Applied Data Science Capstone Project – Week 2
IBM Coursera Data Science Specialization

BATTLE OF THE NEIGHBORHOODS
CAPSTONE PROJECT REPORT

A. Introduction

1. Background

Mr. John Smith is a senior executive at ACME Widgets Inc in Southern California, USA. ACME has experienced severe downturn in its business due to the COVID-19 pandemic and is planning to close its facilities in the state. The company has offered Mr. Smith the choice of being laid off or relocating to its offices in the city of Austin in the state of Texas. Mr. Smith has opted to relocate to Austin.

Austin is a vibrant and dynamic city and is the capital of Texas. It has a population of over 2 million. It is home to the University of Texas and is known for its parks, and lakes, and outdoor activities. The city is a center for performing arts, especially live music, documented in the long running PBS music series *Austin City Limits*. Many of the technology majors have facilities in the city. See Austin's Wikipedia entry for more information.

2. The Problem

Mr. Smith has visited Austin a couple of times for business. He is not familiar with the neighborhoods in the city. Ideally, he would like to rent or buy a home in a neighborhood similar to the one he currently resides in - Westpark, in the city of Irvine in California. He could consult with real estate agents (realtors) in Austin for their opinions on neighborhoods and make a choice based on the feedback. However, the opinions are likely to be biased or highly subjective. It is also unlikely that any of the Austin realtors know about the Westpark neighborhood. Therefore, Mr. Smith would like a recommendation on Austin neighborhoods based on data analysis, which he can compare with the recommendations of realtors in the city.

B. The Data

To provide Mr. Smith with an objective, analytical recommendation, geospatial data from Foursquare and other sources will be used in the analysis.

1. Data Sources

The following datasets will be used:

Data Description	Data Source
1. Neighborhoods in Austin	The city of Austin has a comprehensive data portal. One of the datasets available is a list of neighborhoods in GEOJSON format. The URL for the dataset is here.
2. Geographical coordinate data for neighborhoods in Austin	Nominatim
3. Geographical coordinate data for Westpark	Nominatim
4. Top 100 venues for Austin neighborhoods	Foursquare
5. Top 100 venues for Westpark	Foursquare

2. Data Preparation

The character and nature of a neighborhood is defined not only by the people living in the community, but also by the businesses, venues, and organizations that operate in the area. The starting point was to get a comprehensive list of Austin neighborhoods. A list of 113 Austin neighborhood names was acquired from a geoJSON file on the site. The latitude and longitude coordinates in the file are for multiplot polygons. Therefore, a single latitude and longitude value for the neighborhoods was acquired from Nominatim, a provider of opensource geocoding data. The geographical coordinates for Westpark were acquired similarly.

Nominatim did not have latitude and longitude data on 14 neighborhoods listed in the Austin data file. These were dropped for the analysis.

Data on top 100 venues in each of the remaining neighborhoods was acquired from Foursqure, a location data platform using their application programming interface (API). Of the 89 remaining neighborhoods, Foursquare did not return venues for 3 neighborhoods which where also dropped, leaving 86 Austin neighborhoods.

The data for Westpark was appended to the Austin data.

C. Methodology

1. Analytical Approach

The objective is to classify Austin neighborhoods into groups or clusters based on the features of each neighborhood. This will identify the types of neighborhoods in Austin. Since we are trying to identify groups that have not been labeled, the K-means clustering algorithm will be used. By including Westpark as an "Austin neighborhood", we will be able to see which cluster it is included in for a given optimal K clusters. The optimal K will be determined using the elbow method.

