Pizza

Group members:

Jerrell U2220449D

Melvin U2220256B

Winfred U2221912F

1

Our Motivation

Our goal

We are business owners who want to open a new pizza franchise in the US

Factors affecting success

1. Location
2. Pricing
3. Competition

Dataset used :

1. Pizza Restaurants US by Chirag Desai from Kaggle.

2. List of United States cities by population from wikipedia.

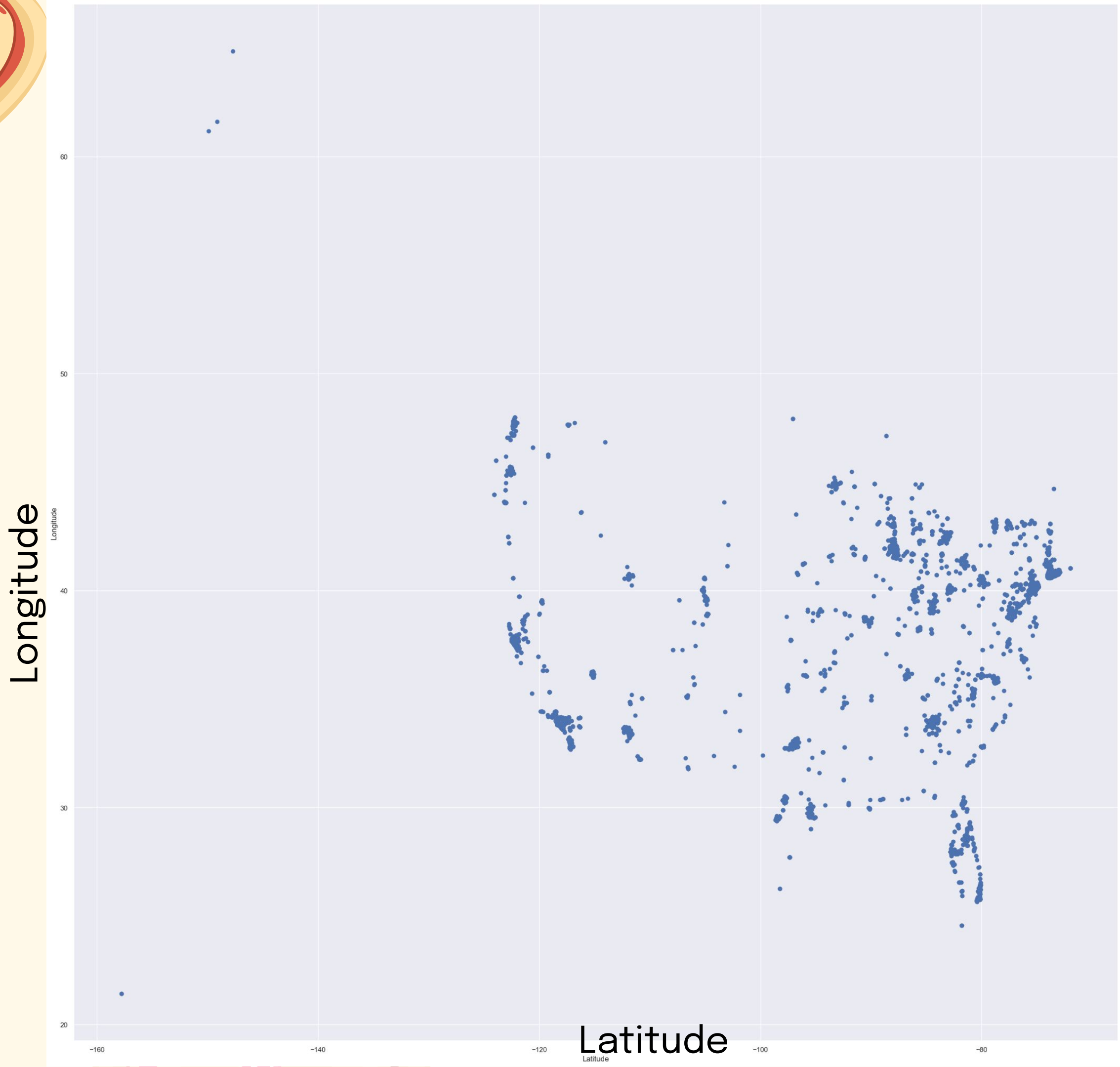
2

Setting the stage

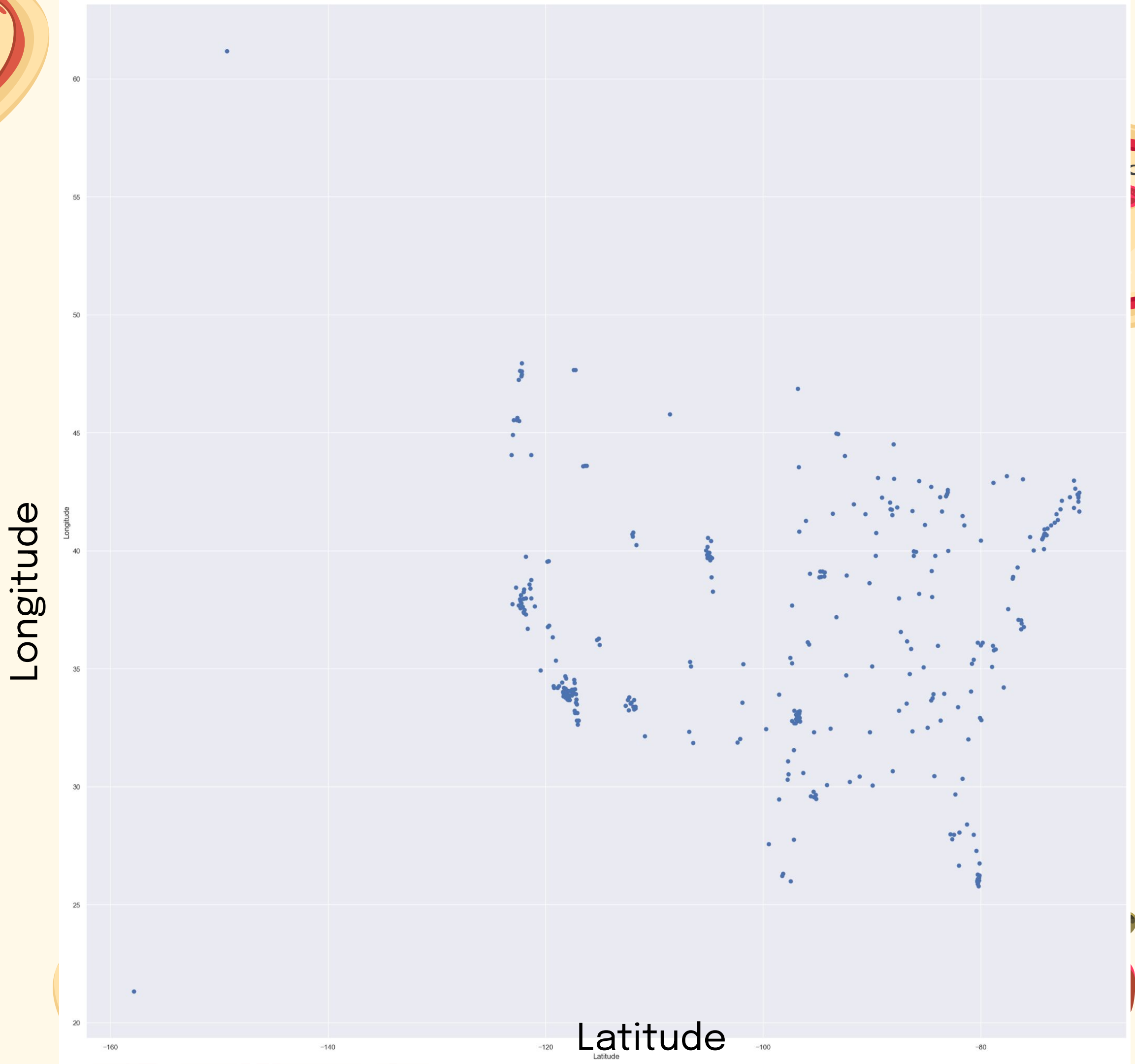
Exploratory Data Analysis

- Scatter plot of pizza shops and cities by location
- Bar plot of cities by population

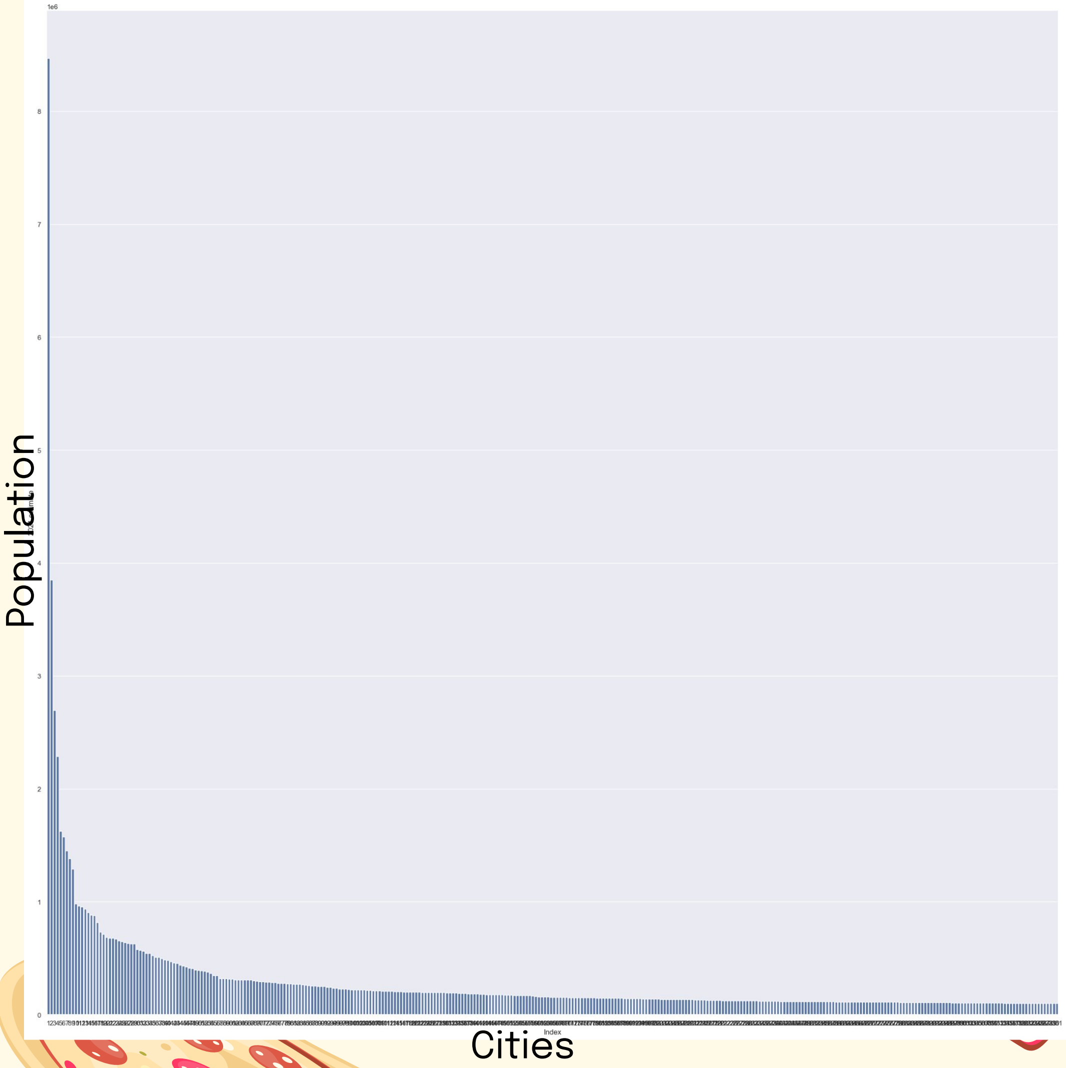
Pizza shops


Longitude
Latitude
60
50
40
30
20
−160
−140
−120
−100
−80
Longitude
Latitude

Cities



Bar plot for cities



Setting up the ML analysis

- Cluster every pizza joint into an optimal number of cluster
- Allocate each city into a cluster
- Calculate the population/ pizza shop in each cluster
- Calculate the median price of a pizza in each cluster

Cleaning & preparation

1 Longitude & Latitude

- Identify US mainland geographical longitude and latitude
- Remove outliers (such as Hawaii & Alaska)

```
cleaned = pizza[(pizza['latitude'] < 50) & (pizza['longitude'] > -140)]
long_clean = cleaned["longitude"]
lat_clean = cleaned["latitude"]
```

```
cities = cities[(cities['Latitude'] < 50) & (cities['Longitude'] > -140)]
long_clean = cities['Longitude']
lat_clean = cities["Latitude"]
```

