Analyzing tech hub real estate investments

Table of contents

1. Introduction

1.1. Background

The city of Berlin, Germany is by many considered as one of the most startup hubs of Europe. In order to leverage the economic potential connected to startups the government body is supporting the development of the necessary infrastructure. In order to strengthen the synergies between science and economics the city is planning to invest in the renovation of a former animal testing laboratory, the "Mäusebunker" and convert it to a med tech hub for startups. The hub will work as an incubator

supporting the young companies with equipped labs as well as a multitude of different business services. As this is a costly investment (approx. 90 million €) a lot of different aspects need to be considered to ensure the likelihood of the success of this investment.

Therefore this analysis especially interesting for future investors in order to get a better understanding of the risk they're taking with their investment.

This work will be part of a master thesis in architecture. After the analysis of the business needs, the infrastructure, the location, the building itself and other general conditions a design for the new usage and reconstruction of the building will be developed.

1.2. Problem

The following work is going to compare the neighborhood and local infrastructure of already successful tech hubs within the city of Berlin with the surroundings of the "Mäusebunker". This is part of the analysis deciding whether the investment in the conversion of the building is worth the financial risk involved in such a big investment. A greater similarity of the surrounding infrastructure with existing hubs is considered a positive indicator for a possible fruitful investment.

1.3. Interest

Considering various data points prior to an investment is essential. Therefore every direct and indirect aspects need to be considered. Analyzing the surrounding infrastructure is an essential part of the development of real estate. Supporting the analysis with data improves the previous manual work and gives a more objective indication if the investment can be successful.

2. Data

2.1. Data sources

In order to segment the city of Berlin into different areas there were multiple options. Berlin has 12 districts consisting of 96 local areas. These can be further divided into 192 postal code areas. Each of these postal code areas have a approx. size of 4 km². To get the finest granularity this work is using postal code areas are reference. A data set containing all German postal codes as well as additional information such as the state, province, community, latitude and longitude can be found in the this dataset of git hub. Data of the surrounding venues and infrastructure can be found on foursquare accessing its explore endpoint of the places API.

2.2. Data cleaning

Downloading the data is needs to be cleaned of unnecessary information regarding postal codes outside of the city of Berlin as well as Information considered irrelevant e.g. country, province and province code. To select the relevant postal codes all data which are not from the state of Berlin (Berlin is a city and a sole state within Germany at the same time) is dropped.

After singling out the relevant postal codes with their corresponding latitude and longitude the venue explore endpoint needs to be requested. In order to get enough relevant data the radius needed for the call of the endpoint is set to 2,000 m as the average size of a postal code area is approx. 4 km². Additionally the number of results is set to 50 per call as this is the maximum number of results per call for free usage of the API. This creates a set of 9,600 samples.

We're using defined reference points based on Berlin's future list containing already successful tech hubs within the city of berlin.

As there was put a lot of effort in selecting the data sources the cleansing of the data took only a minimal amount of time.

3. Methodology

After extracting the relevant postal codes from [the data set](#) of the git hub user Zauberware, the coordinates of latitude and longitude were used to get data about venues surrounding the coordinates. As mentioned earlier the radius is set to 2,000m as the average postal code area is approx.. 4km². After creating a data frame containing every relevant venue one hot encoding is used and afterwards the corresponding postal codes are added back to the frame. These are than grouped per row taking the mean of the frequency of occurrence of each category in consideration. The category of the 50 venues per postal code are than sorted to create a relevant basis for the comparison and clustering.

To define the number of clusters we're using an elbow plot which shows at what value of k, the distance between the mean of a cluster and the other data points in the cluster is at its lowest.

The number of clusters is set to 23 in order to create a great enough spectrum for the comparison. The used k-means clustering is a method of vector quantization that is partitioning observations (in this case the postal codes) into clusters in which each postal code belongs to the cluster with the nearest mean.