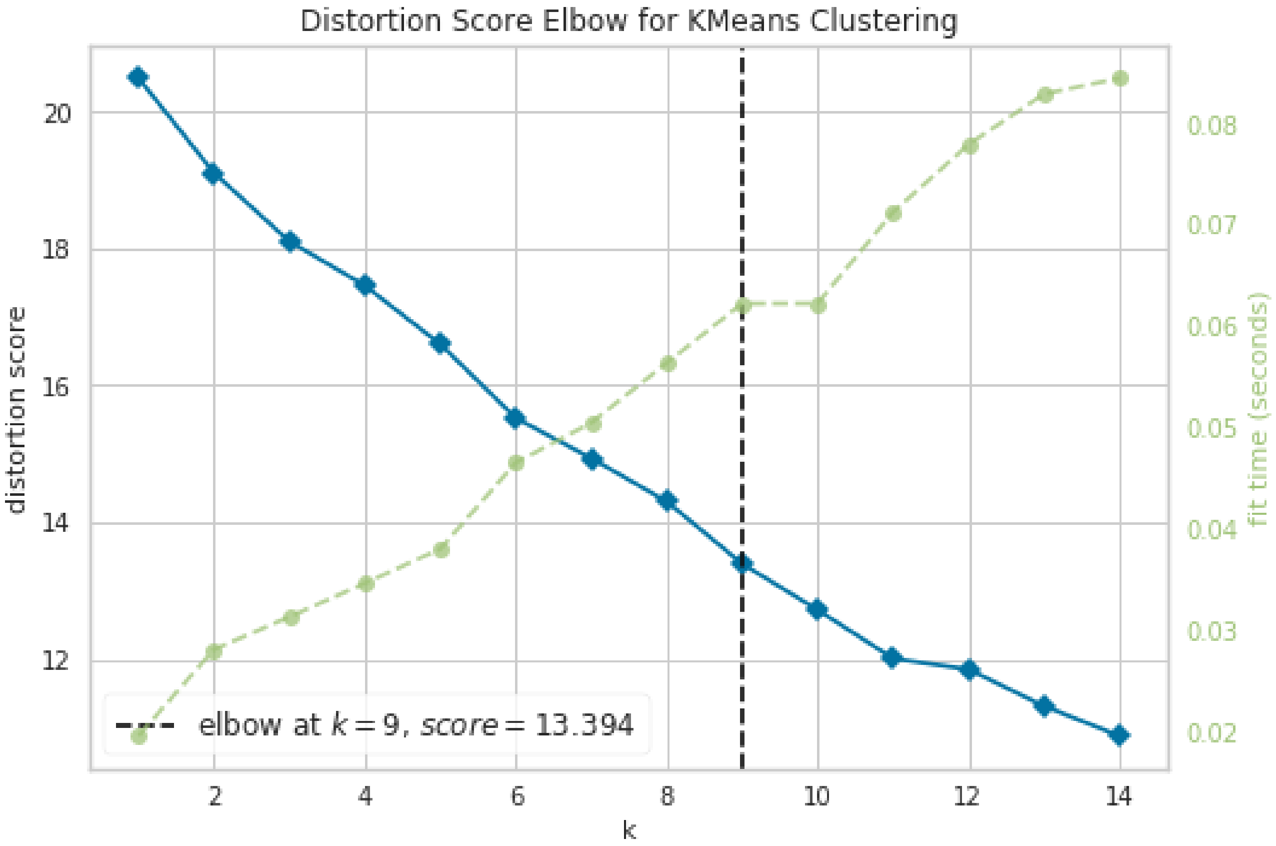
The list of Austin neighborhoods with the same cluster label as Westpark will be recommended to Mr. Smith as the most likely candidates for his relocation plans.

2. Data Analysis

   a. Fit the data to K-means for 1 to 15 clusters

   b. Identify optimal number of clusters using elbow method

The *yellowbrick* Python library was used to run K-means for range of cluster numbers (k) from 1 through 15. The *KElbowVisualizer* function displays the elbow value and score for each value of *k* on a chart. The elbow was identified at 9 clusters.

