EATING MY WAY THROUGH EUROPE

BART ONKENHOUT

A QUEST TO DISCOVER THE BEST FOOD-CITY TO RIVAL CHICAGO

INTRODUCTION

Background

- Plan to move to Europe for work assignment and employer has HQ in several EU cities. Where should I apply?
- I really like food – perhaps I should find the city with the most similar food scene to my current home city of Chicago?

Importance

- Evidence-driven decisions
- Data science as a way to reduce information overload and algorithmically make an optimal decision

PROBLEM STATEMENT

Which city/cities in Europe should I pick for my next work rotation so that I still have similar food options to Chicago?

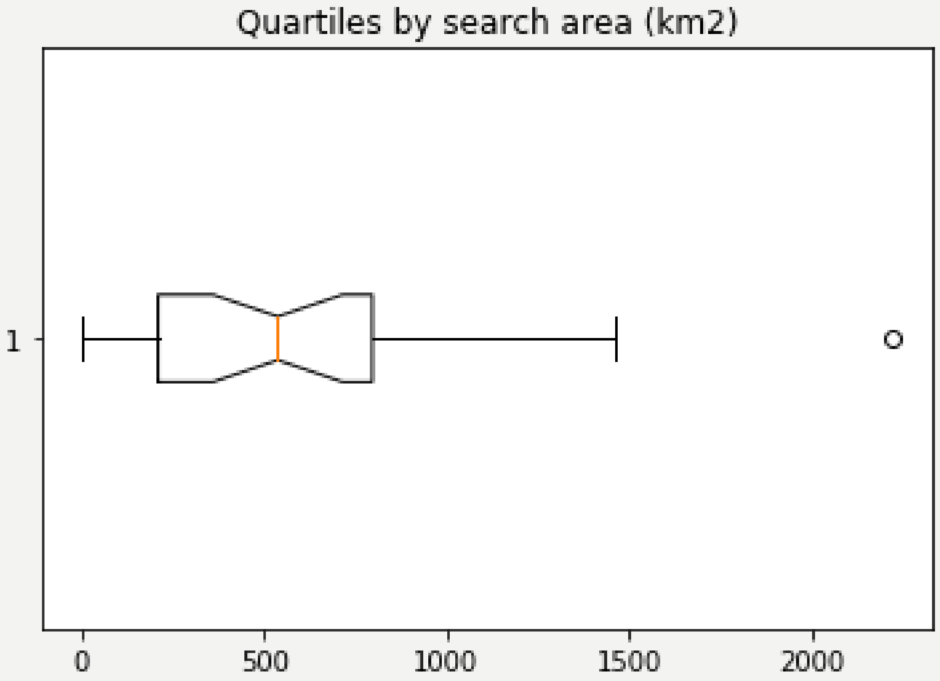
DATA

- List of **European cities** in which employer has offices as a base comparison set
- **Geolocation data** of each city so it can be plotted on a map and fed into the Foursquare API
- **Foursquare API data** for looking up venues in each city
- **GeoJSON shapes** of each city so we can use a GeoJson layer in Folium to outline each city on the map

