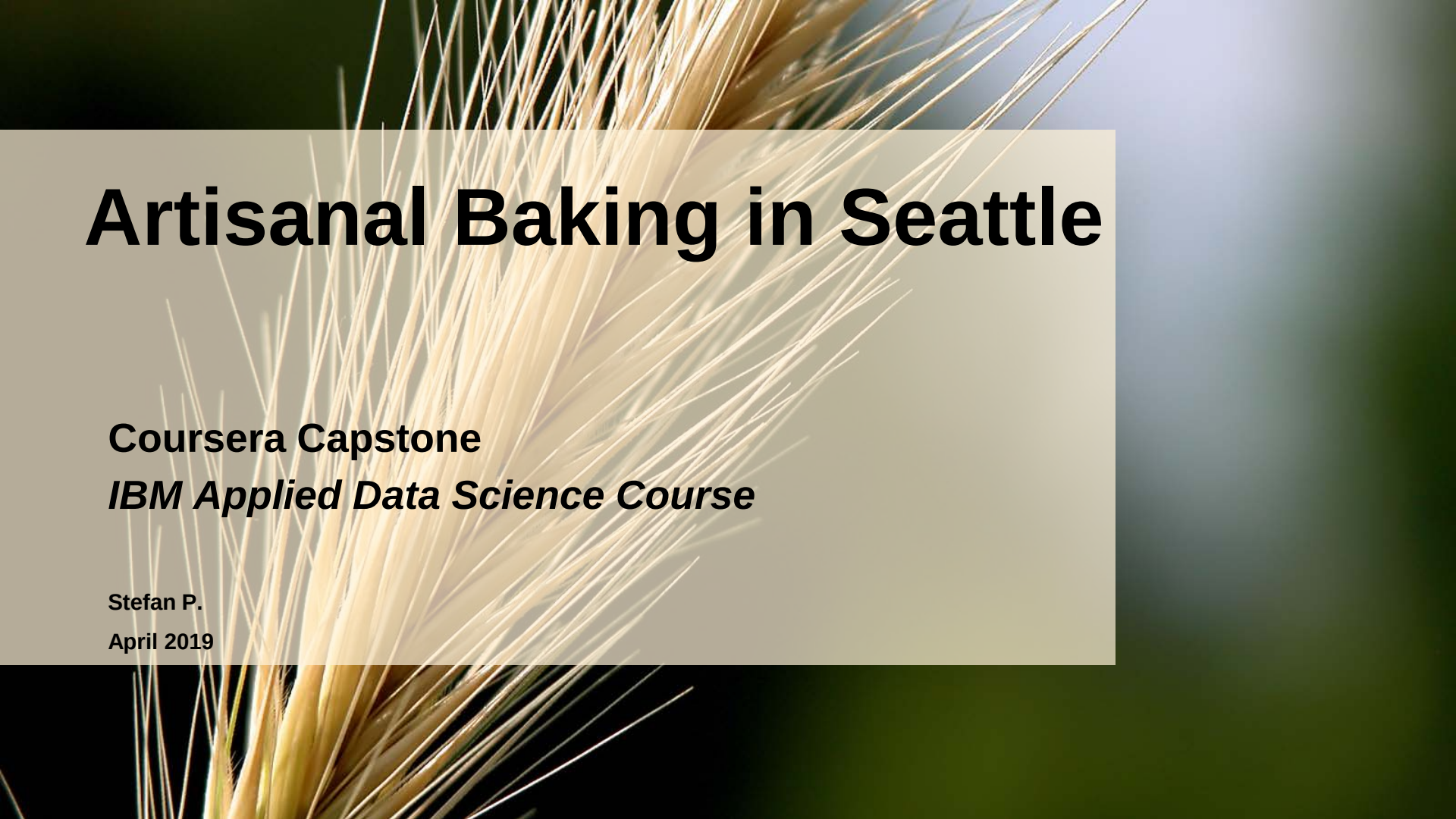


Artisanal Baking in Seattle

Coursera Capstone

IBM Applied Data Science Course

Stefan P.

April 2019

Introduction

Seattle, the ever growing tech city where hipsters, nerds and families flourish surrounded by amazing mountain ranges, ocean and lake waters and clean air. One would say this is the ideal environment to stimulate farmers to grow ancient grains, feed it with the best mountain water and minerals to then create the ultimate healthy bread flower. This flower to then be used by artisanal bakers who make the most beautiful breads feeding the people in Seattle. A dream by a Dutchman who is moving to Seattle.

Objective & Scope

By using machine learning techniques and clustering were are trying to find out:

1. To identify what the coverage of bakeries is across Seattle;
2. To identify where the best location is to open a new bakery;

The scope of this project and it's insights is limited to the "free" sources available for this research (Foursquare API, Wikipedia). Key sources that will be used are a list of neighborhoods in Seattle, coordinates of these neighborhoods and free venue data available in Foursquare.

The outcomes will be useful for anyone who dreams to open an artisanal bakery in Seattle.

