COURSERA IBM DATA SCIENCE CERTIFICATE

CAPSTONE PROJECT

DETERMINING A NEIGHBORHOOD FOR A NEW WINE BAR IN DENVER

INTRODUCTION

- The objective of this project is to determine a location for a new business within the municipality of Denver, Colorado USA. This business will be a Wine Bar. It will feature a tasting menu as well as a variety of interesting wines from around the world. Wines will be available by the bottle, glass, or in flights. Craft beer has been popular in Colorado for a long time, but wine is also popular. Many of the wine bars are outside of Denver in the surrounding suburbs.

- Over the past ten years or so there have been many, trendy neighborhoods that have been attracting more affluent millennials and others who enjoy living in a bustling urban area close to parks and quality restaurants.

- The stakeholders want to capitalize on these trends and place a wine bar in one of these hot, trending neighborhoods. They would also consider a more established neighborhood with the correct demographics that is lacking a similar business.

DATA

- The data to be considered for this project will come from the following sources:

- A list of 25 distinct Denver Neighborhoods found in an article called The 25 Best Neighborhoods in Denver
- An article in the popular local magazine, 5280, called The 25 Best Neighborhoods in Denver
- Data downloaded utilizing the Foursquare API including: most popular venues and locations of wine bars
- Neighborhood Latitudes and Longitudes will be acquired using the Nominatim package from geopy.geocoders.
- Google Maps was used to fill in missing or incorrect data as discovered in the initial exploratory analysis.

DATA CONT

- The criteria for selecting a neighborhood for a new wine bar will include:

- Trending neighborhood - mandatory (all of them meet this criteria)
- Near a wine shop
- Not in a neighborhood with many fast food restaurants or discount stores - mandatory
- In a cluster that contains a neighborhood another wine bar - mandatory (clusters TBD in the Analysis section)
- In a neighborhood with other bars or breweries
- Doesn't already have a wine bar - mandatory
- Near a park

Here is the URL for the neighborhood list and data:

https://www.5280.com/neighborhoods/

The documentation for the Foursquare API is here: https://developer.foursquare.com/docs/places-api/

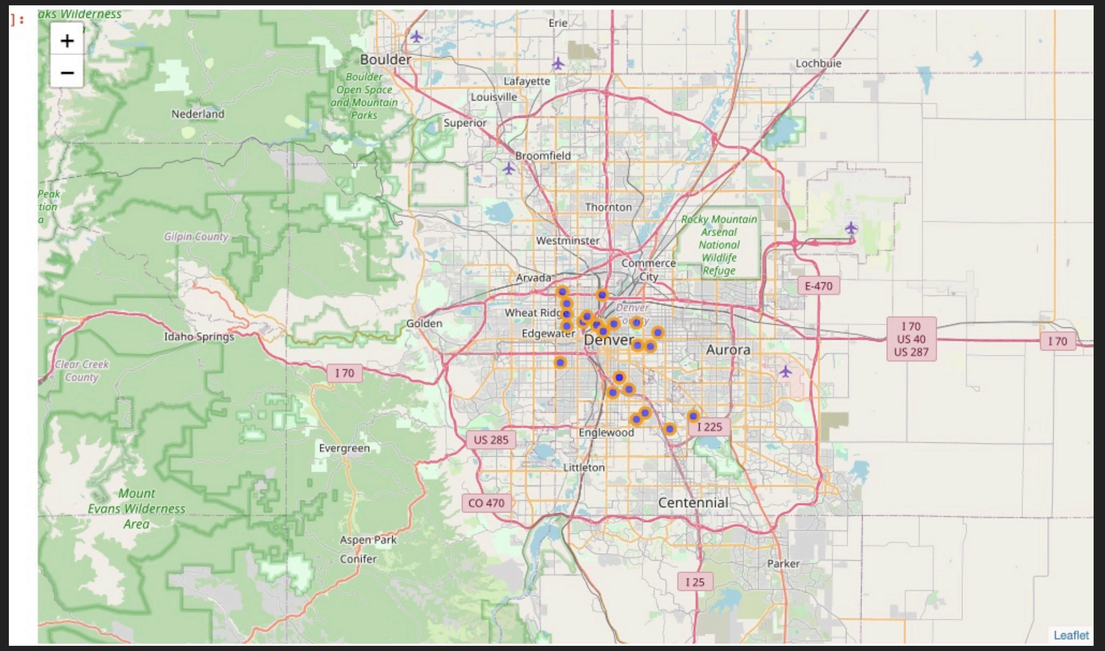
DATA CONT.

