E-Commerce: High Potential Regions

Finding regions within Germany that have a high population density and a low retail store density by utilizing Webscraping, the Foursquare API and Cluster Analysis

Project Report

July 25, 2021

by Kevin Götz

Table of Contents

Abstract

Following soon…
….
….

1. Introduction

1.1 Background

Amid slowing economic activity, COVID-19 has led to a surge in e-commerce and accelerated digital transformation. As closures became the new normal, consumers and businesses have become increasingly "digital," offering and buying more goods and services online, increasing e-commerce's share of global retail from 14% in 2019 to about 17% in 2020. These and other findings are reflected in a new report, “COVID-19 and E-Commerce: A Global Review”, from UNCTAD and eTrade for all partners, which reflects the industry's strong global and regional changes in 2020. (source: https://unctad.org/news/how-covid-19-triggered-digital-and-e-commerce-turning-point, 19.07.2021)

The trend toward e-commerce is likely to continue during the recovery from COVID-19, but the reopening of physical stores will probably have a slowing effect on the growth of E-Commerce companies.

1.2 Problem

Some changes in consumers' shopping habits may be longer-lasting, but there may be several push and pull factors for consumers to leave their house again for an in-person shopping experience offline. After more than one year living in a lockdown the consumers probably are looking forward to experience the normal every day life again and not sit in their flats in front of their screens (push factors). Also, the retail shops are probably offering widespread sales campaigns to attract potential customers and drive revenue again to counteract the lost year of commerce (pull factor).

With the lifting of the restrictions in germany the e-commerce companies no longer have the “monopoly” on shopping and the competition gets more balanced again. The e-commerce industry needs to find strategies to keep up the momentum and not lose their newly gained customers.

1.3 Objective & Relevance

A possible lever for customer retention and customer acquisition could be a geographical analysis of the retail landscape and population. Assuming that consumers with sparse shopping options in their area are more likely to order online, this project aims to find regions within Germany that have a high population density as well as a low retail density. If a e-commerce company is focusing their marketing strategies towards those regions it may maximize their ROI on marketing spending.

2. Methodology

2.1 Data Sources & Collection

The data origins from four different sources in total, which are explained in chronological order below:

1. **Webscraping Zipcode Information**:
   The basis of the analysis are locations which are identified by zipcode, city, quarter, district and region. The data was collected from the website “https://home.meinestadt.de” because it has the most detailed and clean zipcodes for Germany compared to “https://en.wikipedia.org/wiki/List_of_postal_codes_in_Germany” (not detailed enough) or “https://worldpostalcode.com/germany/” (not clean enough).
   The scraping consists of two steps:
   a. Scraping the URLs for the regional pages of Germany
   b. Scraping the regional pages for zipcode data

   Special care was applied when scraping the “meinestadt”-Website. The basic requests from the HTTP-Client (Requests- or Urllib-Library) were blocked from the server (HTTP Error 403 meaning access to the requested resource is forbidden. The server understood the request, but will not fulfill it.) because they were missing the usual request-header containing an Agent-Information when browsing in person and therefore probably detected as a crawler. Using a header (“User-Agent': 'Mozilla/5.0 (Windows NT 6.1)”) and an uniformly randomized time delay (sleep function with numpy.random.uniform) solved the problem.

2. **Webscraping Wikipedia**:
   Wikipedia was used to gain insights about the population and area of the location. In a first step a search string consisting of the zipcode and the quarter name (or city name, if quarter was not applicable for small villages) was build. This search string was then concatenated with the Wikipedia search URL having the settings on "Sort by most relevant" which got the best result page while testing.
   In a second step the result page was crawled for the first entry's URL and finally requested to get to the Wikipedia page of the quarter, city or village (depending on the data).
   Finally the table in the right upper corner (first table on the page) was parsed for area and population information before repeating this process for all zipcodes within Germany (N > 7.200).
   Special care was applied when scraping Wikipedia. Because of regulations to quantity and speed of crawling ("https://de.wikipedia.org/robots.txt") a randomized time delay (sleep function with numpy.random.uniform) was applied again.

3. **Geocoding of Latitude and Longitude Coordinates (OpenStreetMap API)**:
   The distance to retail shops can only be computed (Step 4) when the location data is enriched with coordinates. The library "Geocoder" and an OpenStreetMap API were used to obtain the latitude, longitude and the bounding boxes.

4. **Adding the Retail Information (Foursquare API)**:
   As a last step for the data collection the Foursquare Developer API was used to map Fashion Retail Shops with their distance to the location data. The Foursquare API therefore uses the Latitudes and Longitudes from the previous step and returns all the stores names, categories, latitudes and longitudes in a user-defined circumference. This Data will then be used to geocode the retail density for the zipcode locations.

2.3 Exploring & Cleaning the Data

Following soon…
….
….

2.4 Feature Engineering

Following soon…
….
….

2.5 Model Building

Following soon…
….
….

2.6 Data Visualization

Following soon…
….
….

3. Results

Following soon…
….
….

4. Discussion

Following soon…
….
….

5. Conclusion

Following soon…
….
….