Exploration of Denver Neighborhoods

Tracy Liu

Jun.02.2020

1. Introduction:

ABC Group (an arbitrary name) is a multinational corporate. Its business covers foods, real estate, property management and digital entertainment. And ABC group is considering expanding its business into Denver, a city with a hugely increasing number of immigrants, highly educated labors, travelers, and tech companies. The diverse communities of Denver indicate both promising opportunities and potentials for business.

However, the information at hands for management is limited and far enough to the final decision. In order to penetrate the Denver market, a preliminary analysis is requested by the management. The purpose of this analysis is to better understand the dynamics of Denver neighborhoods and reveal potential business opportunities. Tracy Liu, as an analyst in ABC Group, will be responsible for this analysis and answer the following questions:

- What is the dynamics of Denver neighborhoods
- What kind of business is recommended
- Where or which neighborhood to start the business

Information Description:

To answer the questions and generate meaningful information for the final business decision, the project analyze the information from following two perspective:

- The dynamics of Denver business market
- The crime history of Denver neighborhoods

2. Data and Source:

The dynamics of Denver business market:

- Source: Opendatasoft.com
  - Data description: Opendatasoft.com provides Denver neighborhoods name, and geo location data (latitude and longitude)
- Source: Four-square
  - Data description: Four-square provides venues and categories information in Denver neighborhoods, which is used later in clustering algorithm
- Source: Denver Government website
  - Data description: The Denver Government website provides crime data, which is used later to explore the Denver crime history.

3. **Project and Methodology:**

Project: The dynamics of Denver business market

- This project is aiming to understand the Denver neighborhoods market. Data from Opendatasoft.com and Four-square will be used together to cluster the Denver neighborhoods into similar groups. Then attributes of venues of each cluster are analyzed
- Methodology: Two machine learning algorithms
  - K-means clustering and Agglomerative Clustering (Hierarchical Clustering)

Project: The crime history of Denver neighborhoods

- This project is targeting to explore the Denver crime history of neighborhoods and generate feedbacks to support the business decision
- Methodology: Data visualization

4. **Project Data Details**

The dynamics of Denver business market

There are three parts in this project:

- Create Denver neighborhoods dataframe,
- Cluster the Denver neighborhoods
- Analysis of Denver neighborhoods cluster

Part-one: Create Denver neighborhoods dataframe

The Opendatasoft.com provides the Denver geographic dataset, it is coming from Zillow database and in geojson format. Basically it a dictionary of dictionaries. Taking a initial look of the file, we can easily find the information wanted:

- In the feature key, we wanted to extract the neighborhood name, city, county, latitudes and longitudes.

```
: Denver_data['features'][0]['properties']
```

```
: {'city': 'Denver',
   'name': 'Wellshire',
   'regionid': '268775',
   'geo_point_2d': [39.66055714600584, -104.94958065349705],
   'county': 'Denver',
   'state': 'CO'}
```

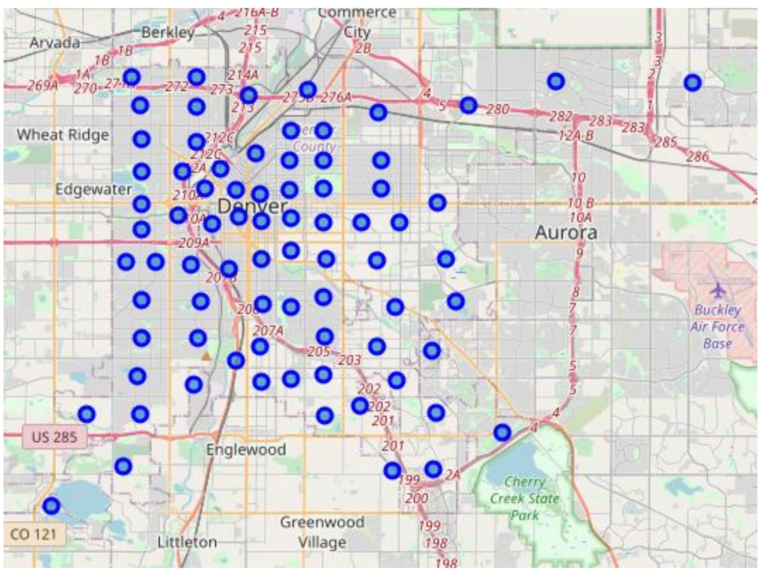
Looping through the datafile, the information above is extracted and entered into a new dataframe: neighborhoods

	City	County	Neighborhood	Latitude	Longitude
0	Denver	Denver	Wellshire	39.660557	-104.949581
1	Denver	Denver	College View - South Platte	39.672460	-105.013201
2	Denver	Denver	West Highland	39.764113	-105.039244
3	Denver	Denver	Whittier	39.756362	-104.966544
4	Denver	Denver	Ruby Hill	39.689994	-105.011162

Then exam the dataframe to check for missing value, duplicated entries and any data that does not make sense in business, following process are executed:

- 2 Hampden neighborhoods, drop the one from Arapahoe county so that all neighborhoods are from Denver county
- DIA neighborhoods is the Denver International Airport neighborhood, due to its different functionality, information of this DIA neighborhood is removed

Then using folium package, the neighborhoods are visualized:



Part-two: Cluster the Denver neighborhoods

To obtain the maximum information of Denver neighborhoods venues, let's do a simple estimation:

According to Wikipedia, A total of 78 counties occupy 155 square miles (401 km2) of area, using math then we can get the average radius of each county: 2km or 2000 meters. Denver population is about 0.7 million, making up about 13% of total population of Colorado State. The number of small business in Colorado in 2018 is about 610,000, using the population percentage we can approximate the number of small business in Denver, which is 610,000 * 0.13 = 79,300, then each neighborhood has about 1,016 small business.