- I used Python's Beautiful Soup package to scrape the list of neighborhoods off of the 5280 article. I noticed it was contained in the table so I looped around to populate lists to create a data frame. I included some of their demographic data to get a little bit better picture of the neighborhoods. I ended up only using the list of names for the scope of this project. The end result was this data frame:

- Below the table is a map of the neighborhoods:

Out[408]:

	Neighborhoods	Rank	AvgSalePrice2019	CrimeRank	XFactorScore
0	South Park Hill	1	$ 804,250	8	8.5
1	Washington Park	2	$ 1,119,585	3	9
2	Congress Park	3	$ 680,522	15	9
3	West Highland	4	$ 661,257	25	8.5
4	Cherry Creek	5	$ 1,165,333	52	8.5
5	Speer	6	$ 505,815	44	8
6	Wellshire	7	$ 812,084	1	7
7	Highland	8	$ 700,576	51	9.5
8	Hilltop	9	$ 983,055	4	8
9	University Hills	10	$ 596,061	46	6.5
10	Berkeley	11	$ 651,844	22	9
11	Union Station	12	$ 823,351	72	9.5
12	Indian Creek	13	$ 335,564	2	4.5



METHODOLOGY

- For this project we are going too need to determine what kind of venues are popular in each neighborhood. Acquiring a Foursquare developer account is mandatory for this project. The advantages of using the Foursquare API is that it has generous API call limits for the free version. Once the Foursquare account is set up and credentials are established for Client ID and Client Secret, then the venue data can be acquired.

METHODOLOGY CONT

- Downloading the venues: Here is a clip of the process using the GetNearbyVenues function created in the Coursera Capstone lab.

```
top25_denver_venues = getNearbyVenues(names=top25_data['Neighborhood_address'],
                                      latitudes=top25_data['Latitude'],
                                      longitudes=top25_data['Longitude']
                                     )
```

```
South Park Hill, Denver, CO
Washington Park, Denver, CO
Congress Park, Denver, CO
West Highland, Denver, CO
Cherry Creek, Denver, CO
Speer, Denver, CO
Wellshire, Denver, CO
Highland, Denver, CO
Hilltop, Denver, CO
University Hills, Denver, CO
Berkeley, Denver, CO
```

- Once the venue data has been acquired and processed into a data frame, then a separate data frame can be created with only Wine Bars. The initial data frame had two wine bars with the exact same coordinates listed in two different neighborhoods. A Google maps search revealed which was the correct one and the duplicate was removed:

Part 3 - Find the wine bars in these neighborhoods

In [422]:
```
wine_bars = top25_denver_venues[top25_denver_venues['Venue Category'] == 'Wine Bar']
wine_bars
```

Out[422]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
287	Speer, Denver, CO	39.756877	-105.018439	The Truffle Table	39.758129	-105.011643	Wine Bar
412	Highland, Denver, CO	39.761583	-105.012500	The Truffle Table	39.758129	-105.011643	Wine Bar
591	Berkeley, Denver, CO	39.775231	-105.039261	BookBar	39.775213	-105.043888	Wine Bar
736	Union Station, Denver, CO	39.753630	-105.000748	Cru Wine Bar	39.747963	-104.998908	Wine Bar
907	Five Points, Denver, CO	39.754658	-104.977986	Mile High Winery	39.761417	-104.983637	Wine Bar
1237	Central Business District, Denver, CO	39.747378	-104.992737	Cru Wine Bar	39.747963	-104.998908	Wine Bar

There is a mistake on Foursquare. The Truffle Table, listed twice, is actually in the Highland neighborhood, not Speer. Removing that row.

