# Analysis for Business Entrepreneurs in Berlin

**Azeer Esmail**
**july/2019**

## 1. Introduction

### 1.1 Background Description

Berlin is the largest city in Germany by both area and population, and 2nd largest in Europe by population within the city limits.
Tourism figures have more than doubled within the last ten years and Berlin has become the third most-visited city destination in Europe, In 2018, the GDP of Berlin totaled €147 billion, an increase of 3.1% over the year of 2017.
All that coupled with its very high diversity where foreign residents originate from about 190 different countries, and the fact that it's a very active city with a booming economy in many sectors, made it a desired destination not only for tourists and expatriates but also for entrepreneurs and businesses.
Therefore studying the opportunities and understanding the status quo before starting a venture in this city is of big importance and the cornerstone of the business's future success.

### 1.2 Problem

Despite its very promising characteristics for entrepreneurs, Berlin is a very dynamic city and ever changing, as a resident here, I noticed how frequently small businesses start and close their doors after a relatively short time.
This is because of the lack of knowledge of the demographics and other variables of a certain area.
To avoid such a scenario of loss, and to ensure higher possibility of success, Data-Science will be used to try to understand and decide where a certain kind of business could be successful or not.

### 1.3 Interested Stakeholders

The interested stakeholders will be those who want to check different variables of success before starting a business in Berlin, whether it's a small business or housing/real estate business.

# 2. Data Description

I decided to use **neighborhood density**, **business density**, and **tourism** as the deciding variables in this project.

### 2.1 Data Acquisition

The data was acquired from different sources with different methods and in some cases extrapolated from the already existing data.

The **neighborhoods**, **boroughs** and **neighborhood density** data was scraped from the wiki page: https://en.wikipedia.org/wiki/Boroughs_and_neighborhoods_of_Berlin

**Tourism data** was acquired from:
https://www.statistik-berlin-brandenburg.de/webapi/jsf/tableView/tableView.xhtml

**Venues data** was acquired from the Foursquare API

**Geojson of neighborhoods** (Ortsteile) of Berlin city from:
https://common-data.carto.com/tables/berlin_ortsteile/public/map

### 2.2 Data Processing

The acquired data from the wiki page was used to create the column of the **neighborhoods**, **neighborhood density** and a column of the corresponding **borough** in the dataframe.

Since berlin has very big 12 boroughs i decided to take a more granular approach so the comparison can be between smaller neighborhoods(ortsteile), however i did not consider all of the neighborhoods in the city, but only the more populous half of them.

As for the **coordinates** of each neighborhood i tried to acquire them by scraping the geohack pages that were linked in each neighborhood wiki page, but unfortunately when I plotted them on the map they were inaccurate, so i tried with geocoder and also those were inaccurate, I finally resorted to calculating them myself using the geojson file that divides Berlin neighborhoods.

I used the geojson file to divide Berlin into its neighborhoods in order to create the choropleth map of the population density.

**Tourism data** was consisting of **number of guests** and **number of overnight stay** for each **borough**, out of these 2 variables i created the **Tourism metric** variable and attached it to the dataframe.
Since the tourism metric was for the boroughs and not neighborhoods, for each neighborhood I assigned the tourism metric corresponding to its borough.

**Venues data** was used to create a dataframe with columns of **venue, venue coordinates**, and **venue category** plus the columns of neighborhood and neighborhood coordinates, then it was used to calculate the **most common venues** in each neighborhood, and eventually to cluster the data using K-means algorithm.
In addition this data was used to calculate the **business density** in each neighborhood and out of it the **business density metric** that was created and attached to the dataframe.

Certain measures were taken in order to ensure that the data used was valid and consistent:
When a call is made to fetch venues from Foursquare, the circle of coverage with radius 1000m from a certain neighborhood's coordinates, could cover an area of a nearby neighborhood, this could result in assigning an 'outsider' venue to this neighborhood, this is accounted for and corrected.