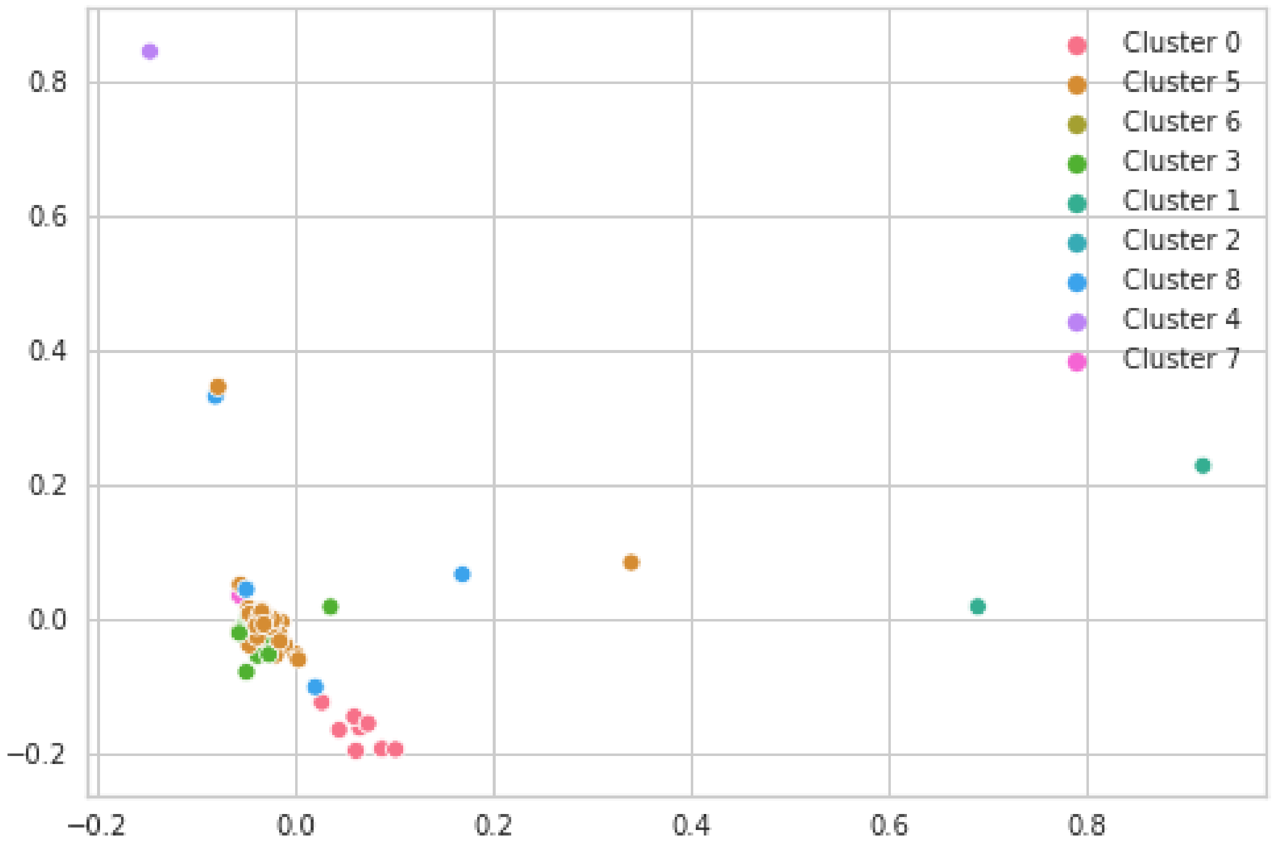


c. Fit the data to K-means for optimal number of clusters determined in step 2

The K-means clustering was re-run for 9 clusters.