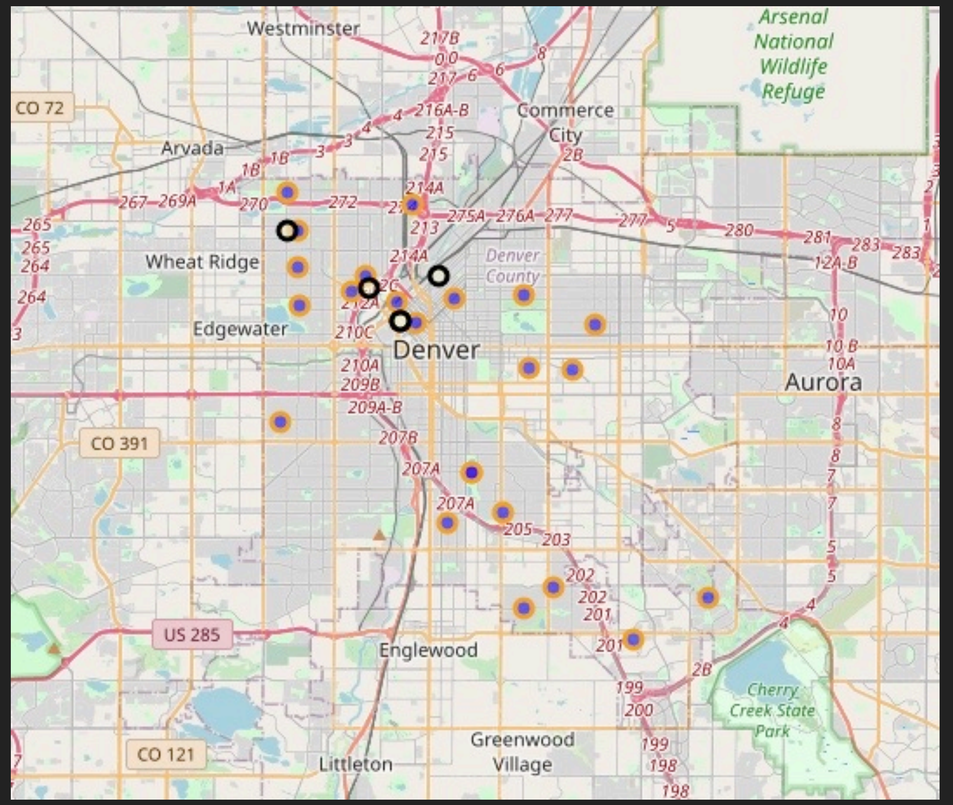
In [423]:
```
wine_bars = wine_bars.drop(index=287)
wine_bars
```

Out[423]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
412	Highland, Denver, CO	39.761583	-105.012500	The Truffle Table	39.758129	-105.011643	Wine Bar
591	Berkeley, Denver, CO	39.775231	-105.039261	BookBar	39.775213	-105.043888	Wine Bar
736	Union Station, Denver, CO	39.753630	-105.000748	Cru Wine Bar	39.747963	-104.998908	Wine Bar
907	Five Points, Denver, CO	39.754658	-104.977986	Mile High Winery	39.761417	-104.983637	Wine Bar
1237	Central Business District, Denver, CO	39.747378	-104.992737	Cru Wine Bar	39.747963	-104.998908	Wine Bar

METHODOLOGY CONT.

- Map of the neighborhoods with an overlay of the wine bars.
- The wine bars are the black circles with the yellow centers.
- The next step will be to create clusters of similar neighborhoods using Kmeans clustering.



PREPROCESSING BEFORE KMEANS

- The neighborhood and venue data will then be processed so that we can perform Kmeans clustering to group the neighborhoods into clusters based on their most popular venues. This was done using the one hot encoding method to reshape the data to get a count of each venue in each neighborhood as shown here:

- Next, a new data frame is created to include a column each for the top ten venue types in each neighborhood.

