# Coursera Capstone Project
# Choose The Right Hotel To Stay In

Radoslaw Cholewinski, May 2020

# Business Case

- Hotel search engine provide users with simple filters
- It would be useful to differentiate between hotels based on venues, which are located within a walking distance (e.g. close to restaurants, close to sightseeings)
- *k-means* algorithm could be implemented to cluster hotels using location data from the Foursquare API
- Commercial applications:
  - More efficient hotel search engines
  - Targeting users of booking websites
  - Price discrimination for booking websites
  - Differentiation of tourist tax by local authorities

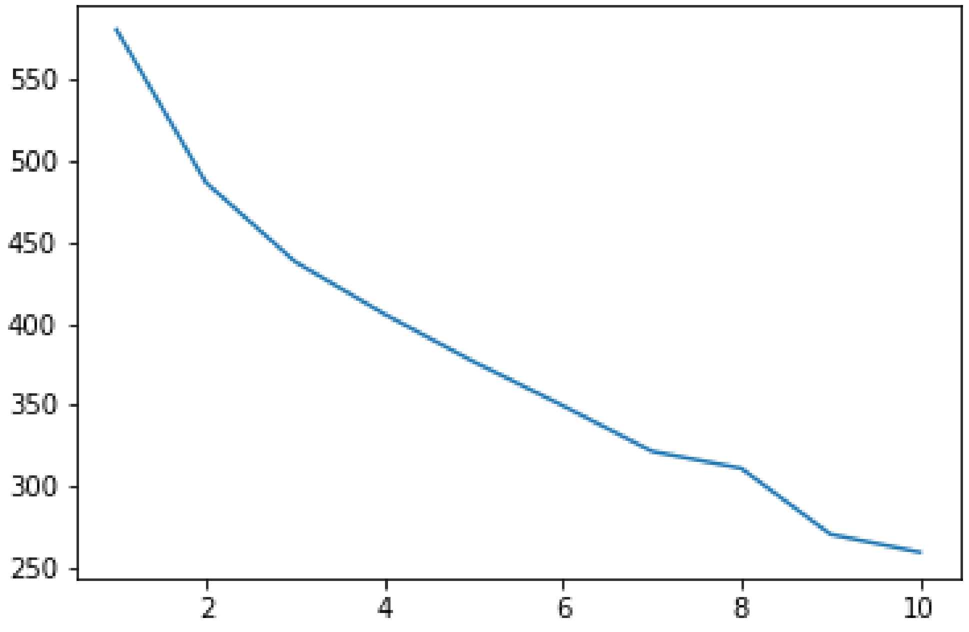
# Data

- Data for Naples, Italy
- 98 recommended hotels by Foursquare, of which 89 analyzed
- 3600 different venues within 500 m distance from each hotel
- 139 categories of venues
- In each hotel’s neighborhood there are on average 40 venues

# Methodology

- Simple *k*-means clustering algorithm
- Features:
  - One-hot encoding of venues based on their category
  - For each hotel the number of occurrences of a given venue is divided by the largest numer of occurrences of that venue for all hotels – **score**
- Choice of *k*
  - More an art than science
  - It is desirable that clusters can be assigned with meaningful labels
  - Value of *k* may affect the shape of clusters
  - Too many clusters don't really provide information

