Coursera Capstone Project:

Neighborhood opportunity for a Security Service in Denver, Colorado

Fabian A Williams

November 2019

Table of Contents

Acknowledgements

The success and final outcome of this project required a lot of documentation review and practice which was provided by IBM Data Science learning and development team using the Coursera learning platform. I am extremely grateful to have gotten this support all throughout the IBM Profession Data Science Certificate course.

To my online learning colleagues I say thank you for reviewing and commenting on my course submissions.

Part 1: Business problem and data description

Introduction

The intent of this project is to provide investors information to support their decision to establish a security service in Denver, Colorado (Denver, CO). This document was prepared based on several months of research, data extraction and analysis using Python. The information used in this report is real and was accessed from the Denver, CO county database and clusters using the FourSquare API.

About Denver, Colorado

Denver, is the capital of the U.S. state of Colorado. Denver is located in the South Platte River Valley on the western edge of the High Plains just east of the Front Range of the Rocky Mountains. The Denver downtown district is immediately east of the confluence of Cherry Creek with the South Platte River, approximately 12 mi (19 km) east of the foothills of the Rocky Mountains. Denver is named after James W. Denver, a governor of the Kansas Territory. It is nicknamed the Mile High City because its official elevation is exactly one mile (5280 feet or 1609.3 meters) above sea level.

Denver is ranked as a Beta world city by the Globalization and World Cities Research Network. Estimated population of 716,492 in 2018, Denver is the 19th-most populous U.S. city, and with a 19.38% increase since the 2010 United States Census, it has been one of the fastest-growing major cities in the United States.

- The 10-county Denver-Aurora-Lakewood, CO Metropolitan Statistical Area had an estimated 2018 population of 2,932,415 and is the 19th most populous U.S. metropolitan statistical area.
- The 12-city Denver-Aurora, CO Combined Statistical Area had an estimated 2018 population of 3,572,798 and is the 15th most populous U.S. metropolitan area.
- In 2016, Denver was named the best place to live in the United States by U.S. News & World Report.

The Business Problem¶

Though rated the best place to live, can Denver, CO considered a safe place to operate a business. Investors in a security firm would like to perform a study of the neighborhood crimes within Denver, CO. The company's intent is to identify neighborhoods with burglary crime, across multiple business locations/venue categories. The segmentation of these businesses will also form apart of the project planning process, as all neighborhoods will not be launched at the same time. Other uses of the project data:

- Families or individuals relocating to Denver, CO
- Real Estate companies seeking key selling features of the city, as the venues are clustered
- City Offices/Governor looking to gentrify a high crime area.

Success Criteria: Use the data to recommend the first neighborhood to be launched and there after the order of subsequent neighborhoods.¶

Data Processing & Description¶

This project will require the analysis of Denver, CO neighborhood crime statistics, specifically focusing on burglary crimes and the clustering of the neighborhoods in order to recommend the first neighborhood. The databases accessed for this project was from www.denvergov.org. This database provided the Denver, CO crime data: https://www.denvergov.org/media/gis/DataCatalog/crime/csv/crime.csv.

- Author: City and County of Denver, Denver Police Department / Data Analysis Unit
- Maintainer: City and County of Denver, Technology Services / DenverGIS Data
- Version: 1.0.2921
- Data was last updated on 11/21/2019

This dataset includes criminal offenses in the City and County of Denver for the previous five calendar years plus the current year to date. The data is based on the National Incident Based Reporting System (NIBRS) which includes all victims of person crimes and all crimes within an incident.