In [426]:
```
# one hot encoding
denver_onehot = pd.get_dummies(top25_denver_venues[['Venue Category']], prefix="", prefix_sep="")

# add neighborhood column back to dataframe
denver_onehot['Neighborhood'] = top25_denver_venues['Neighborhood']

# move neighborhood column to the first column
col_name="Neighborhood"
first_col = denver_onehot.pop(col_name)
denver_onehot.insert(0, col_name, first_col)

denver_onehot.head()
```

Out[426]:

	Neighborhood	ATM	Accessories Store	American Restaurant	Antique Shop	Aquarium	Arcade	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Automotive Shop	BBQ Joint
0	South Park Hill, Denver, CO	0	0	1	0	0	0	0	0	0	0	0	0	0	0
1	South Park Hill, Denver, CO	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	South Park Hill, Denver, CO	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	South Park Hill, Denver, CO	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	South Park Hill, Denver, CO	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Part 6 - Preprocess and perform kmeans clustering on the selected cluster from part 5

This is the cluster containing the neighborhoods with all of the existing wine bars

In [475]:
```
cluster1_neighborhoods = cluster1['Neighborhood_address']
cluster1_grouped = denver_grouped[denver_grouped['Neighborhood'].isin(cluster1_neighborhoods)]
cluster1_grouped.head()
```

Out[475]:

	Neighborhood	ATM	Accessories Store	American Restaurant	Antique Shop	Aquarium	Arcade	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Automotive Shop
0	Barnum West, Denver, CO	0.000000	0.0	0.125000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0
1	Berkeley, Denver, CO	0.000000	0.0	0.034884	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0
2	Central Business District, Denver, CO	0.000000	0.0	0.070000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.010000	0.000000	0.0
4	Congress Park, Denver, CO	0.011494	0.0	0.022989	0.0	0.0	0.0	0.0	0.0	0.0	0.011494	0.011494	0.000000	0.0
5	Cory-Merrill, Denver, CO	0.000000	0.0	0.058824	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.029412	0.058824	0.0

KMEANS CLUSTERING

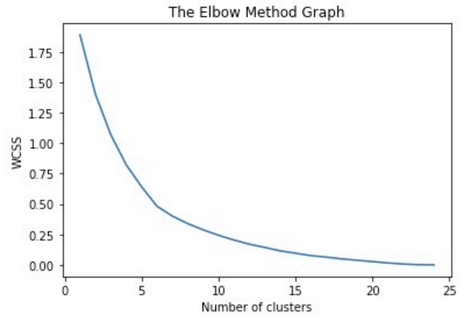
- Here is the processing for the Kmeans including methodology for the selection of the number of clusters and generating the cluster matrix:

Create the dataset for the Kmeans. Loop through 25 times to create the elbow graph to select the optimum number of clusters

In [433]:
```
X = denver_grouped.drop('Neighborhood', 1)
```

In [434]:
```
wcss = []
for i in range(1,25):
    kmeans = KMeans(n_clusters=i, init ='k-means++', max_iter=600,  n_init=25,random_state=0 )
    kmeans.fit(X)
    wcss.append(kmeans.inertia_)
```

In [435]:
```
import matplotlib.pyplot as plt
plt.plot(range(1,25),wcss)
plt.title('The Elbow Method Graph')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
plt.show()
```



There is a bend in the elbow graph just past 5 so we will use 6 clusters.

In [436]:
```
kmeans = KMeans(n_clusters=6, init ='k-means++', max_iter=300, n_init=6,random_state=0 )
y_kmeans = kmeans.fit_predict(X)
kmeans.labels_[0:10]
```

Out[436]:
```
array([0, 0, 0, 3, 0, 0, 5, 0, 0, 0], dtype=int32)
```

DATA FRAME AND MAP OF CLUSTERS

- Next I created a new data frame with the cluster numbers inserted:
- Here is the map of the clusters
- Cluster 1(red) was selected so the past few steps were repeated for that cluster. This needed to happen because cluster 1 had 20 neighborhoods including all neighborhoods that contained wine bars.