Cleaning & preparation

2 Price of Pizzas

- Remove the duplicates
- Calculating the median price / shop

```
pizza_clean = cleaned.drop_duplicates(subset=['address'])
```

```
grp = cleaned.groupby('address')
median_price = []
for add in cleaned.address.unique():
    median_price.append(grp.get_group(add)["menus.amountMax"].median())

pizza_clean['Median_Price'] = median_price
```

3

Core Analysis

Process

Step 1

Choosing an appropriate clustering algorithm

Step 2

Optimisation of algorithm through cost function

Step 3

Overlay population dataset

Optimised

Step 4

Cluster level analysis

Conclusion

Clustering Algorithms

K-means

Centroid-based algorithm

Hierarchical

Non-spherical clusters

Fuzzy

Multi-cluster assignment

Density

Density of data points and handles outliers well

Model

Probabilistic models to identify clusters

Spectral

Data with complex structures

K-Means

Benefits over Others

- Simple to implement
- Easily interpreted and analysed
- Centroid-based algorithm allows for overlay of secondary dataset (population)
- Can form clusters quickly and efficiently for large datasets

Cost Functions / Metrics

1. **Inertia (SSE)**
   Total distance

2. **Distortion**
   Average distance

3. **Silhouette**
   Individual
   Intercluster : Intracluster Distance

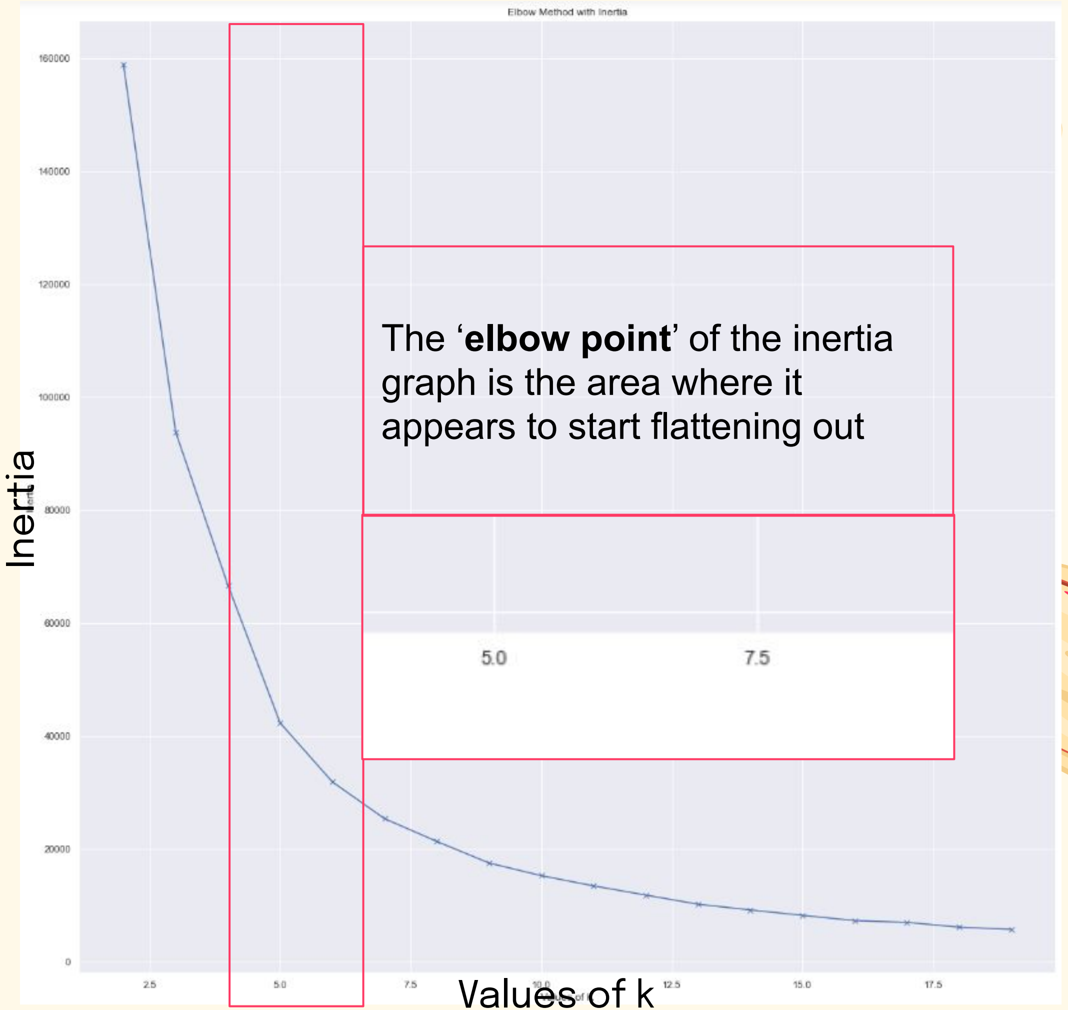
4. **David Bouldin**
   Cluster
   Intercluster : Intracluster Distance

Inertia

Description

- **Sums** the distance between all data points and its assigned centroid
- Look for ‘**elbow point**’

Ideal k [4:8]