METHODOLOGY - ETL

Scraped necessary data from various sources and cleaned/scrubbed everything

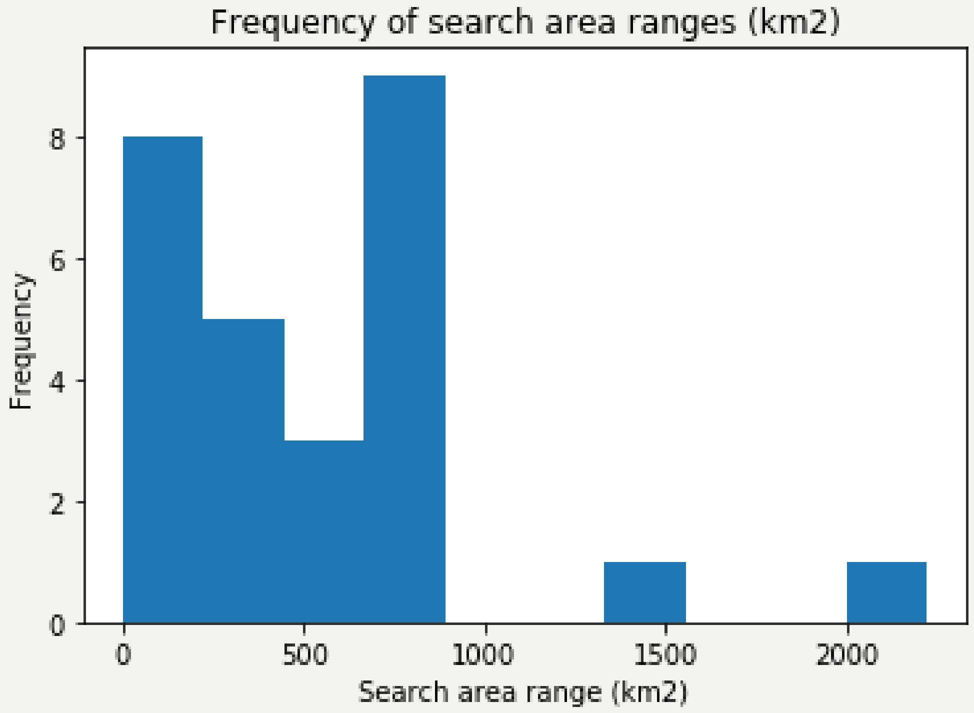
METHODOLOGY – EXPLORATORY ANALYSIS



- Calculated square km area for each city based on geolocation boundaries

- Checked interquartile ranges for square km area for each city and found **Moscow** to be an outlier.

METHODOLOGY – EXPLORATORY ANALYSIS



- Checked frequencies for each search area range and found an additional outlier in **Chicago.**

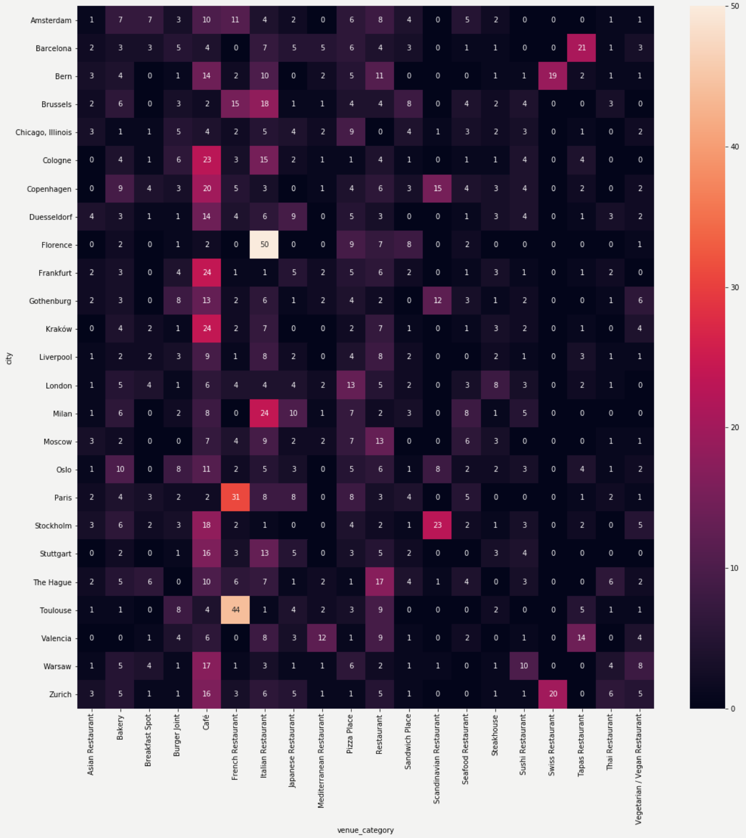
- Checked interquartile ranges for square km area for each city and found **Moscow** to be an outlier.

METHODOLOGY – EXPLORATORY ANALYSIS

	city	dist_from_ctr_ne	dist_from_ctr_sw	search_area_km2	area_quartile	possible_iqr_outlier	z_score
10	Helsinki	0.622382	0.622401	0.618574	bottom	False	-1.215256
19	Rotterdam	0.006541	0.006541	0.000077	bottom	False	-1.216581
23	Moscow	31.800554	35.451290	2220.463037	top	True	3.542410
26	Chicago, Illinois	18.376116	36.710079	1457.424863	top	False	1.907034

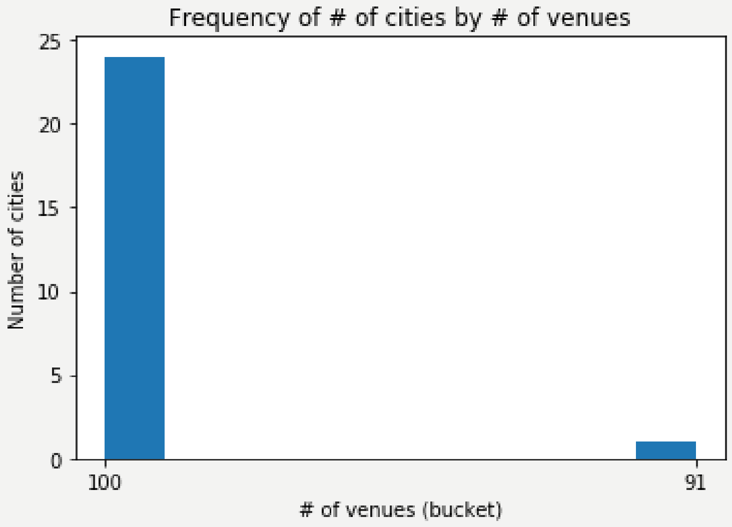
- Also checked z-scores for each search square km area to check for outliers.
- Decided to drop **Helsinki** & **Rotterdam** due to likely errors in OSM data.
- Kept **Moscow** because no likely errors.
- Kept **Chicago** because it is the basis of comparison, and thus required.

