Berlin – The Pulse of a City

Tobias Gerken

May 12, 2019

1 Introduction

According to U.S. Small Business Association [1], approximately 50% of new businesses fail within five years of opening. Given the high risk and challenging retail environment, finding the right target demographic and business environment is therefore key. At the same time, cities are becoming increasingly dynamic and diverse which provides ample opportunities for business segmentation.

Combining data-science approaches with the increasing availability of real time consumer data and open data from city governments can provide valuable insight that can businesses help attract and retain customers.

1.1 Business question

Berlin, the capital of Germany, is home to approximately 3.6 million people and widely recognized to be one of Europe's most dynamic cities. Using the city of Berlin as example, we segment its districts by their population demographics and analyze what type of businesses thrive is the respective neighborhoods. The *trending* feature of the Foursquare API, we analyze locations and neighborhoods in Berlin that are trending within a 24-h window letting us experience the *Pulse of the City*.

In detail we ask:

- Which neighborhoods have the highest density of trending venues?
- Does the age distribution of district residents have a discernible impact on venues density and type?
- Can we segment neighborhoods based on the type of business venues they attract and is there a relationship to the demographics Berlin's districts?

2 Data

Berlin is a city state divided into 12 districts, each with populations between 200,000 and 400,000 (Figure1).

The Berlin state government provides demographics data for Berlin through its Open Data Portal [2]. Geographic shape files of the Berlin districts [3] and postcode [4] areas are available through the *Technologie Stiftung Berlin* (Berlin Technology Foundation).

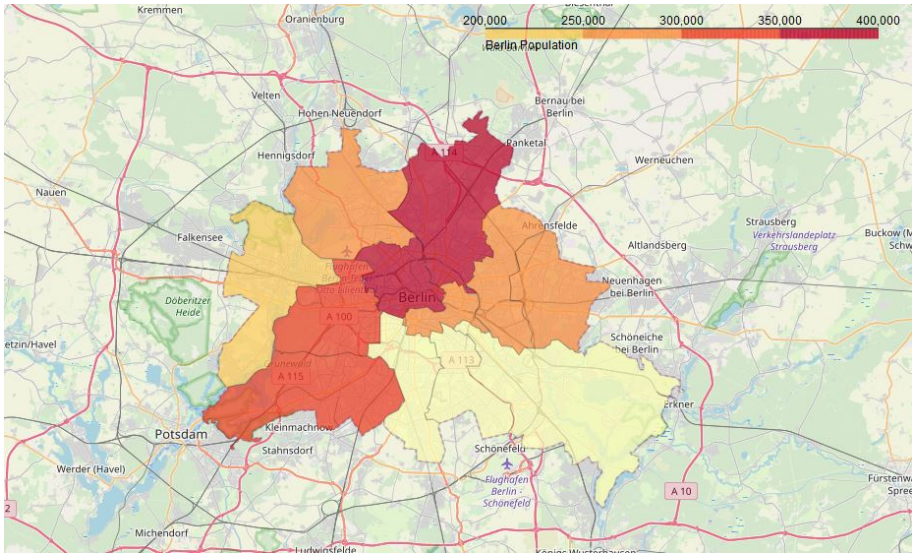

Figure 1: Berlin is home to 3.6 million people, who live in 12 districts with 200,000 to 400,000 inhabitants each.

Foursquare's API [5] allows to query business venue data based on geographical coordinates. Using the *explore* functionality, we receive information about nearby venues, including their postcode and geographic coordinates, which allow them to be linked to Berlin's demographic data. The *trending venues* functionality, allows us to find types of business venues and associated geographic areas that are popular during the course of a day.

2.1 Approach

Demographic data at the city neighborhood-level (*Ortsteil*) were downloaded in CSV-format and aggregated to district level. The demographic data is provided in age groups of 5 years. We suppose that similar age groups have similar consuption patterns and aggregated the data to age groups 0-15, 15-30, 30-50, 60-65, and over 60 years old (Table 1), approximately representing children, young adult, middle age, older, and senior consumers.

2.1.1 Retrieving Venues

To capture variation in business venues, we execute a gridded sampling approach (Figure 2), where Berlin is covered with an equidistant grid of 1 km length for which Foursquare search queries are being executed.