- According to the information and approximation above, set up the radius as 2000 (meters) and limit as 1000
- After trying multiple times, I soon find out that the max venue output is 100, so limit here can be any number larger than 100

Then, through the Four-square API, Denver neighborhoods and their venue information is downloaded and input into a new dataframe:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wellshire	39.660557	-104.949581	Chipotle Mexican Grill	39.666417	-104.939887	Mexican Restaurant
1	Wellshire	39.660557	-104.949581	Patxi's Chicago Pizza	39.654451	-104.959771	Pizza Place
2	Wellshire	39.660557	-104.949581	Glacier Ice Cream	39.654489	-104.959988	Ice Cream Shop
3	Wellshire	39.660557	-104.949581	Sprouts Farmers Market	39.664247	-104.939432	Grocery Store
4	Wellshire	39.660557	-104.949581	Schlessman Family YMCA	39.668171	-104.941526	Recreation Center

What we care about is the Venue category, it represents the business category of each venue. Then a quick summary shows that more than half of 77 neighborhoods have 100 categories:

	Venue Category
count	77.000000
mean	88.025974
std	18.641334
min	38.000000
25%	74.000000
50%	100.000000
75%	100.000000
max	100.000000

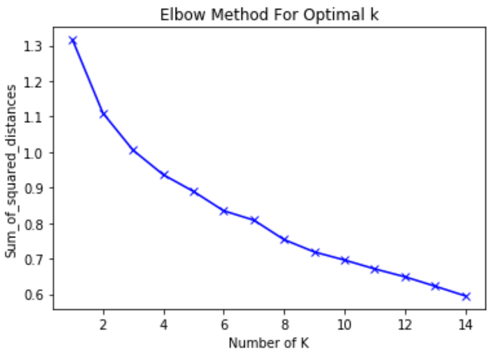
Now all the data needed is acquired, the next step is to cluster the neighborhoods using the venue categories. Before running the algorithm, several data manipulation process are executed to prepare the data for use:

- Use one-hot coding to convert categorical variable (venue) into binary variable
- Group by the neighborhood using the average frequency of each venue category
    - The average frequency is used by the machine learning algorithm to determine the similarities among neighborhoods
- Find out the top 30 venues and for each neighbor and convert into data frame for use later

Clustering the neighborhoods:

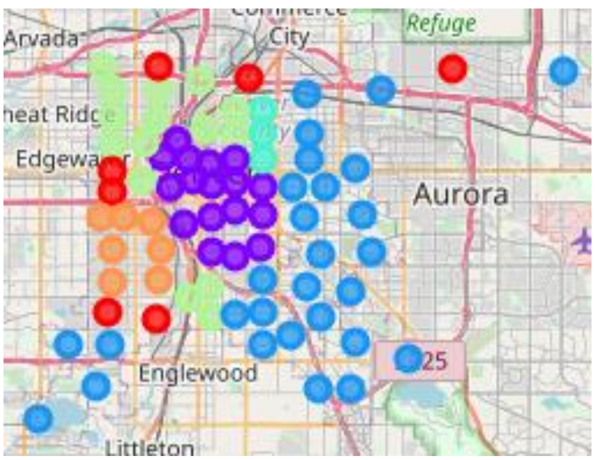
Both the K-means and Agglomerative algorithm requires a k (number of clusters) decides in advance. To determine the K, two methods are used:

- Empirical Method: A simple empirical method of finding number of clusters is Square root of N/2 where N is total number of data points, so that each cluster contains square root of 2 * N Number of Cluster: (77/2)^(0.5) ~ 6
- Elbow Method: Find the point where Sum of Squared distance change steeply
    - Elbow method also indicates the K = 6

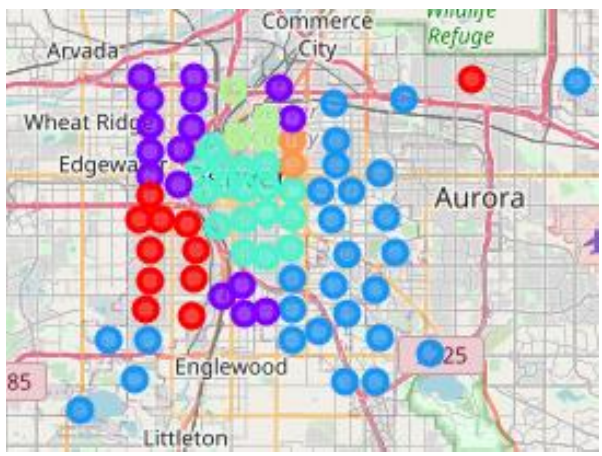


Clustering results:

K-means



Agglomerative



K-means Cluster	Neighborhoods	Agg Cluster - ward linkage	Neighborhoods
2	28	2	27
3	17	3	17
1	15	1	16
0	7	0	11
4	7	4	4
5	3	5	2

At first, the algorithms do not differentiate each other too much. Visual comparison shows that Agglomerative is slightly better than K-means since it cluster the neighborhoods better in circle. From the results of K-means, there are some neighborhoods (red dots in the left graph) are spreading far from others within the same cluster.

Besides the visual comparison, statistic output is also necessary to evaluate the two algorithms, the following three measurements are used:

- Silhouette Coefficient: a higher Silhouette Coefficient score relates to a model with better defined clusters
- Calinski-Harabasz Index: a higher Calinski-Harabasz score relates to a model with better defined clusters
- Davies-Bouldin Index: a lower Davies-Bouldin index relates to a model with better separation between the clusters

Results from statistic comparison:

```
The K-means silhouette_score: 0.121416031756513
The Agglom silhouette_score: 0.1306063199035374

The K-means Calinski-Harabasz Index: 8.177786576441273
The Agglom Calinski-Harabasz Index: 8.232882841289406


The K-means Davies-Bouldin Index: 1.9052393299627661
The Agglom Davies-Bouldin Index: 1.813738882011263
```

From the three measurements above:

1. Agglomerative is better than K-means with higher Silhouette Coefficient
2. Agglomerative is better than K-means with higher Calinski-Harabasz Index
3. Agglomerative is better than K-means with lower Davies-Bouldin Index

Overall Agglomerative model is better than K-means and results from Agglomerative model is used for cluster analysis.