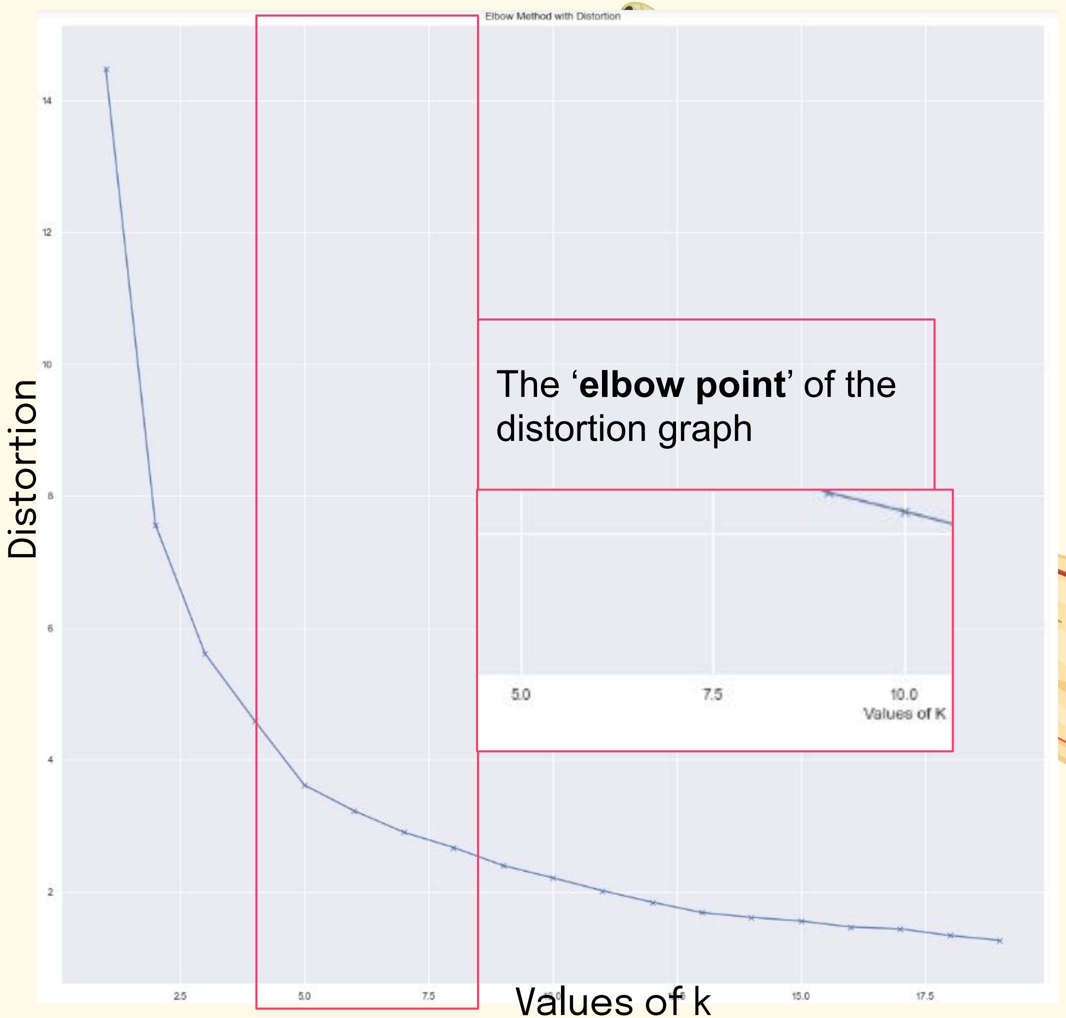


Distortion

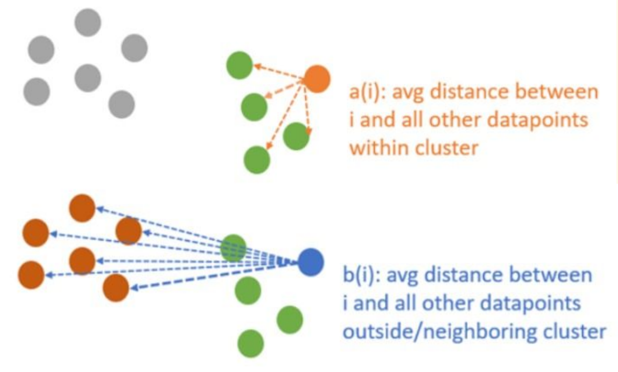
Description

- **Averages** the distance between all data points and its assigned centroid
- Look for ‘**elbow point**’

Ideal k [5:10]