# Results

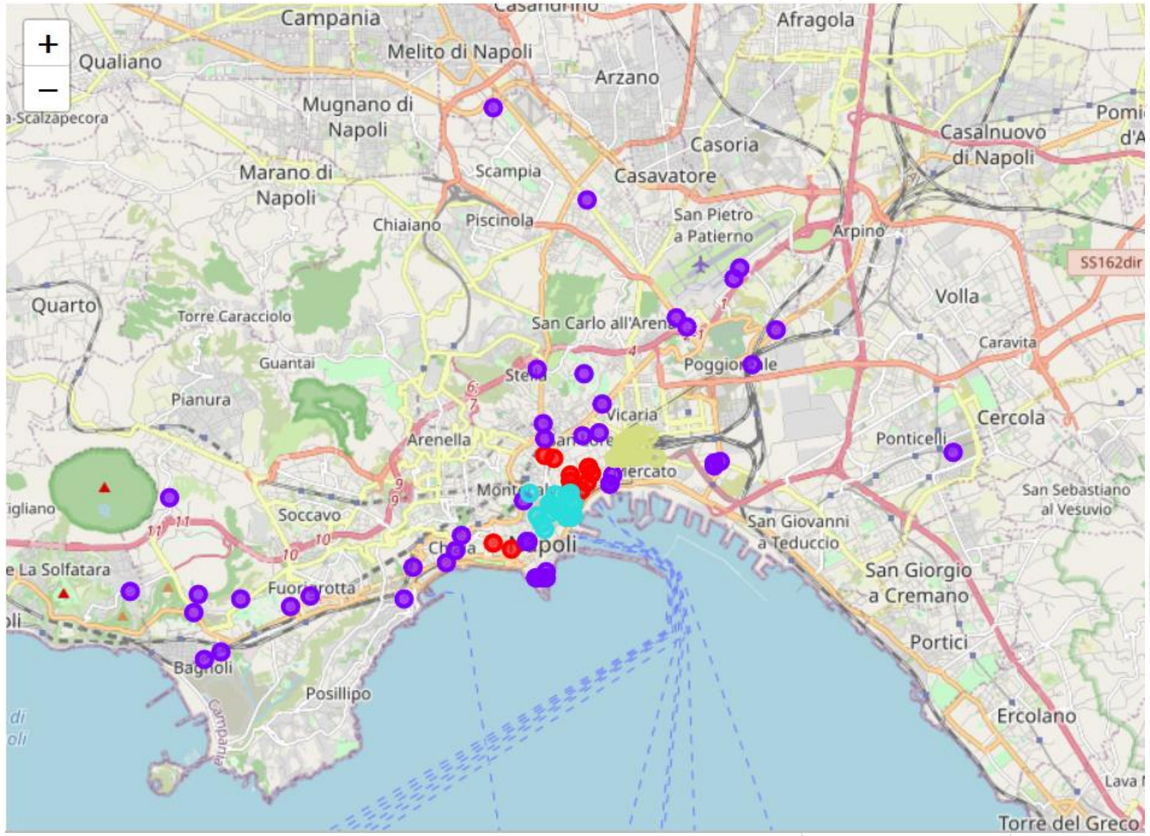



Inertia shows the steepest decline for $k$ between 2 and 7

# Results

- For k = 3 and k = 4 hotels within a cluster are physically located close to each other, and there exists one cluster of other hotels
- When k = 5, additional cluster consists of only two hotels
- For k = 4 clusters can be meaningfully labeled:
  - #0 – delis, sandwiches, fried chicken, plazas – tourist area (or tourist traps)
  - #1 – pizza places, convenience stores, diners, public transport, Italian restuarant – great for independent travellers
  - #2 – Asian restaurants, convenience stores, fountains, hotels, hostels – typical hotel district
  - #3 – bakeries, food & drink shops, gift shops, candy stores – main train station neighborhood

# Results – clusters visualized




Red – tourist area (traps)
Blue – independent travelers
Pink – hotel district
Green – train station area

# Discussion

- Each cluster caters for different kinds of tourists
- Three clusters located in physical neighborhoods (tourist sights, hotel district and train station)
- One cluster spread across the city (not one physical neighborhood) with hotels close to local restaurants, convenience stores and public transport – choice for independent travelers
- Based on existing filters in hotel search engines, it would be difficult to identify the ‚independent traveler cluster’

# Recommendations

- Hotel search engines could be enhanced by clustering hotels using simple *k*-means algorithm
- Hotel clusters could be labelled on a basis of the most common venues near hotels within each cluster
- Great tool for booking sites (targeting users, price discrimination)
- Local authorities could use cluster analysis to differentiate tourist tax