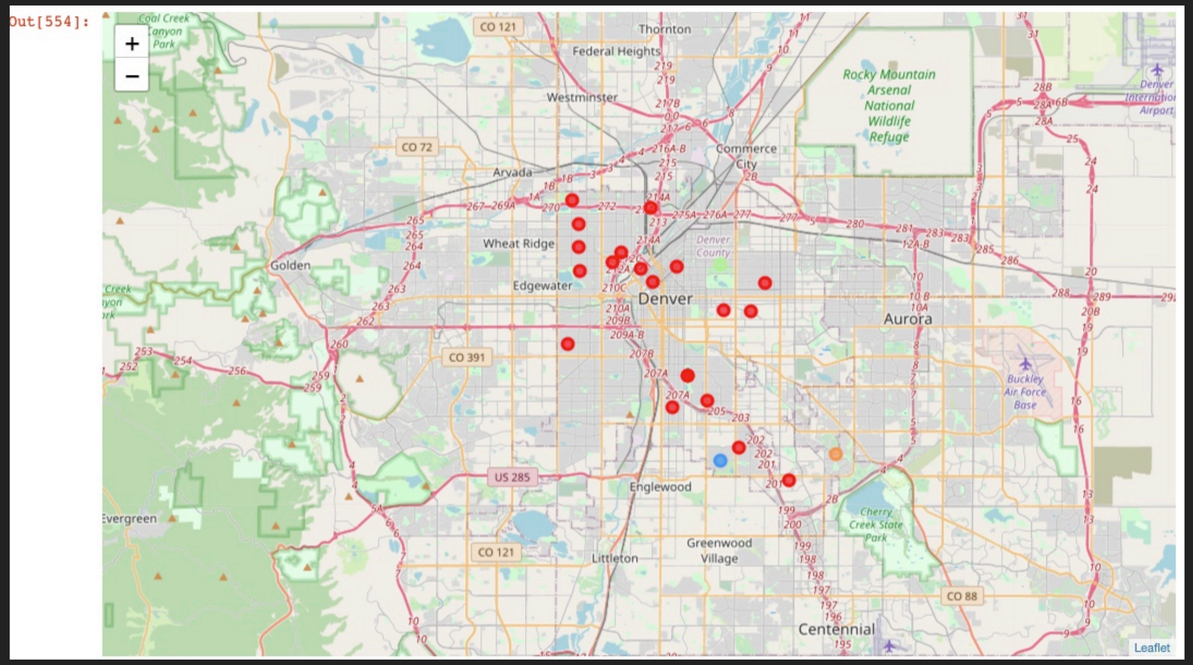
Insert the cluster labels into the combined neighborhood and top ten venue data frame.

In [494]:
```
# add clustering labels
neighborhoods_venues_sorted.insert(0, 'Cluster Labels', kmeans.labels_)
top25_denver_merged = top25_data
top25_denver_merged = top25_denver_merged.join(neighborhoods_venues_sorted.set_index('Neighborhood'), on='Neighborhood')
top25_denver_merged.head() # check the columns!
```

Out[494]:

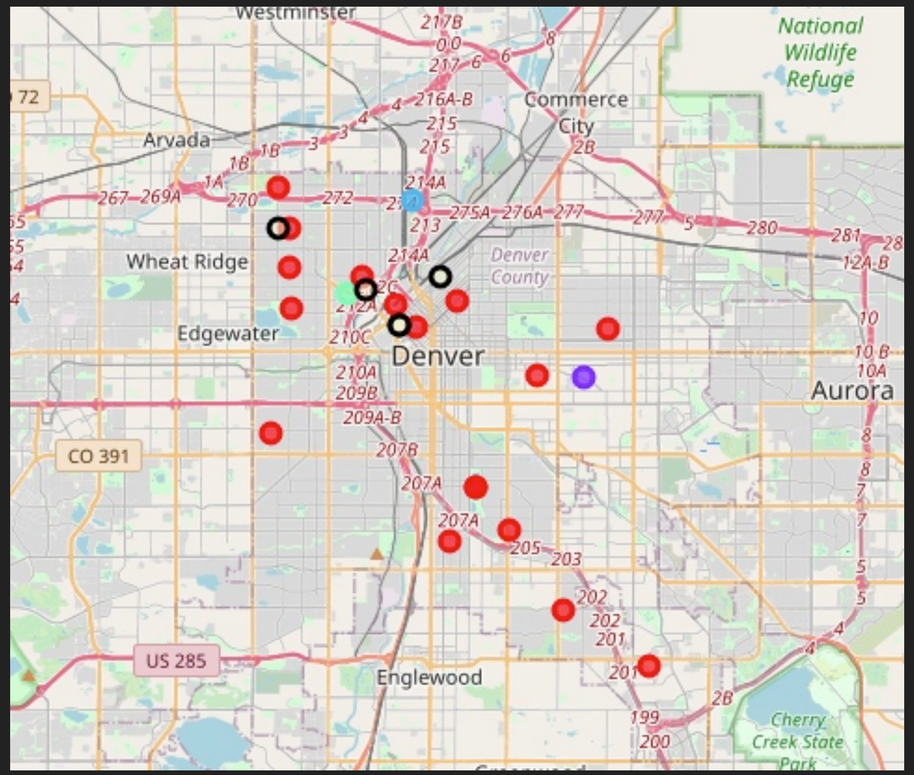
	Neighborhood_address	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue
0	South Park Hill, Denver, CO	39.746650	-104.922043	0	Coffee Shop	Bus Stop	Fast Food Restaurant	Yoga Studio	Gym / Fitness Center	American Restaurant	Marijuana Dispensary	Discount Store	Ethiopian Restaurant
1	Washington Park, Denver, CO	39.702081	-104.971034	0	Coffee Shop	Spa	American Restaurant	Park	Ice Cream Shop	Boutique	Italian Restaurant	Breakfast Spot	Electronics Store
2	Congress Park, Denver, CO	39.733720	-104.948367	0	Coffee Shop	Pizza Place	Bar	Italian Restaurant	Greek Restaurant	Grocery Store	Fast Food Restaurant	Mexican Restaurant	Chinese Restaurant
3	West Highland, Denver, CO	39.764466	-105.039271	0	Pizza Place	Coffee Shop	Mexican Restaurant	Sushi Restaurant	American Restaurant	Italian Restaurant	Fast Food Restaurant	Gym / Fitness Center	Liquor Store
4	Cherry Creek, Denver, CO	39.663610	-104.877444	3	Park	Automotive Shop	Gym	Zoo Exhibit	Event Space	Donut Shop	Electronics Store	Ethiopian Restaurant	Exhibit

Out[554]:



METHODOLOGY CONT.

- Using KMeans on the new cluster yielded five different groupings of neighborhoods.
- Map to the right shows the new clusters with an overlay of the existing wine bars



RESULTS AND DISCUSSION

- This new Wine Bar should be located in a neighborhood in the following criteria as mentioned in the data section:
- Trending neighborhood - mandatory (all of them meet this criteria)
- Near a wine shop
- In a cluster that contains a neighborhood another wine bar - mandatory
- Not in a neighborhood with many fast food restaurants or discount stores - mandatory
- In a neighborhood with other bars or breweries
- Doesn't already have a wine bar - mandatory
- Near a park
- We can already eliminate the following clusters:
- Cluster 1 - Only one neighborhood which already has a wine bar
- Cluster 2 - No wine bars in the cluster
- Cluster 4 - Only one neighborhood and has many fast food restaurants and a discount store

CHART WITH NEIGHBORHOOD SCORES

- While all of these neighborhoods would be an excellent location based on their similarity with other neighborhoods with wine bars, only one had all seven criteria. The stakeholders were fairly particular on checking all of the boxes for this location. The reasoning behind their criteria is as follows.
- They wanted a walkable neighborhood. Neighborhoods with parks are generally going to attract more foot traffic.
- They wanted to attract people willing to sit down for a meal rather than wanting to grab a quick bite at a fast food restaurant.
- They wanted to be located near other brewery or pub type restaurants to give customers a wider selection in alcohol based dining.
- And, finally, they wanted to have wine shops near by so that they could partner with them for marketing purposes.

Neighborhood	Criteria
Washington Park	5
University Hills	4
Wash Park West	5
Southmooor Park	4
Speer	6
Sloan Lake	7

CONCLUSION – SLOAN LAKE WINS!

- Sloan Lake is the winning neighborhood because it is the only one that has all seven of the criteria.

- The stakeholders are satisfied with this selection. This neighborhood has several new developments that are a combination of business on the first floor and residential space on the upper floors. They will be selecting a space in one of these locations. They will make a selection based on the availability of space for an outdoor patio and with good nearby parking.

- Here is a snapshot of the map showing the location of this neighborhood. The black and yellow circles are the existing wine bars. The cluster color for this neighborhood cluster is red.