METHODOLOGY – EXPLORATORY ANALYSIS



- Heat mapped Foursquare API data to each city to find frequency.
- Data looks to be in order and of high quality due to high visual correlation in accordance with expectations – many French restaurants in French cities, Italian restaurants in Italian cities, lots of cafés in each city, etc.

METHODOLOGY – EXPLORATORY ANALYSIS

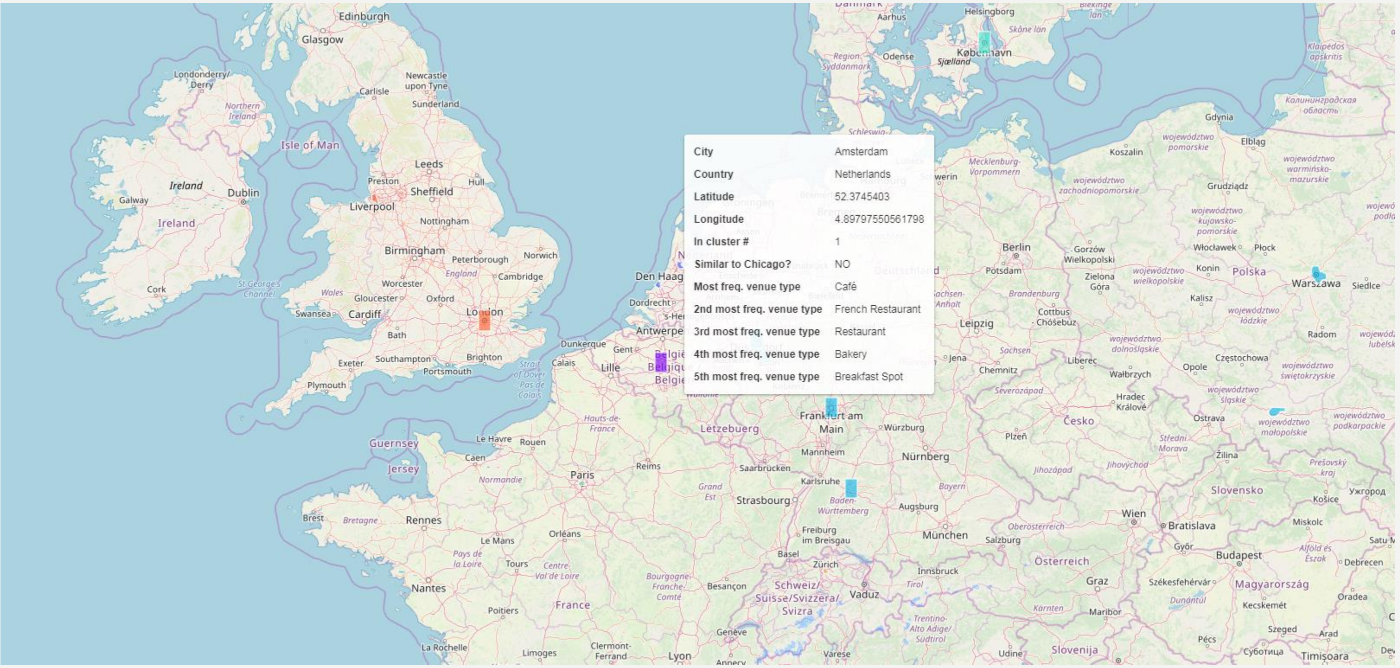


- Checked each city to see if it is normalized against the others, as well as returning a representative sample of venues from Foursquare API (see notebook and report for details)
- Bern was underrepresented and had one-hot encoding corrected by underrepresentation factor

METHODOLOGY -- CLUSTERING

- K-means clustering used due to good performance and applicability in general clustering
- 116 features in one-hot encoded data results in computationally expensive dimension reduction for many other algorithms
- Possible other algorithms considered: DB SCAN, SVM, binary trees, etc.

RESULTS



- Enriched GeoJSON data with k-means clustering results and embedded into Folium Map
- Added GeoJson layer to Folium map, colored according to each cluster
- Allows visual exploration of the model – similar cities to Chicago same color as Chicago

DISCUSSION

- Model returned London or Liverpool as most similar to Chicago in terms of food scene
- However, there may be systemic bias in the model – requires further investigation
- K-means clustering is sensitive to **vanishing gradient** problem and initial centroid coordinates, but London appears most often in the list of cities similar to Chicago.
- Intuitively, London makes sense and I should start investigating London as a possible next assignment.
- Important to use model as supplement to decision-making, not replacement for.