coordinates By City:** https://public.opendatasoft.com/explore/dataset/1000-largest-us-cities-by-population-with-geographic-coordinates/table/?sort=-rank

3) **Database(s) of City Venue Detail: *FourSquare API*:** https://developer.foursquare.com/
FOURSQUARE

Libraries Used

Folium: Visualization and mapping

Geocoder/Nominatim: Generate and Read Location Data.

JSON: Analyze JSON files

Matplotlib: Python plotting and graphing

Numpy: Arrays and Data Set Functions

Pandas: Misc. DataFrame Functions

Requests: Generate API requests

Seaborn: Statistical Data Visualization

SkLearn: K-means clustering.

______________.

Please Note: The data sources I used did not require the use of Beautiful Soup for data scraping