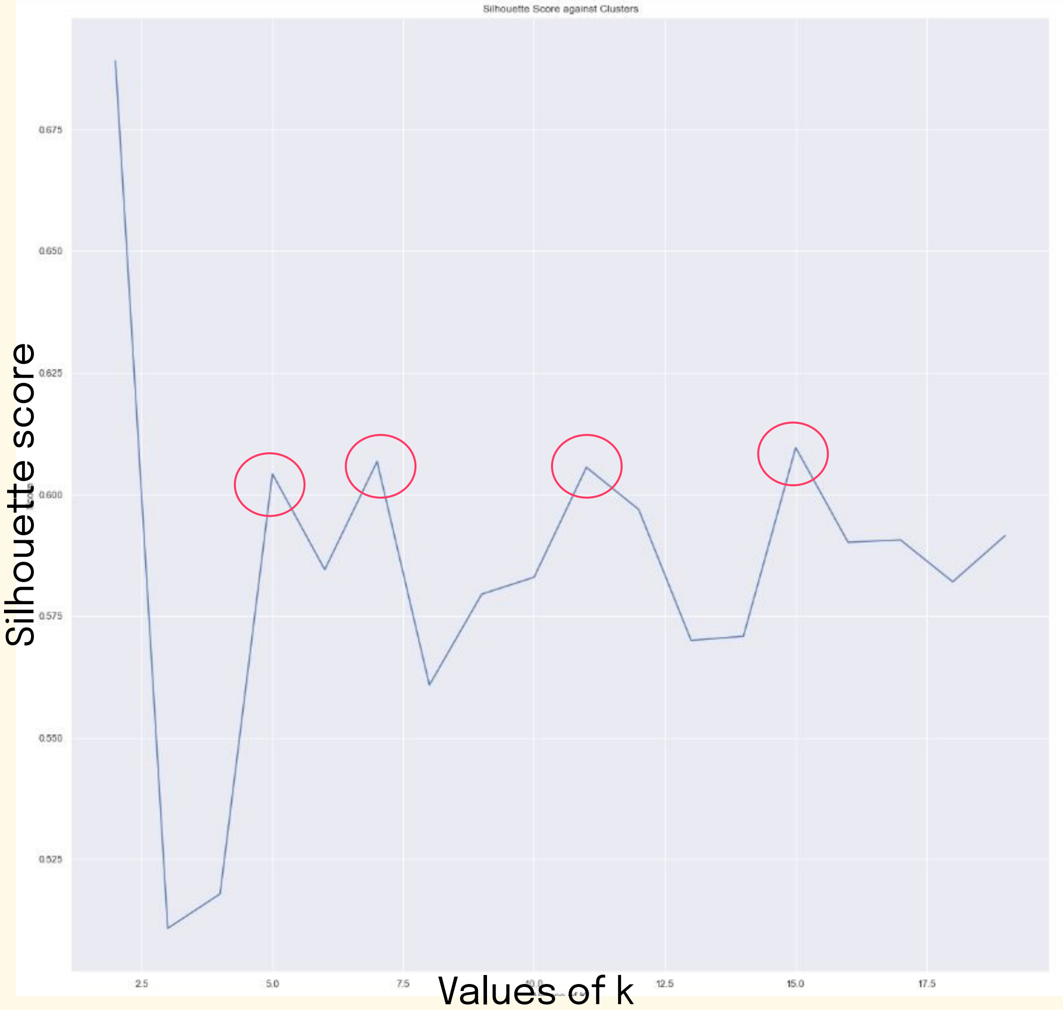


Silhouette



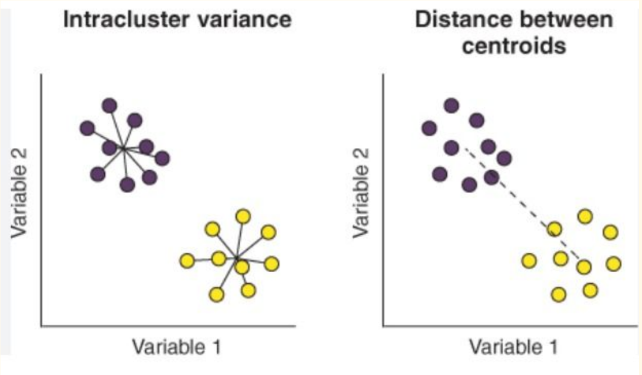
Description

- Measure of similarity between **data points**
- Compares *intercluster* to *intracluster* distances
- Look for **highest points**



Ideal k [5,7,11,15]