Data Sources (& their extraction)

To generate the initial insights I used three different data sources. The table below show the source, a brief description and the method of extraction.

Data source	Description	Extraction
Seattle Neighborhoods webpage (https://seattle.findwell.com/seattle-neighborhoods/)	Leverage a complete list of Seattle neighborhoods published online to build a list.	BeautifulSoup
Neighborhood Latitude / Longitudes	Connect the list of Seattle Neighborhoods to Geocoder to pinpoint their locations and provide input for KMeans to cluster.	Python Geocoder
Venue Data (Foursquare)	Leveraging Foursquare data to identify which areas are high and low on bakeries to set a first step to identify a good new location.	Foursquare API

The Methodology

The methodology for this Capstone project consists of 4 key steps:

1. Grab a list of neighborhoods in the city from a reliable web-source. In order to do this Python & Beautifulsoup packages are used. This however will only result in a list of neighborhoods. To generate meaningful results we need to also get the geographical data and enrich this with Bakery information in clusters (2,3,4);

2. Use the Geocoder & Folium packages to convert the neighborhoods into geographical coordinates (latitude & longitude) and add the data to the initial list from step 1 and generate a map to validate the correctness of the data;

3. To enrich the neighborhoods the Foursquare API is invoked to get the top 100 venues that are within a radius of 2000 meters and capture the venue name, unique venue categories, venue latitude and longitude for each neighborhood.

4. Analyze the data to determine for each neighborhood the frequency of occurrence of each venue category and filter by “Bakery”. Apply K-means to identify k number of centroids, allocate every data point to the nearest cluster (while keeping the centroids as small as possible). This machine learning approach will help identifying the density of bakeries to find opportunity for a new one.

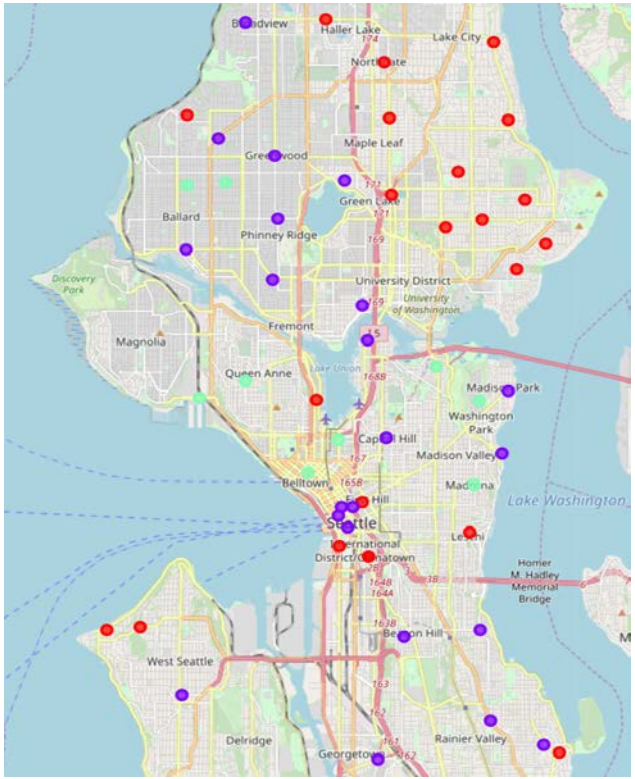
Results

Three key clusters have been identified in Seattle with respect to the bakeries:

1. Cluster 0 (red): Neighborhoods with moderate number of bakeries;

2. Cluster 1 (purple): Neighborhoods with low number to no existence of bakeries;

3. Cluster 2 (green): Neighborhoods with high concentration of bakeries;



Discussion, Limitations (& next steps)

The most central band of Seattle has the highest concentration of bakeries in the neighborhoods whereas the South, North East and North West of Seattle (cluster 0 and 1), have a moderate to low degree of bakeries in the neighborhoods. This could pose a great opportunity to open another bakery and potentially a new bakery chain (Starbucks started small too!). There are a few limitations to this type of approach of just looking at density and not including other factors. There may be reasons for low density of bakeries beyond just simple not having one. There may be low demand or demand may be fulfilled by large grocery stores like WholeFoods. People may simply not choose bread as their source for breakfast or lunch because they prefer other types of breakfast /lunch foods (e.g. yoghurt, oatmeal). Lastly by just using free sources of data there is potential critical data missing to build a more complete model and/or apply different Machine Learning techniques.

This first step therefore was just exploratory by nature and needs more variables to show a more complete input for decision making.

Summary / Conclusion

Opening up a bakery can be a fun and exciting journey. The steps as set forth in this project are a good first reconnaissance to assess if opening up a bakery in Seattle can be a viable business venture. The approach (although it has some limitations) shows that in cluster The neighborhoods in cluster 1 have a low concentration of bakeries hence a good opportunity for first mover advantage. As with opening any type of business other factors like actual demand for artisanal bread compared to other breakfast / lunch options, perceived quality of existing bakeries in more crowded clusters, prices of commercial real estate in neighborhoods, income levels and for example existence of major grocery stores are factors to include in future research.