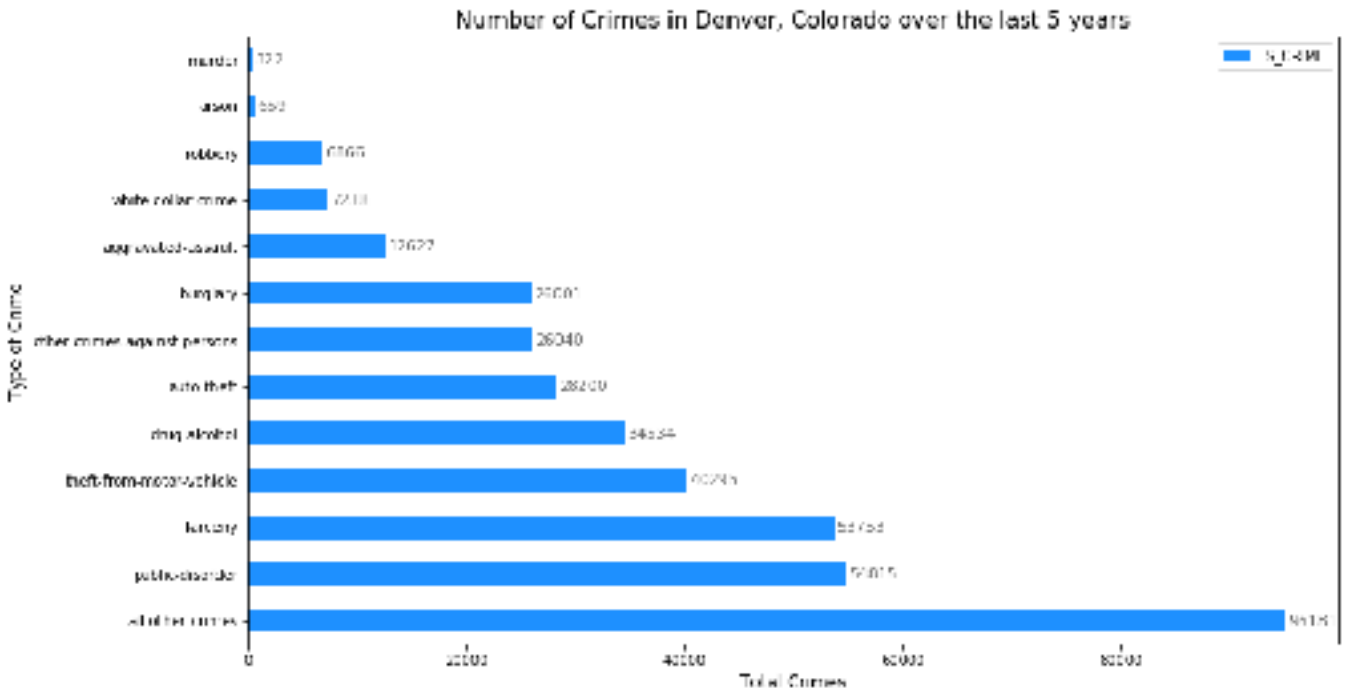

FIGURE 1: CRIMES IN DENVER, COLORADO

The data is dynamic, which allows for additions, deletions and/or modifications at any time, resulting in more accurate information in the database. Due to continuous data entry, the number of records in subsequent extractions are subject to change. Crime data is updated Monday through Friday. Crimes that occurred at least 30 days ago tend to be the most accurate, although records are returned for incidents that happened yesterday.

The table consists of the columns listed in Figure 2. The data will be wrangled to only indicate neighborhoods where only burglary crimes have been committed.

```
INCIDENT_ID                int64
OFFENSE_ID                 int64
OFFENSE_CODE               int64
OFFENSE_CODE_EXTENSION     int64
OFFENSE_TYPE_ID           object
OFFENSE_CATEGORY_ID       object
FIRST_OCCURRENCE_DATE     object
LAST_OCCURRENCE_DATE      object
REPORTED_DATE             object
INCIDENT_ADDRESS          object
GEO_X                    float64
GEO_Y                    float64
GEO_LON                  float64
GEO_LAT                  float64
DISTRICT_ID                int64
PRECINCT_ID                int64
NEIGHBORHOOD_ID           object
IS_CRIME                   int64
IS_TRAFFIC                 int64
dtype: object
```

FIGURE 2: DATA TABLE COLUMNS

For the purpose of this report the relevant columns are:

•OFFENCE_CATEGORY_ID: the type of crime

•GEO_LON: longitude

•GEO_LAT: latitude

•NEIGHBORHOOD_ID: name of the neighborhood

•IS_CRIME: crime committed using a boolean values 1 or 0

The data in these columns will be wrangled to provide an initial view of the burglary data which will then be used with the FourSquare API to explore the world around where the crimes have been committed. The Foursquare API allows application developers to interact with the Foursquare platform. The API itself is a RESTful set of addresses to which you can send requests, the request returns XML or JSON format.