With the help of folium these results are visualized in order to allow an easy visual verification of the clustering results.
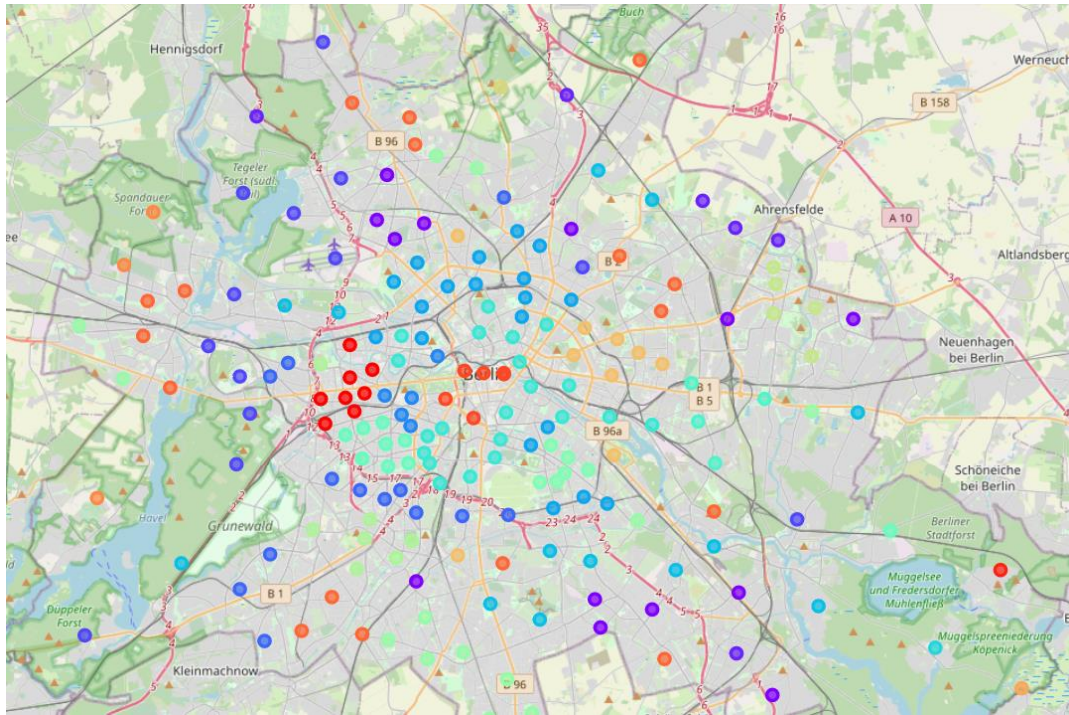
Clustered map of the city of Berlin
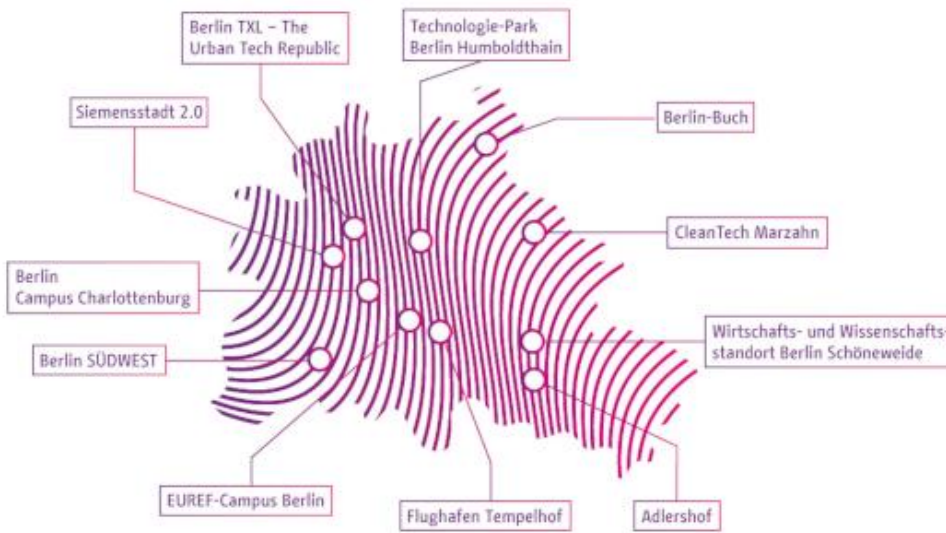
In the next step the postal codes and thereby the relevant reference clusters need to be identified.

These are based on the Berlin's future list of the ministry of economic and energy the following:

Object	Postal code	Cluster
Charlottenburger Innovations-Centrum – CHIC	10625	0
Existenzgründerzentrum Technische Dienstleistungen an der HTW Berlin	10318	10
Gründerinnenzentrum WeiberWirtschaft	10115	9
Gründerzentrum der HWR Berlin	13629	7
Gründungszentrum „GründerScout" der Beuth Hochschule für Technik Berlin	13409	17
Innovations- und GründerZentrum Berlin-Adlershof – IGZ	12489	1
Innovations- und Gründerzentrum – IGZ – Campus Berlin-Buch	13125	20

Innovationspark Wuhlheide	12555	3
PHÖNIX Gründerzentrum Am Borsigturm	13507	3
Technologie- und Gründerzentrum Spreeknie – TGS	12459	20
Unternehmerinnen und Gründerinnen Zentrum in Charlottenburg-Wilmersdorf – UCW	10713	11
Wissenschafts- und Technologiepark Berlin-Adlershof	12489	1
Mäusebunker	12203	14

4. Results

The cluster containing the "Mäusebunker" is number 14. None of the other tech hub we're using as reference for an supporting infrastructure is in the same cluster as the "Mäusebunker". Thereby we can see that the surrounding venues of the "Mäusebunker" and the other objects are not similar to each other.

5. Discussion

Using "only" 50 venues per 4 km² could be not enough data points in a highly dense city like Berlin. This can result in a not optimal data basis for the research. In order to improve the results a foursquare plan allowing to access more data could be useful. Additionally are the surrounding venues only part of the indication whether a real estate should be developed. The used reference objects could have a suboptimal surrounding infrastructure and be successful regardless of that fact. Thereby a clear recommendation cannot be given based on the taken analysis. It can only be part of a wider evaluation.

6. Conclusion

Comparing the infrastructure surrounding real estate objects is an essential part of every investment evaluation and can support the process of figuring out if the development of a real estate is worth it. Supporting infrastructure can be important when it comes to logistics, the way to work for employees, local supply of goods and multiple other reasons. Using existing objects as a reference is quite a stretch as these objects could be successful regardless of the surrounding. It could be more reasonable using an expert opinion to define which venues are most supportive for a venture like the researched one. Thereby the results will be taken into consideration and will be followed by a qualitative research to support the findings.