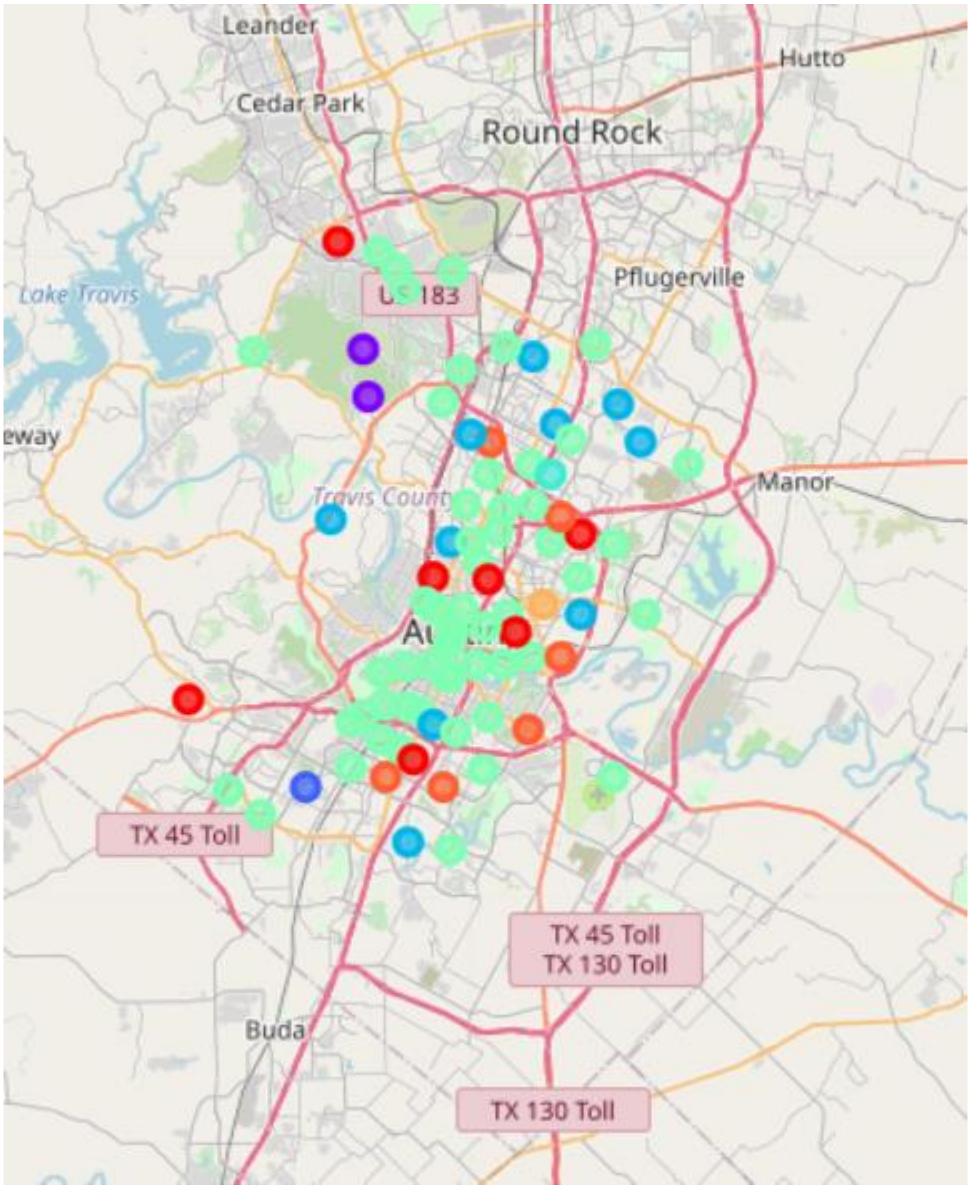
d. Visualize clusters on 2D chart

The data consists of top 10 venues (features) for each neighborhood. To view the clusters on a 2D chart, Principle Component Analysis (PCA) was used. The feature data was reduced to two dimensions and plotted.



e. Visualize location of clusters on map of Austin

   Using the *folium* library, the K-means clusters were plotted on a map of Austin



f. List the neighborhoods in Austin in the same cluster as Westpark, ranked in ascending order of distance to the cluster centroid.