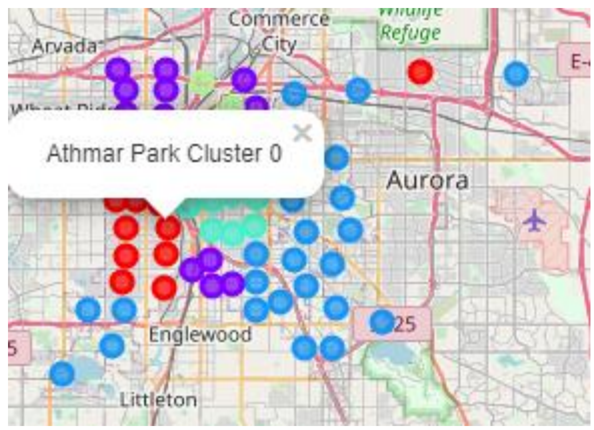
	Agg Cluster	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	0	Mexican Restaurant	Convenience Store	Vietnamese Restaurant	Fast Food Restaurant	Grocery Store	Discount Store	Marijuana Dispensary	Pizza Place	Gas Station	Sandwich Place
1	1	Mexican Restaurant	Coffee Shop	Brewery	Pizza Place	Park	Bar	Breakfast Spot	Convenience Store	Italian Restaurant	Marijuana Dispensary
2	2	Coffee Shop	Sandwich Place	Pizza Place	Mexican Restaurant	Park	Fast Food Restaurant	Convenience Store	Liquor Store	Grocery Store	Discount Store
3	3	American Restaurant	Coffee Shop	Mexican Restaurant	Italian Restaurant	Brewery	Sandwich Place	Pizza Place	Bar	Park	Hotel
4	4	Brewery	Bar	Coffee Shop	Pizza Place	Burger Joint	Cocktail Bar	New American Restaurant	Restaurant	Convenience Store	Park
5	5	Zoo Exhibit	Bar	Science Museum	Coffee Shop	Park	Greek Restaurant	Brewery	Pizza Place	Mexican Restaurant	American Restaurant

Part-three: Analysis of Denver neighborhoods cluster

Adding the labels to the neighborhoods and then summarized the frequency of venue categories by each cluster, taking the top 10 venue categories:

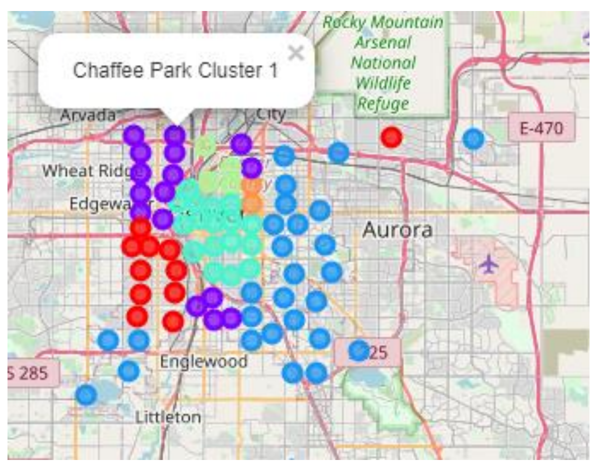
Analysis of each cluster:

Cluster 0: 11 Neighborhoods



- Cluster 0 contains neighborhoods in the west area of Denver
- Mexican Restaurant, Convenience Store, Vietnamese Restaurant, Fast Food Restaurant, Grocery Store, Discount Store, Pizza Place, Gast Station: Choice of cheap food, small amount purchase and maybe poor living environment
- Marijuana Dispensary: Indicator of crimes

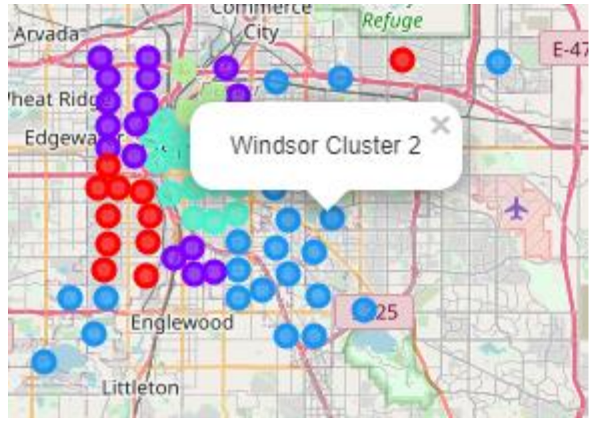
Cluster 1: 16 Neighborhoods



- Cluster 1 contains neighborhoods mostly in the west and northwest area, a few in the north and south area of Denver
- Coffee, brewery, park, bar: Places to hang out and chill.
- Mexican Restaurants, Pizza Places, and Convenience Store: A low living expenses.
- Marijuana Dispensary: Indicator of crimes
- Italian Restaurants: Premium food consumptions

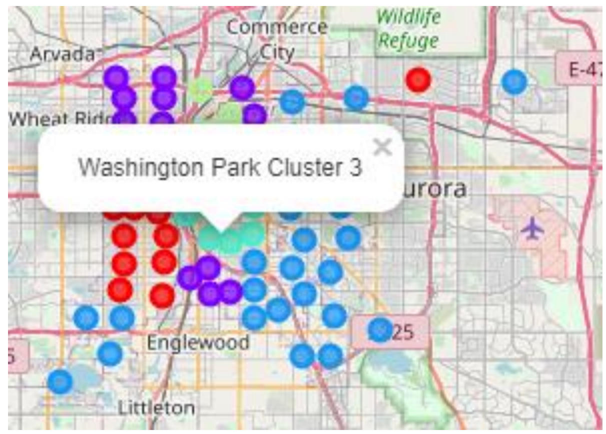
Cluster 2: 27 Neighborhoods

- Cluster 2 contains neighborhoods from northeast to east and south area of Denver
- Coffee shop, park: Place to hang out and chill
- Sandwich Place, Pizza Place, Mexican Restaurant, Fast Food Restaurant: Choice for cheap food
- Convenience Store, Liquor Store, Grocery Store, Discount Store: Small amount purchase