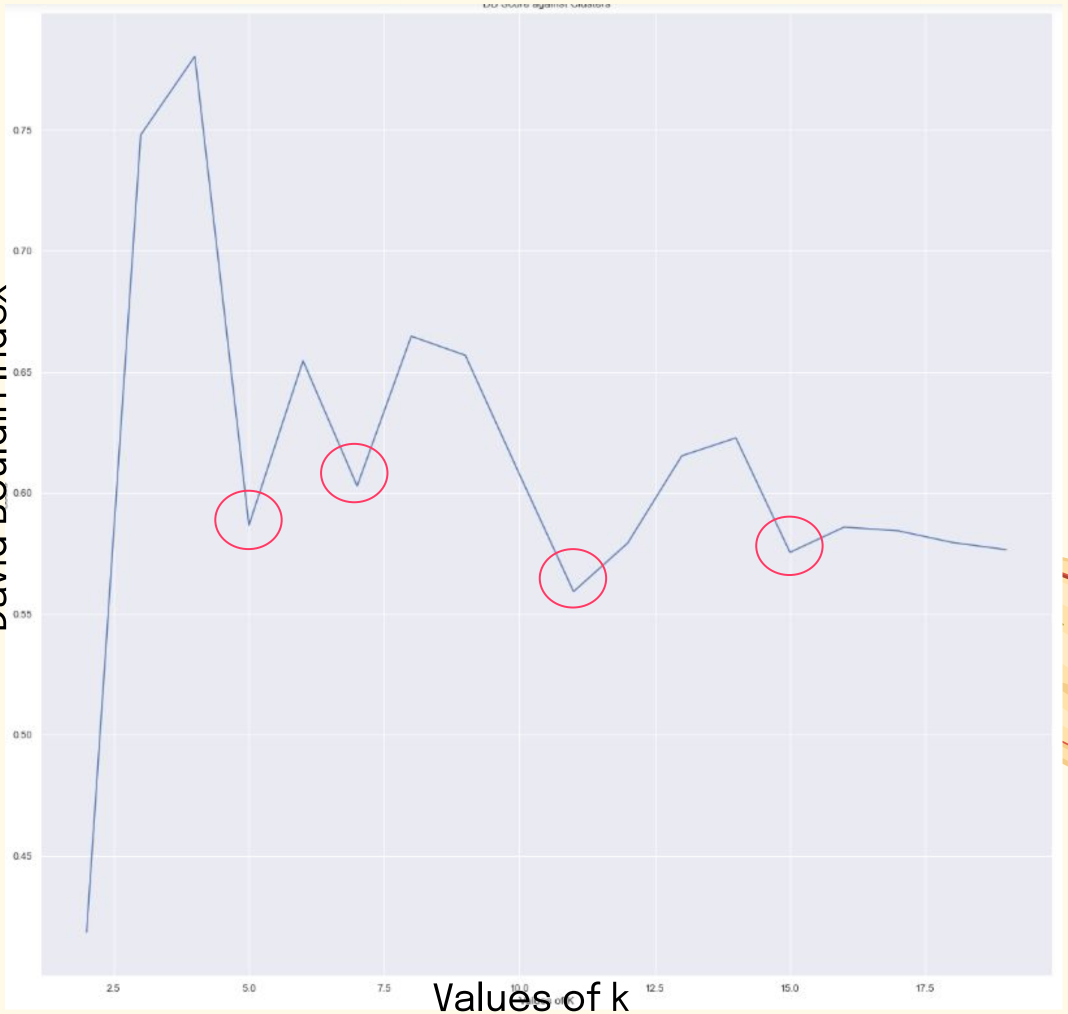
David Bouldin



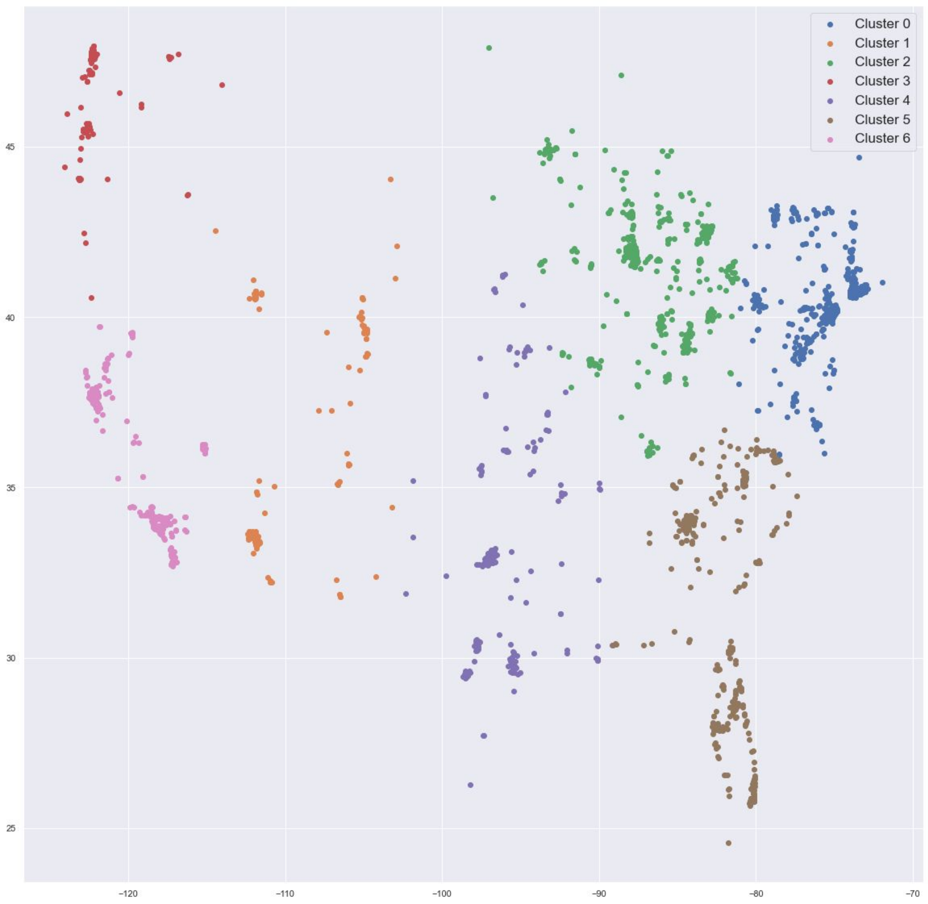
Description

- Measure of similarity between **clusters**
- Compares *intercluster* to *intracluster* distances
- Look for **lowest points**

Ideal k [5,7,11,15]