This dynamic nature of crime data means that content provided here today will probably differ from content provided a week from now. Likewise, content provided on this site will probably differ somewhat from crime statistics published elsewhere by the City and County of Denver, even though they draw from the same database. Crime locations reflect the approximate locations of crimes but are

not mapped to actual property parcels. Certain crimes may not appear on maps if there is insufficient detail to establish a specific, mappable location.

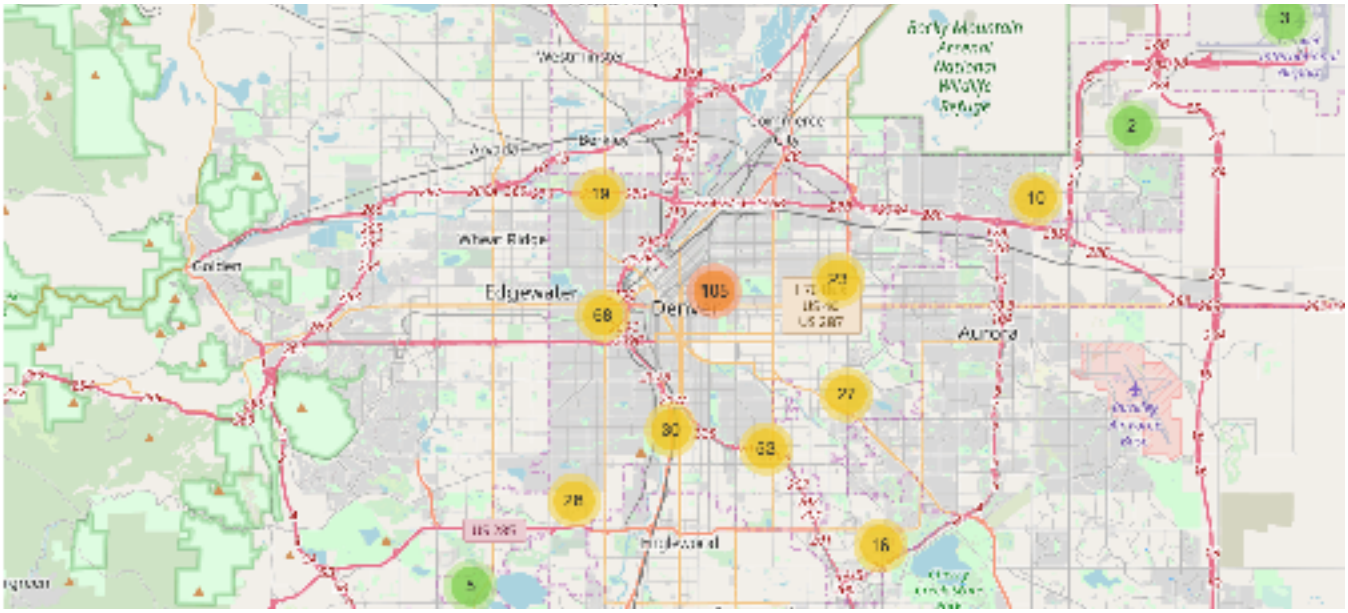

FIGURE 3: DENVER, COLORADO CRIME MAP

After wrangling the data to extract only the burglary crime within each neighborhood we can look at the dispersion of crimes across the Denver, CO geography.