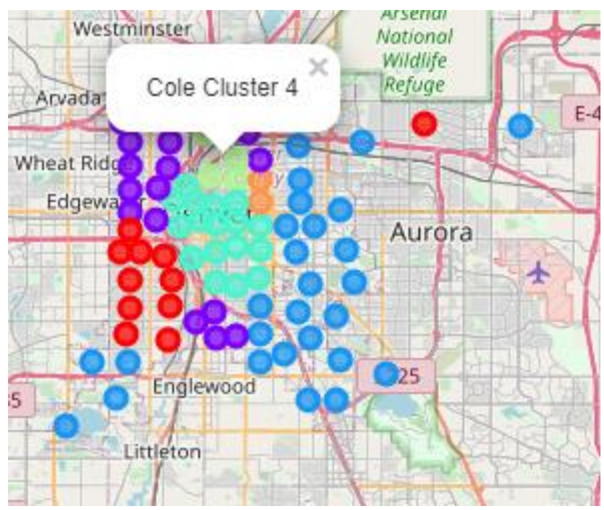
Cluster 3: 17 Neighborhoods

- Cluster 3 contains neighborhoods in downtown (center) area of Denver
- American Restaurant, Italian Restaurant: Premium food consumption
- Coffee Shop, Brewery, Park, Bar: Place to handout and chill
- Mexican Restaurant, Sandwich Place, Pizza Place: Choice for cheap food
- Hotel: Travelers' choice



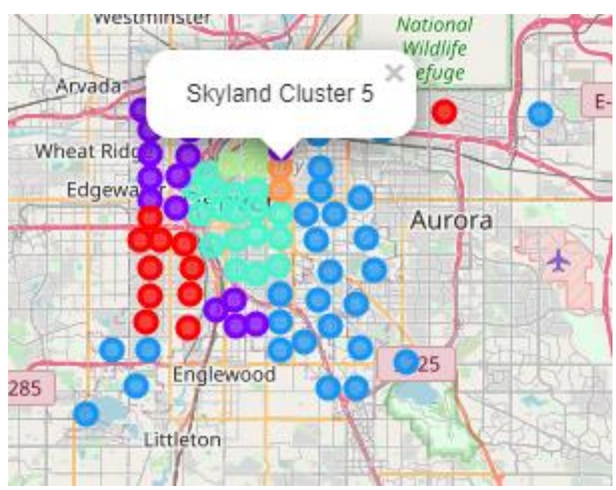
Cluster 4: 4 Neighborhoods

- Cluster 4 contains neighborhoods north of downtown area
- Brewery, Bar, Coffee Shop, Cocktail Bar, Park: Place to hang out and chill
- Burger Joint, New American Restaurant: Explore the different taste
- Convenience Store, Pizza Place: Choice for cheap food and small amount purchase



Cluster 5: 2 Neighborhoods

- Cluster 5 contains neighborhoods north of downtown area, next to cluster 4
- Zoo Exhibit, Science Museum: Place for Learning and family activities
- Bar, Coffee Shop, Park, Brewery: Place to hang out and chill



- Greek Restaurant, American Restaurant: Diverse food consumption
- Pizza Place, Mexican Restaurant: Choice for cheap food

The crime history of Denver neighborhoods

There are two parts of the project:

- Create a dataframe of Denver crimes history
- Analysis of Denver neighborhoods clusters crime history

Part-one: Create a dataframe of Denver crimes history

There are two data files to work on, the first one is the crime data file, the second one is the crime code data file. The part one cleans the both file, merges then into one dataframe for the analysis for part two.

Data description and clean: Crime data file

INCIDENT_ID 486830 non-null int64

- Unique identifier of each incident
- Clean procedure: keep

OFFENSE_ID 486830 non-null int64

- Incident id + offense code + 0 + offense code extension
- Clean procedure: drop

OFFENSE_CODE 486830 non-null int64

- The offense category code
- Clean procedure: keep

OFFENSE_CODE_EXTENSION 486830 non-null int64

- The extension of offense category, 1 or 0
- Clean procedure: keep and combine with offense code to form as a unique key

OFFENSE_TYPE_ID 486830 non-null object

- Specific offense type name
- Clean procedure: drop

OFFENSE_CATEGORY_ID 486830 non-null object

- Offense category name
- Clean procedure: drop

FIRST_OCCURRENCE_DATE 486830 non-null object

- The first time the incident occurred
- Clean procedure: drop

LAST_OCCURRENCE_DATE 155228 non-null object

- The last time the incident occurred, missing value identified
- Clean procedure: drop

REPORTED_DATE 486830 non-null object

- The date time the incident is reported
- Clean procedure: convert to datetime and keep

INCIDENT_ADDRESS 441162 non-null object

- The address the incident happened
- Clean procedure: keep

GEO_X 482610 non-null float64

- Geographic coordinate
- Clean procedure: drop

GEO_Y 482610 non-null float64

- Geographic coordinate
- Clean procedure: drop

GEO_LON 482610 non-null float64

- Longitude
- Clean procedure: keep

GEO_LAT 482610 non-null float64

- Latitude
- Clean procedure: keep

DISTRICT_ID 486830 non-null int64

- District ID
- Clean procedure: drop

PRECINCT_ID 486830 non-null int64

- Precinct ID
- Clean procedure: drop

NEIGHBORHOOD_ID 486830 non-null object

- Neighborhood name in lower case and hyphen between words
- Clean procedure: replace the hyphen with space, title the first letter, add the label to the n eighborhood

IS_CRIME 486830 non-null int64

- Is a crime or not, 1 or 0
- Clean procedure: filter the row with value = 1

IS_TRAFFIC 486830 non-null int64

- Is a traffic violation or not, 1 or 0
- Clean procedure: drop

Data description and clean: Crime code file

OFFENSE_CODE 299 non-null int64

- The offense category code
- Clean procedure: keep

OFFENSE_CODE_EXTENSION 299 non-null int64

- The extension of offense category, 1 or 0
- Clean procedure: keep and combine with offense code to form as a unique key

OFFENSE_TYPE_ID 299 non-null object

- Specific offense type name
- Clean procedure: drop

OFFENSE_TYPE_NAME 299 non-null object

- Specific offense type name

- Clean procedure: drop

OFFENSE_CATEGORY_ID 299 non-null object

- Specific offense category name
- Clean procedure: drop

OFFENSE_CATEGORY_NAME 299 non-null object

- Specific offense category name
- Clean procedure: keep

IS_CRIME 299 non-null int64

- Is a crime or not, 1 or 0
- Clean procedure: drop

IS_TRAFFIC 299 non-null int64

- Is a traffic violation or not, 1 or 0
- Clean procedure: drop

Use the unique key in each dataframe and merge the two dataframe, then drop the unnecessary columns:

	Neighborhood	REPORTED_DATE	Agg Cluster Labels	OFFENSE_TYPE_NAME	OFFENSE_CATEGORY_NAME
0	Baker	2018-07-23 03:42:00	3.0	Homicide by a family member	Murder
1	Bear Valley	2016-07-31 15:26:00	2.0	Homicide by a family member	Murder
2	Chaffee Park	2020-02-09 21:41:00	1.0	Homicide by a family member	Murder
3	City Park West	2019-05-31 17:33:00	3.0	Homicide by a family member	Murder
4	East Colfax	2017-03-18 04:12:00	2.0	Homicide by a family member	Murder

Part – two: Analysis of Denver neighborhoods clusters crime history

Count of Crimes by Cluster:

	Neighborhood Label	Count of Crimes	Percentage
3	3.0	104,171	29%
2	2.0	100,918	28%
0	0.0	61,198	17%
1	1.0	54,656	15%
4	4.0	32,099	9%
5	5.0	3,202	1%

- Cluster 2 and cluster 3 each contributes nearly 30% of total crimes
- Considering the number of neighborhoods, cluster 1 and cluster 3 have almost the same number of neighborhoods, but the number of crimes from cluster 3 is nearly twice the number from cluster 1
- Cluster 1 and cluster 0 are very close

Overall count of crimes:

	OFFENSE_CATEGORY_NAME	Count_of_crimes	Percentage
1	All Other Crimes	85,889	24%
9	Public Disorder	48,811	14%
6	Larceny	48,126	14%
12	Theft from Motor Vehicle	38,977	11%
5	Drug & Alcohol	30,337	9%
3	Auto Theft	26,733	8%
4	Burglary	23,799	7%
8	Other Crimes Against Persons	23,764	7%
0	Aggravated Assault	12,102	3%
13	White Collar Crime	6,336	2%
10	Robbery	6,293	2%
11	Sexual Assault	4,167	1%
2	Arson	588	0%
7	Murder	322	0%

- Almost 24% of crimes belong to "All other Crimes"
- Excluding the "All other crimes":
  - Stealing related (Larceny + Theft) counted for 33% of crimes
  - Public disorder comes as the biggest single category as 14%
  - Drug & Alcohol comes as the second biggest single category as 9%
  -

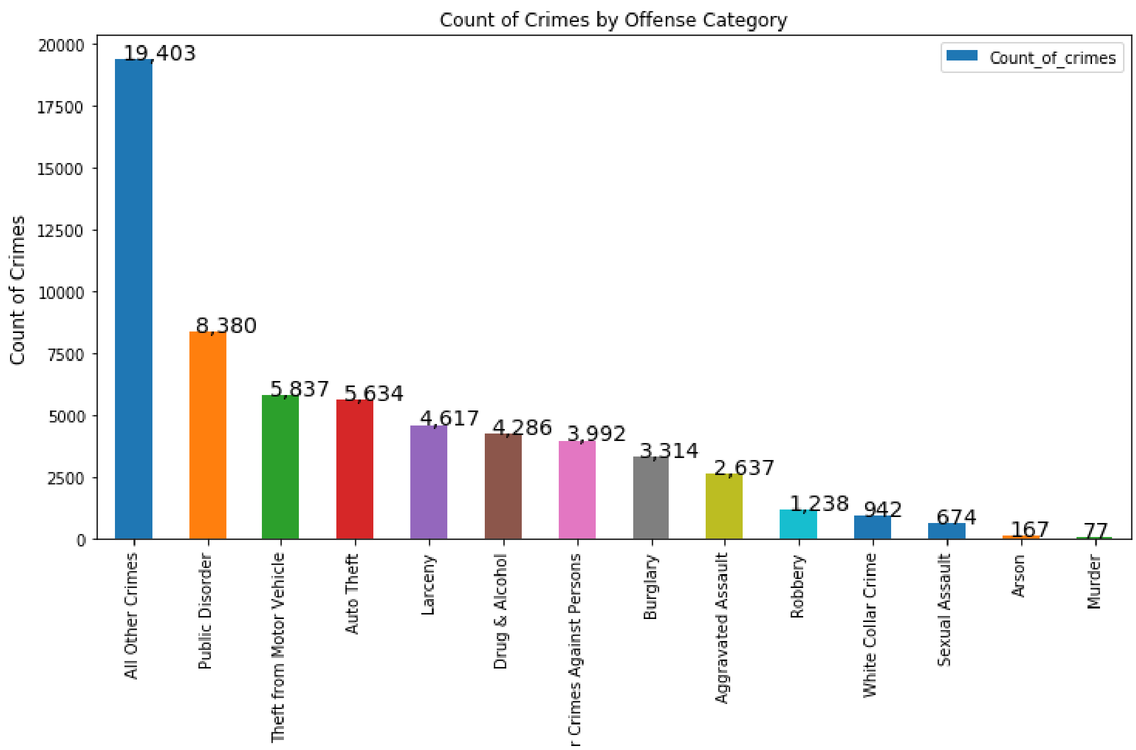
Crimes category distribution among clusters

OFFENSE_CATEGORY_NAME	Aggravated Assault	All Other Crimes	Arson	Auto Theft	Burglary	Drug & Alcohol	Larceny	Murder	Other Crimes Against Persons	Public Disorder	Robbery	Sexual Assault	Theft from Motor Vehicle	White Collar Crime
Agg Cluster Labels														
0.0	22%	23%	28%	21%	14%	14%	10%	24%	17%	17%	20%	16%	15%	15%
1.0	14%	15%	18%	18%	21%	12%	13%	14%	13%	16%	14%	14%	17%	14%
2.0	26%	22%	25%	35%	35%	20%	31%	32%	27%	29%	28%	28%	36%	37%
3.0	26%	30%	19%	18%	23%	39%	38%	17%	32%	28%	28%	30%	23%	28%
4.0	12%	9%	8%	7%	7%	15%	8%	11%	10%	9%	9%	10%	8%	6%
5.0	1%	1%	2%	1%	1%	1%	1%	2%	1%	1%	1%	1%	1%	1%
Category_Total	12,102	85,889	588	26,733	23,799	30,337	48,126	322	23,764	48,811	6,293	4,167	38,977	6,336