   Square of distance to centroid was calculated for all neighborhoods. The Austin neighborhoods in the “Westpark cluster” were ranked in ascending order of distance to the cluster centroid.

D. Results

neighborhood	latitude	longitude	cluster labels	sqdist	1st Most Common Venue
Westpark, Irvine CA	33.6913524	-117.8088444	0		Playground
WEST OAK HILL	30.2384802	-97.8890123	0	5.35	Brewery
EAST CONGRESS	30.2103976	-97.7660519	0	5.39	Restaurant
HANCOCK	30.2958956	-97.7247678	0	5.4	Park
OLD WEST AUSTIN	30.296822	-97.7548514	0	5.52	Park
UNIVERSITY HILLS	30.3175801	-97.6739168	0	7.05	Bridal Shop
ANDERSON MILL	30.4558345	-97.8070957	0	7.95	Park
ROSEWOOD	30.2713704	-97.7101117	0	9.57	Park

	neighborhood	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
0	Westpark, Irvine CA	Park	Other Repair Shop	Pool
1	WEST OAK HILL	Playground	Park	Farm
2	EAST CONGRESS	Convenience Store	Taco Place	Park
3	HANCOCK	Mexican Restaurant	Golf Course	Yoga Studio
4	OLD WEST AUSTIN	Shop & Service	Food Truck	Tanning Salon
5	UNIVERSITY HILLS	Arts & Entertainment	Park	Fast Food Restaurant
6	ANDERSON MILL	Food Truck	Pool	Dog Run
7	ROSEWOOD	Pool Hall	CafÃ©	Gym / Fitness Center

	neighborhood	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
0	Westpark, Irvine CA	Moving Target	Escape Room	Food Truck
1	WEST OAK HILL	French Restaurant	Frame Store	Fountain
2	EAST CONGRESS	Yoga Studio	Escape Room	Fountain
3	HANCOCK	Farm	Frame Store	Fountain
4	OLD WEST AUSTIN	Yoga Studio	Farm	Frame Store
5	UNIVERSITY HILLS	Fried Chicken Joint	French Restaurant	Frame Store
6	ANDERSON MILL	Yoga Studio	Farm	Frame Store
7	ROSEWOOD	Soccer Field	Yoga Studio	Farm

	neighborhood	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Westpark, Irvine CA	Food Court	Food & Drink Shop	Food
1	WEST OAK HILL	Food Truck	Food Court	Food & Drink Shop
2	EAST CONGRESS	Food Truck	Food Court	Food & Drink Shop
3	HANCOCK	Food Truck	Food Court	Food & Drink Shop
4	OLD WEST AUSTIN	Fountain	Food Court	Food & Drink Shop
5	UNIVERSITY HILLS	Fountain	Food Truck	Food Court
6	ANDERSON MILL	Fountain	Food Court	Food & Drink Shop
7	ROSEWOOD	Fountain	Food Truck	Food Court

E. Discussion

The neighborhoods identified by K-means in the "Westpark Cluster" appear to share similar attributes - restaurants, parks, and fitness places. Mr. Smith will have a choice of neighborhoods to potentially relocate to. Of course, there is an emotional factor in choosing where to stay. As he starts visiting the neighborhoods in person, he may prefer a neighborhood that is not at the top of the list. He might even prefer a neighborhood in another cluster.

F. Conclusion

At a minimum, the list of neighborhoods above provides an excellent starting point for Mr. Smith's exploration of Austin. The neighborhoods are listed in ascending order of squared distance to the cluster centroid. In his exploration of Austin, he can start with neighborhoods at the top (closest to centroid) and go down the list.

G. Acknowledgments

I would like to thank the team at IBM and Coursera for putting together the Data Science Specialization. The courses in the specialization have demystified Data Science and Machine Learning and given me a good foundation for further studies in the field. I also want to thank my peers who responded to questions posted by me and others in the Discussion Forums.