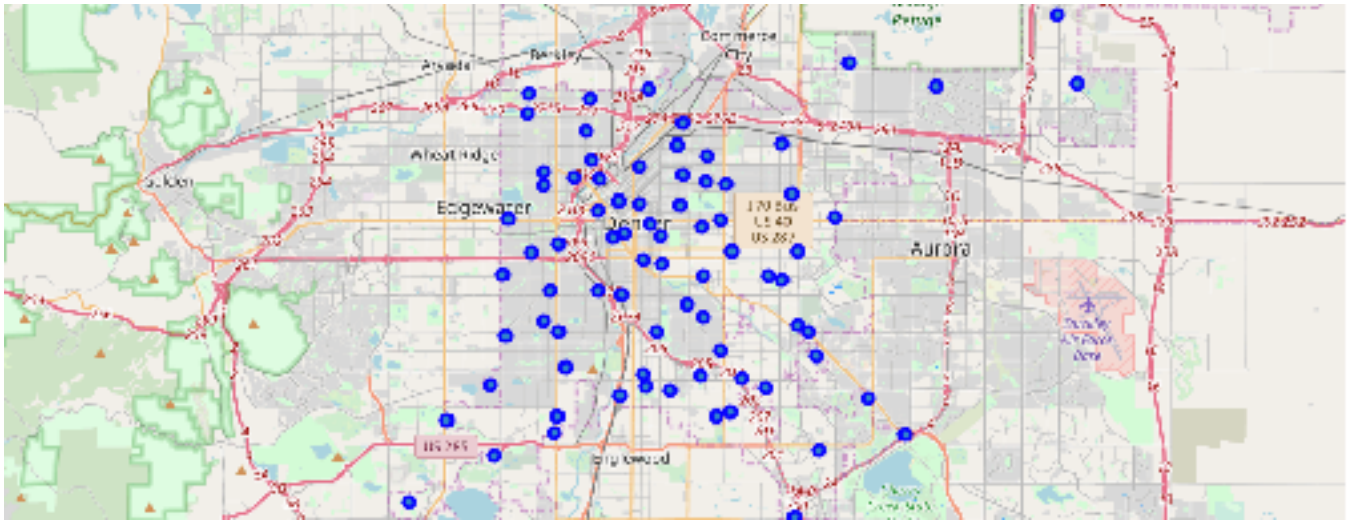

FIGURE 4: BURGLARY CRIME BY NEIGHBORHOOD

Part 2: Connect to the FourSquare API and Data Clustering

Methodology

This analysis will be done using Python. Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Python has built in libraries that will be installed in a Python Notebook to perform the analysis. The primary libraries are as follows:

- Pandas data frame: the data will be aligned in a tabular fashion in rows and columns.
- Numpy is a general-purpose array-processing package
- Matplotlib: is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms.
- Folium Python: helps to create several types of Leaflet maps.
- Scikit-learn provides many unsupervised and supervised learning algorithms such as K Means clustering. Given a data set of items, with certain features, and values for these features (like a vector). The kMeans algorithm; an unsupervised learning algorithm, is used to categorize these items.

The database is imported using Pandas, and an initial view of data was created using Matplotlib, a horizontal bar chart (see Fig. 1). The data was wrangled to identify burglary crimes and then grouped by neighborhood. The data frame included the longitude and attitude of burglary crimes. The data extracted using from the Denver, CO.org crime database will be merged with FourSquare API data to explore the world around where burglary crimes have been committed.

For feature extraction One Hot Encoding is used in terms of categories. Therefore, each feature is a category that belongs to a venue. Each feature becomes binary, this means that 1 means this category is found in the venue and 0 means the opposite. Then, all the venues are grouped by the neighborhoods, computing at the same time the mean. This will give us a venue for each row and each column will contain the frequency of occurrence of that particular category.

Clustering

Clustering is one of the most common exploratory data analysis technique used to get an intuition about the structure of the data. It can be defined as the task of identifying subgroups in the data such that data points in the same subgroup (cluster) are very similar while data points in different clusters are very different. In other words, we try to find homogeneous subgroups within the data such that data points in each cluster are as similar as possible according to a similarity measure such as euclidean-based distance or correlation-based distance. The decision of which similarity measure to use is application-specific.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	elyria-park	Vietnamese Restaurant	Chinese Restaurant	Thai Restaurant	Grocery Store	Sandwich Place
1	aurora	American Restaurant	Theater	Fast Food Restaurant	Brewery	Basketball Stadium
2	baker	Bar	Chinese Restaurant	Breakfast Spot	Coffee Shop	Shipping Store
3	berkeley	Fast Food Restaurant	Bakery	Vietnamese Restaurant	Mexican Restaurant	Dim Sum Restaurant
4	berkeley-west	Convenience Store	Donut Shop	American Restaurant	Mexican Restaurant	Liquor Store