- Cluster 2 and 3 are highest almost among all the categories
- Cluster 2 takes 35% of auto theft and burglary, 36% of theft from motor vehicle, and 37% of white collar crime
- Cluster 3 takes 39% in Drug & Alcohol, 38% in Larceny

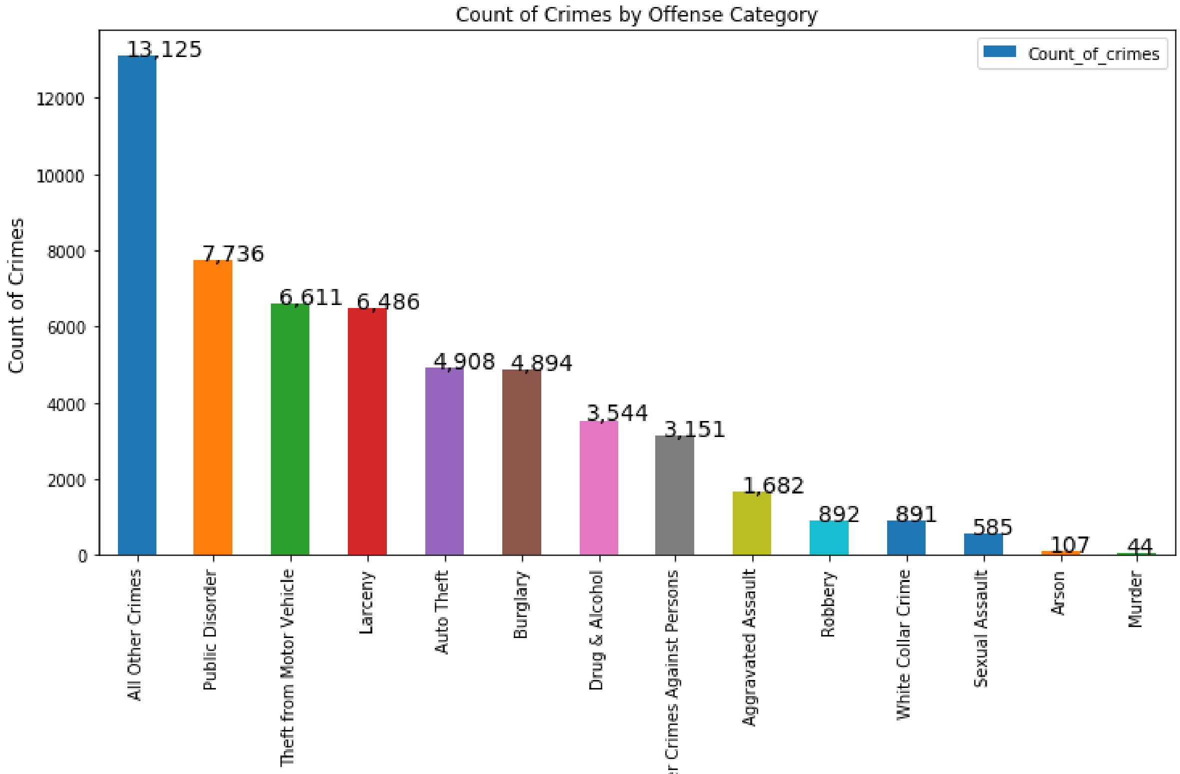
Analysis of Clusters:

Cluster 0:



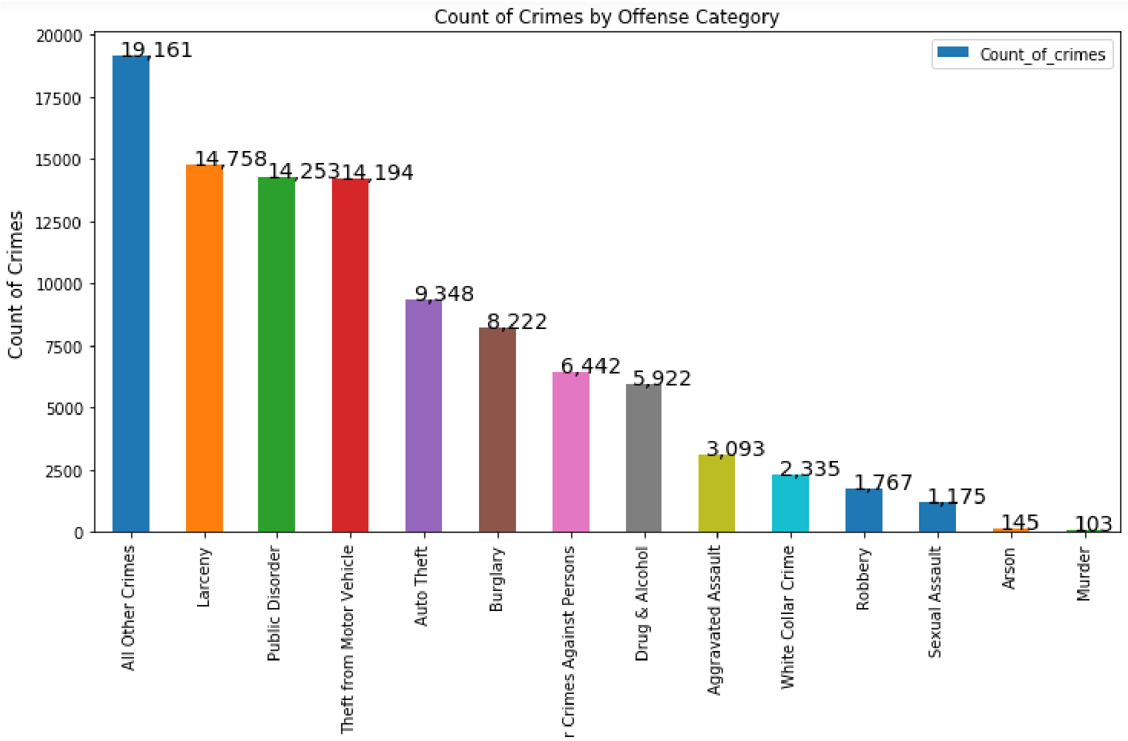
- Excluding the 'All other Crimes', the primary category of crimes are public disorder, steal related (theft, larceny, burglary) and drug & alcohol
- The amount of Robbery and murder is relatively small
- There is no change of trend in terms of specific category