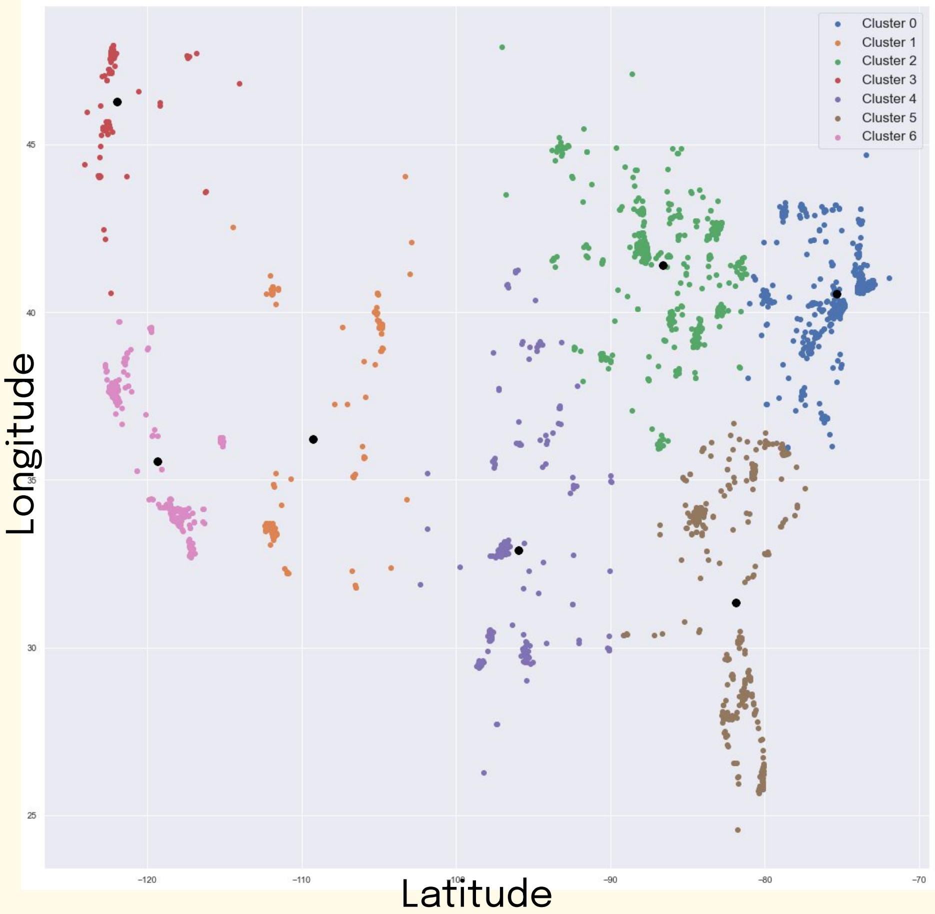


Choosing k=7



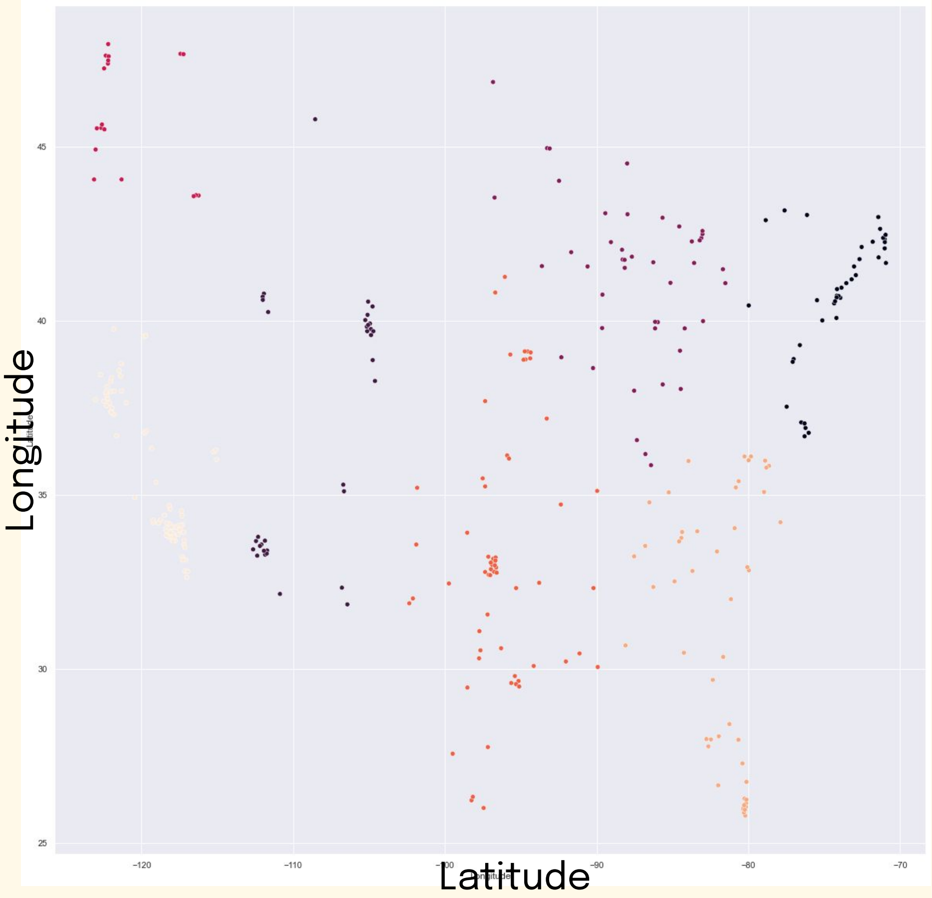
- k=7 is optimal as supported by all the other metrics
- k=5, k=11 and k=15 are good as well, but they are far from the elbow point in the inertia & distortion graph