FIGURE 4: TOP 5 ROWS OF VENUE CLUSTERS TABLE

K Means Clustering Algorithm

K Means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to

```
# Run k-means to cluster the neighborhood into 5 clusters

# set number of clusters
kclusters = 5

denver_grouped_clustering = denver_grouped.drop('Neighborhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(denver_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

```
array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1], dtype=int32)
```

FIGURE 5: K MEANS ALGORITHM, K = 5

make the inter-cluster data points as similar as possible while also keeping the clusters as different

(far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

	NEIGHBORHOOD_ID	GEO_LON	GEO_LAT	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	alamo-park	-105.004940	39.697324	1	Vietnamese Restaurant	Chinese Restaurant	Thai Restaurant	Grocery Store	Sandwich Place
1	aurora	-105.004569	39.744035	1	American Restaurant	Theater	Fast Food Restaurant	Brewery	Basketball Stadium
2	baker	-104.992522	39.711524	1	Bar	Chinese Restaurant	Breakfast Spot	Coffee Shop	Shopping Store
3	berkeley	-105.039488	39.717925	1	Fast Food Restaurant	Bakery	Vietnamese Restaurant	Mexican Restaurant	Dim Sum Restaurant
4	berkley-west	-105.052578	39.719239	1	Convenience Store	Donut Shop	American Restaurant	Mexican Restaurant	Liquor Store

FIGURE 6: TOP 5 ROWS OF DATA AFTER K MEANS

Results

Intuitively the results seem promising, in terms of holding some patterns about the dataset. The clusters seem generally dispersed geographically.

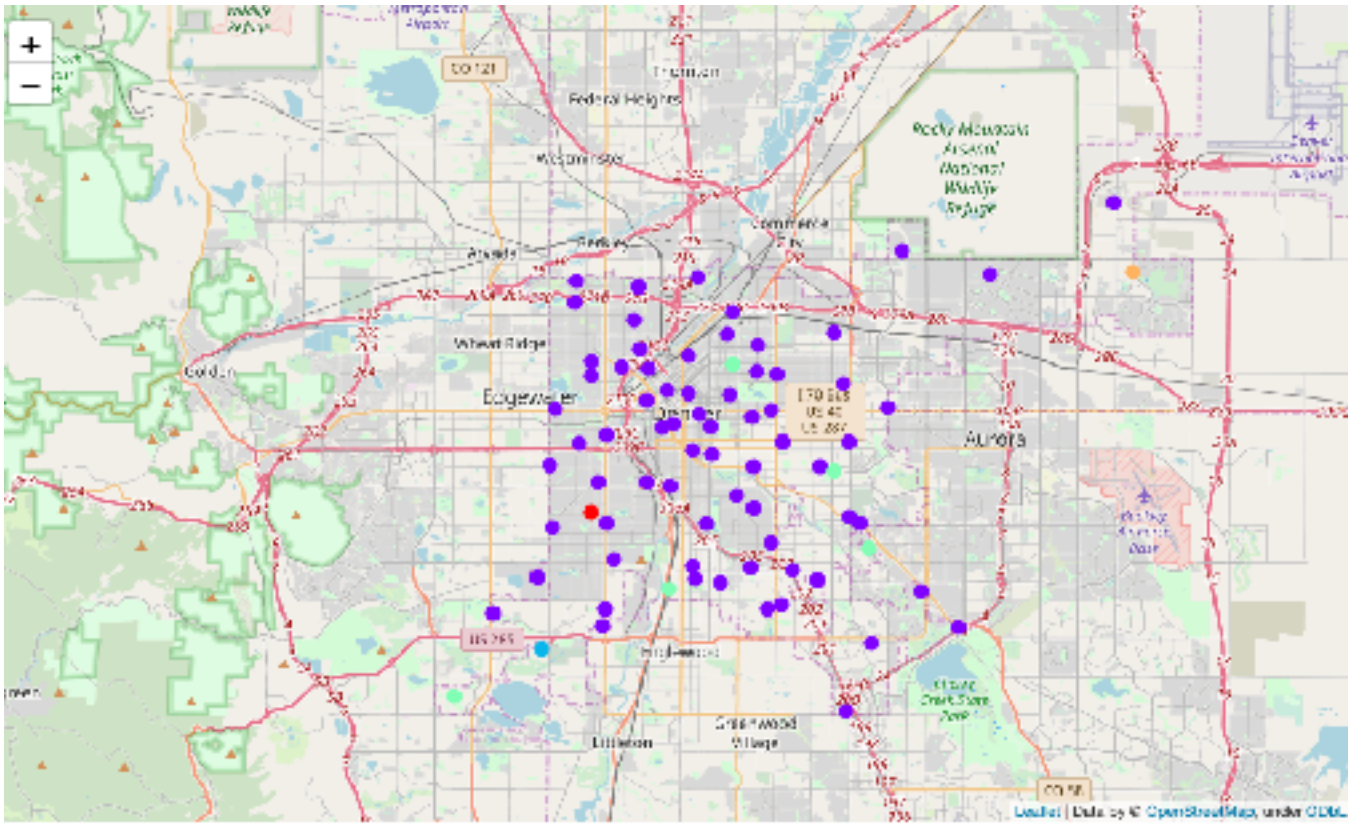

FIGURE 7: CLUSTER MAP

Cluster 1: Red

The 1st most common venue in this neighborhood is a Seafood Restaurant, followed by Women's Store, Flower Shop, Fishing Spot and Fish Market. Only one(1) neighborhood was associated with this cluster.

Cluster 2: Purple

Largest cluster with 1st most common venue ranging from Vietnamese and American Restaurants, Pharmacy, Convenience Stores, Liqour Store, Bars and others; followed by Chinese, Mexican and