Cluster 1



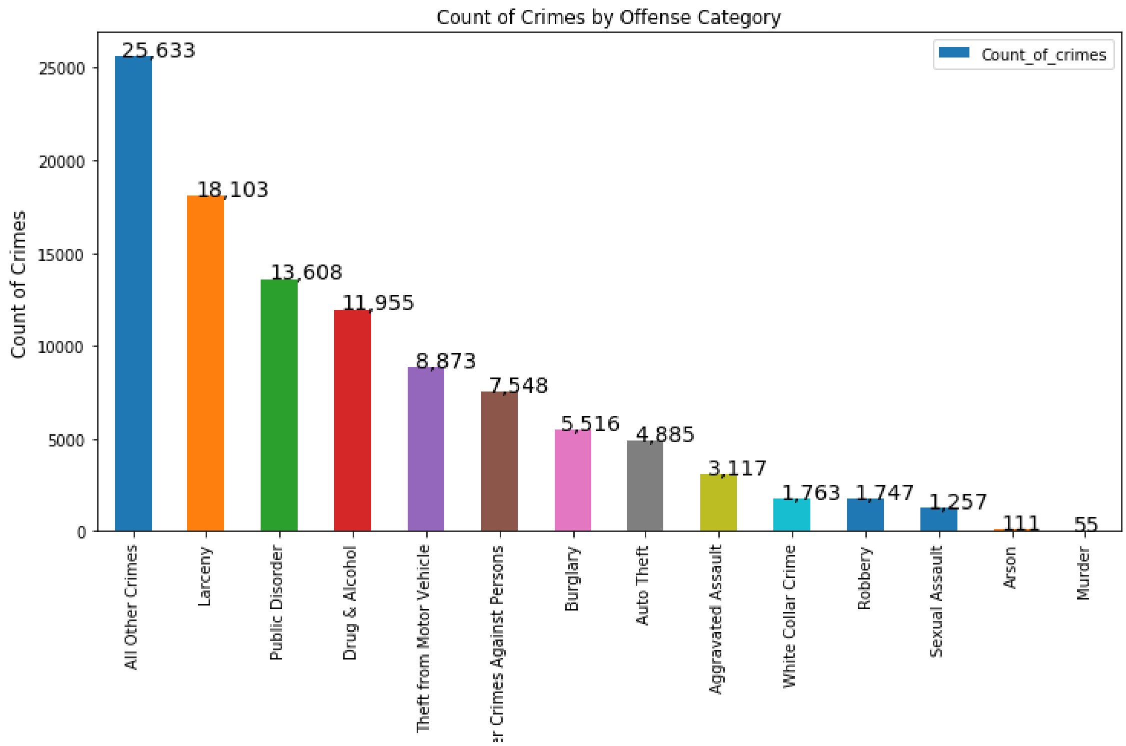
- Unlike Cluster 0, Cluster 1 has less "All other Crimes" (13,125 vs 19,403)
- Public Disorder, steal related (theft, larceny, burglary) and drug & alcohol are the primary crimes
- Larceny and Theft from Motor Vehicle increase continuously

Cluster 2



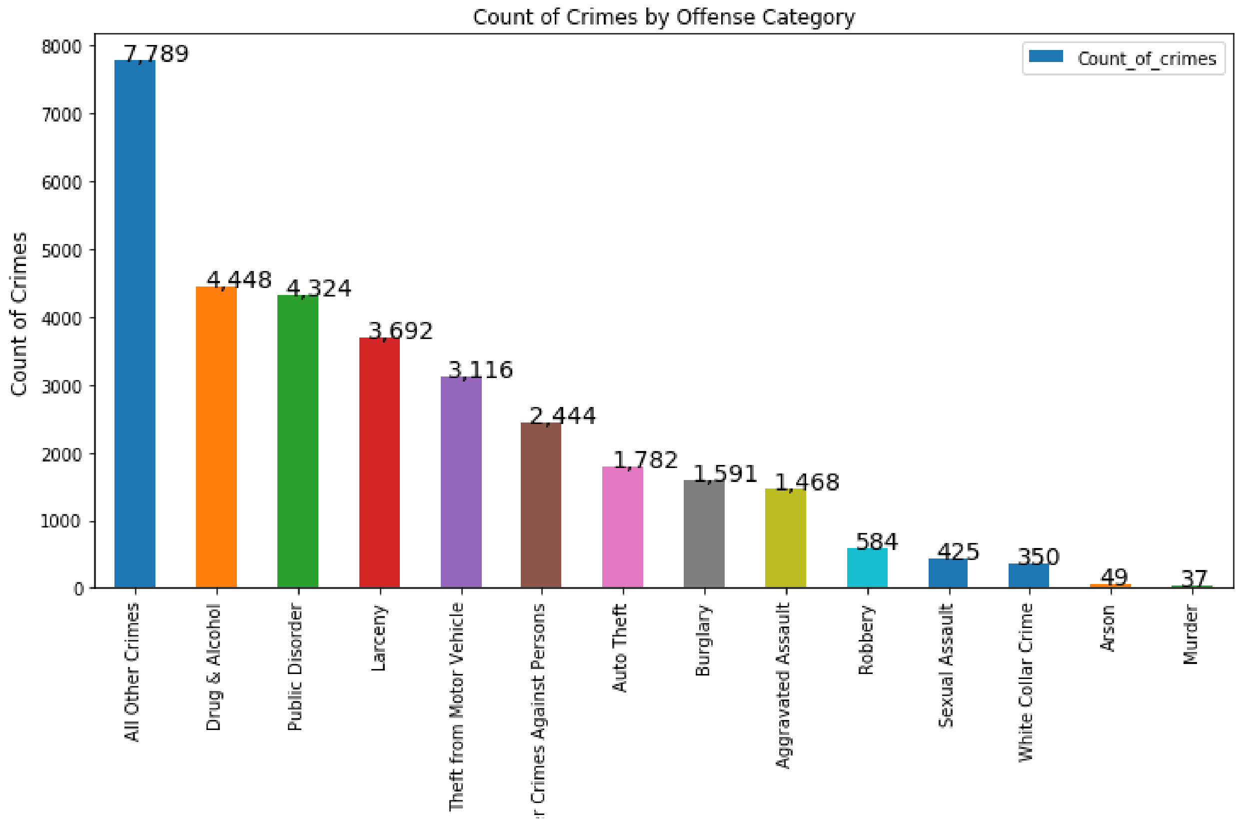
- In cluster 2, Larceny, Public Disorder, and Theft from Motor Vehicle are very high
- Larceny and Theft from Motor Vehicle increase continuously over the years