Find the centroids of the clusters

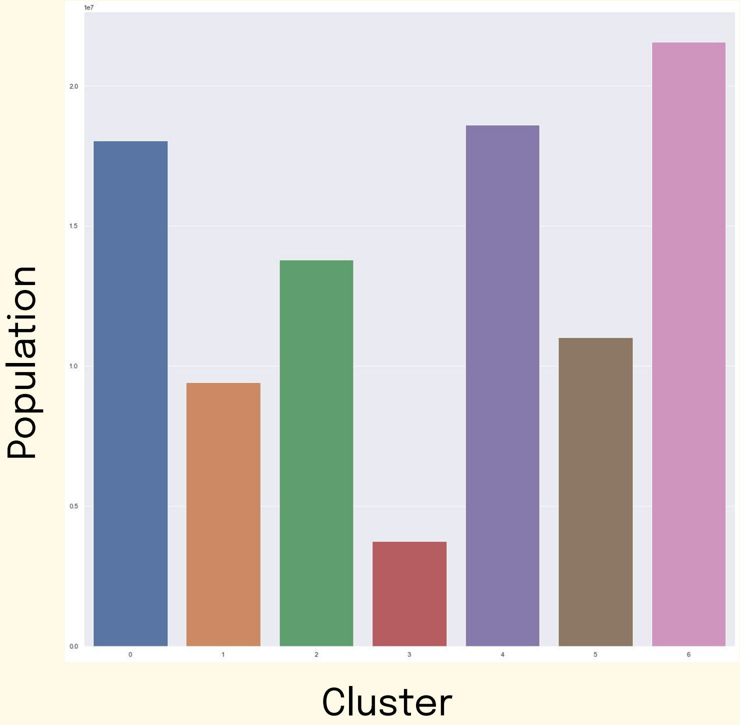


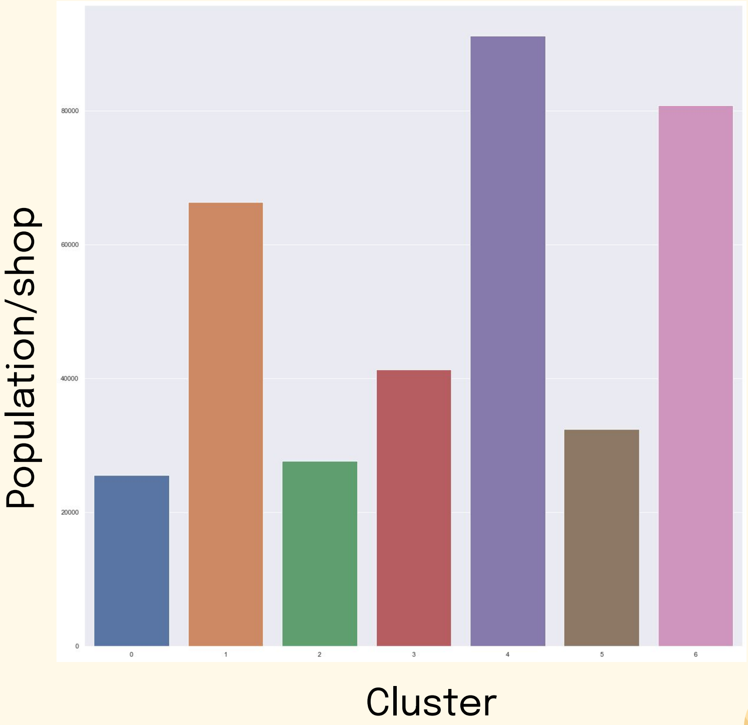
Step #3: Population Overlay

Assign city to cluster centroids

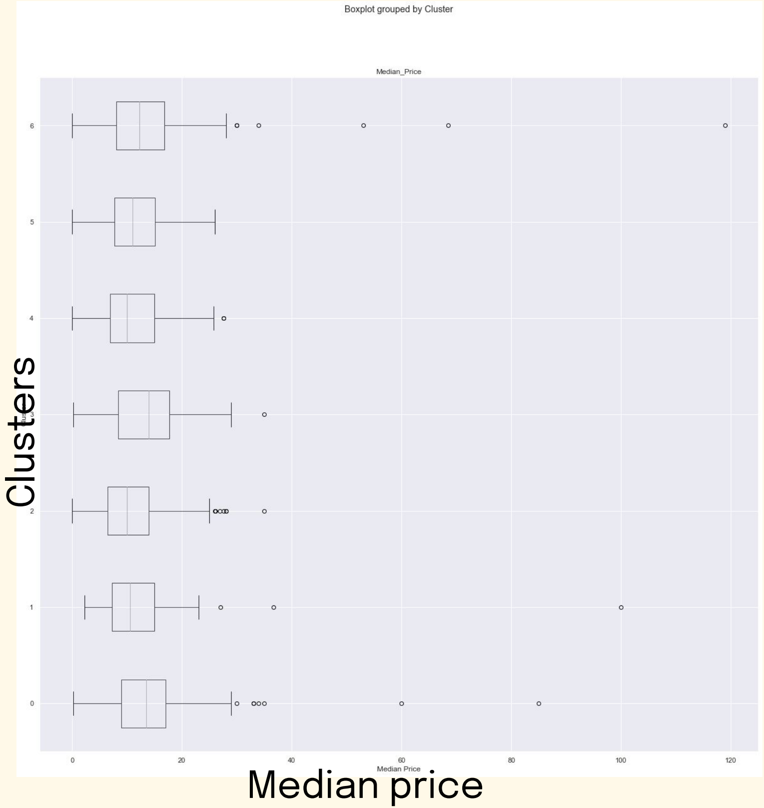


Step #4: Cluster analysis

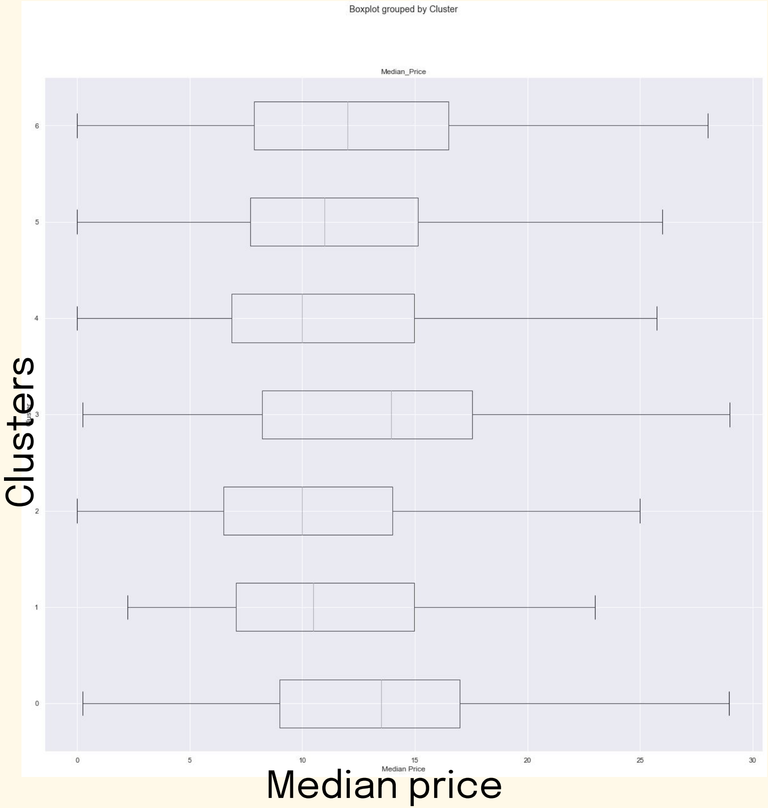




Median prices



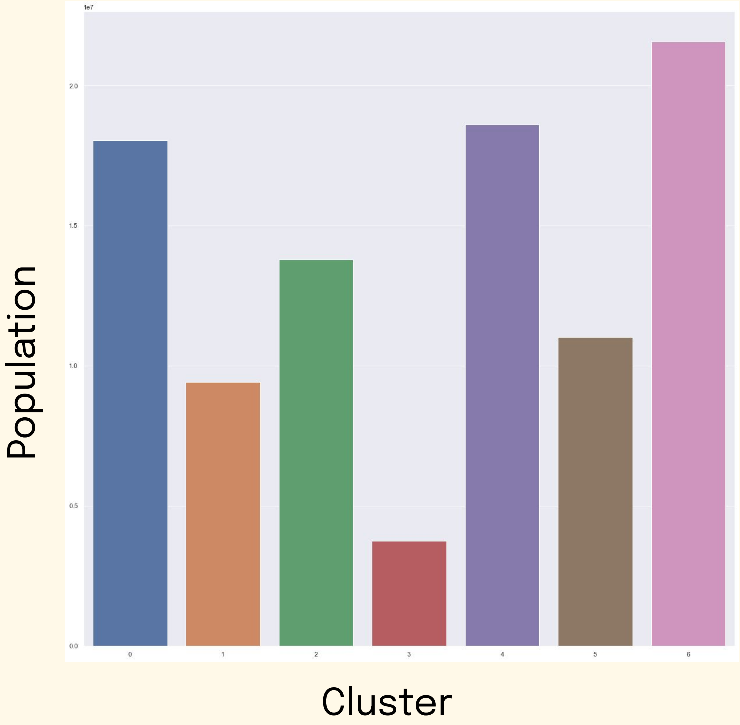
Before



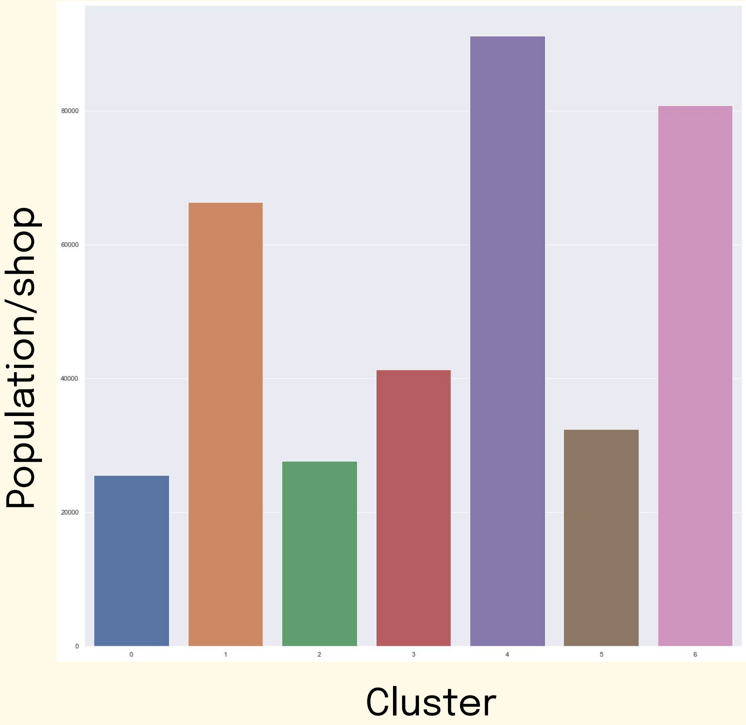
After

4

Conclusion



Population
Cluster



Population/shop
Cluster

Competitors' menu item prices

Median price of rival shops set ~**$10**. To be used as a guideline for menu pricing

Outcome

- Cluster 4 has the highest pop/shop
- Lowest level of competition with rivals
- It also has 2nd highest population
- $10 serves as a good median price in this cluster

What we learned

- Using K-Means clustering
- Using different metrics to optimise the number of clusters

Thank you!