Asian Restaurants, Spa, Stadium and Coffee Shop. This cluster seems to be include well established venues across the city centre. There are seventy(70) neighborhoods associated with this cluster.

Cluster 3: Orange

The 1st most common venue in this neighborhood is Park, followed by Women's Store, Donut Shop, Flea Market and Fishing Spot. Only one(1) neighborhood was associated with this cluster.

Cluster 4: Pale Green

The 1st most common venue are Parks, followed by varying common venues from Trail, Lake, Dog Run and Fishing Spot. There are five(5) neighborhoods associated with this cluster.

Cluster 5: Blue

The 1st most common venue Gym, followed by Women's Store, Dog Run, Fish Market and Fishing Spot. Only one(1) neighborhood was associated with this cluster.

Discussion

The Denver, CO Burglary data when clustered are generally dispersed across the geography. The clusters do not overlap, however there are clusters that only have one(1) neighborhood. Cluster 4, primarily consist of outdoor venues whilst Cluster 2 is associated with venues found in a busy city centre. This analysis provides context for the security company to establish a business to secure these venues. Cluster 2 provides the target neighborhoods for the company to roll out their service.

Conclusion

The analysis of the Denver, CO crime data; specifically the Burglary data has indicated the neighborhoods that have the potential to develop a security service. The analysis could be further used for other categories of crime data in Denver, CO. The company may develop additional services based on the analysis of the other crime data.

Cluster 2, would be the likely opportunity to begin deploying the security service. This cluster has common venues that is associated with a city centre and likely to be affected by burglary crimes.

References

- https://en.wikipedia.org/wiki/Denver
- www.denvergov.org
- https://labs.cognitiveclass.ai/tools/jupyterlab/lab/tree/labs/coursera/ML0101EN/ML0101EN-Clus-K-Means-Customer-Seg-py-v1.ipynb
- https://labs.cognitiveclass.ai/tools/jupyterlab/lab/tree/labs/DV0101EN/DV0101EN-3-5-1-Generating-Maps-in-Python-py-v2.0.ipynb
- https://matplotlib.org
- https://python-visualization.github.io/folium/modules.html