Cluster 3



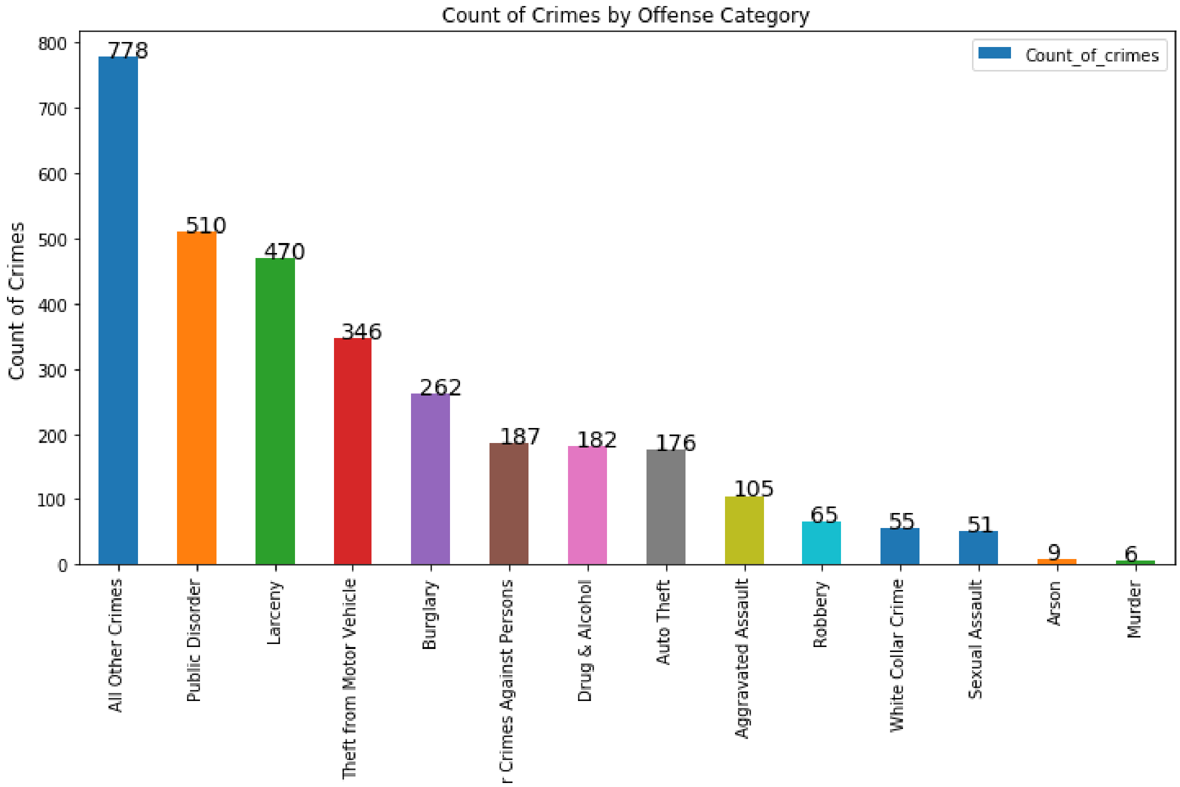
- Larceny, Public Disorder and Drug & Alcohol are very high
- Larceny and Theft from Motor Vehicle increase continuously over the years
- Drug & Alcohol decrease over the years

Cluster 4



- Cluster 4 has much less crimes, the highest one is less than 10,000
- Drug & Alcohol, Public Disorder and Larceny are the top three
- Drug & Alcohol decreases over the years

Cluster 5



- Cluster 5 is the best cluster, the biggest one is still small than 1000
- Top three are Public Disorder, Larceny, and Theft from Motor Vehicle
- Larceny and Public Disorder increase over the years

Conclusion

Based on the average frequency of venue category in every neighborhood, Denver neighborhoods are clustered into 6 clusters, K-means clustering and Agglomerative clustering are used, comparing the results from visual comparison and statistic measurements, result from Agglomerative clustering is better.

Cluster 0:

- Recommended for small business, such as fast food, convenience store, and pizza store
- Major crime: Public Disorder
- Among all clusters, crimes in cluster 0 is moderate, no trend up of any major crime type.

Cluster 1:

- Recommended for small and medium business, Italian restaurant, bar, and coffee shop are popular.
- Major crime: Public Disorder, Theft from Motor Vehicle, Larceny
- Among all clusters, crime in cluster 1 is moderate, but Larceny and Theft from Motor Vehicle are increasing

Cluster 2:

- Recommended for small business such as pizza shop, Mexico restaurants, and convenience store
- Major crime: Public Disorder, Theft from Motor Vehicle, Larceny
- Among all clusters, crime in cluster 2 is high, it has the largest number of crimes in terms of Public Disorder and Theft from Motor Vehicle
- Larceny and Theft from Motor Vehicle increase continuously over the years

Cluster 3:

- Recommended for medium and premium business
- Major crime: Larceny, Public Disorder, Drug & Alcohol
- Among all clusters, crime in cluster 3 is the highest one. It has the largest number of crimes in terms of drug & alcohol, larceny, and sexual assault
- Larceny and Theft from Motor Vehicle increase continuously over the years

Cluster 4

- Recommended for entertainment and premium business
- Major crime: Larceny, Public Disorder, Drug & Alcohol
- Among all clusters, crime in cluster 4 is small

Cluster 5:

- Recommended for public education and entertainment business
- Major crime: Public Disorder, Theft from Motor Vehicle, Larceny
- Among all clusters, crime in cluster 5 is extremely low