Table 1: There is considerable diversity in age group distribution between Berlin's districts. For example Friedrichshain-Kreuzberg has the larges share of middle age consumers, while Steglitz-Zehlendorf has a much older population.

AgeGroup	0-15	15-30	30-50	50-65	>65
District					
Charlottenburg-Wilmersdorf	0.11	0.17	0.29	0.23	0.24
Friedrichshain-Kreuzberg	0.14	0.23	0.42	0.16	0.11
Lichtenberg	0.14	0.19	0.3	0.21	0.21
Marzahn-Hellersdorf	0.15	0.17	0.27	0.27	0.2
Mitte	0.14	0.25	0.35	0.18	0.14
Neukölln	0.14	0.21	0.32	0.19	0.19
Pankow	0.16	0.17	0.39	0.19	0.16
Reinickendorf	0.14	0.18	0.26	0.22	0.25
Spandau	0.15	0.19	0.27	0.22	0.23
Steglitz-Zehlendorf	0.13	0.16	0.26	0.22	0.27
Tempelhof-Schöneberg	0.13	0.18	0.29	0.22	0.22
Treptow-Köpenick	0.13	0.16	0.29	0.22	0.24

Using the Foursquare API's *explore* functionality, we query recommended venues within a radius of 750 m around each grid-point. Complete coverage of Berlin is ensured through overlap of the sampling points. Subsequently, duplicate venues are removed. A total of unique 9777 venues were returned. We extracted venue names, venue location, venue address, venue-category and the corresponding city and postcode for each venue. For 1429 venues, for which no postcode was recorded in the data-set, the postcode was retrieved by geographic location using a shapefile of Berlin's postcodes*PLZ*. Finally, we removed any venues for which no postcode could be retrieved, or which had a venue city other then Berlin. This resulted in a final dataset of 9311 venues.

Figure 3 gives an approximation of the venue density in Berlin. However, there is a caveat. The Foursquare API limits the number of returned venues to 50 per search and will only return recommended venues, as such this constitutes a lower bound on the total number of venues in Berlin.

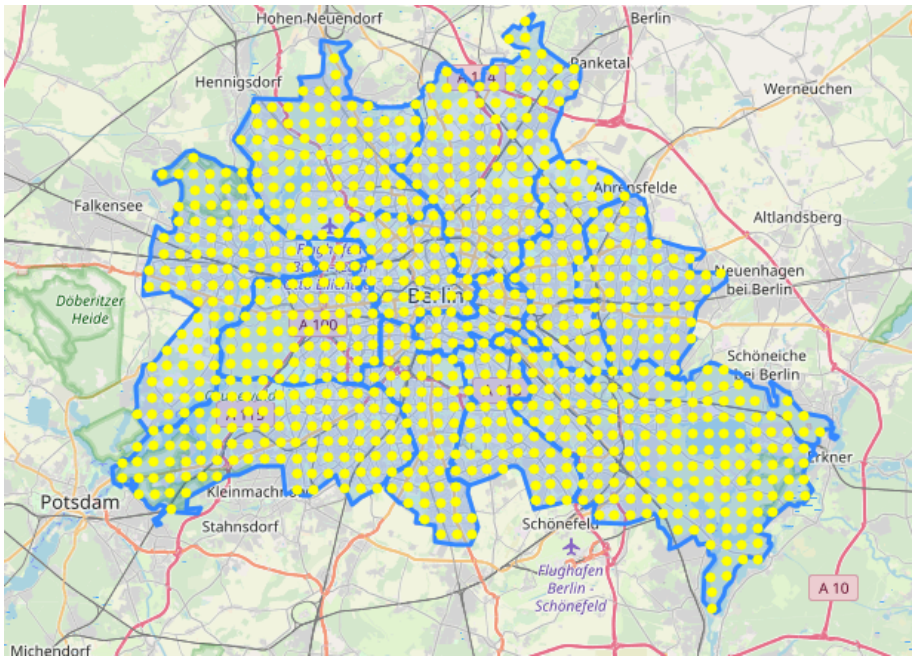

Figure 2: The Foursquare API is queried for a large number of locations in Berlin, which are arranged on a regular grid.

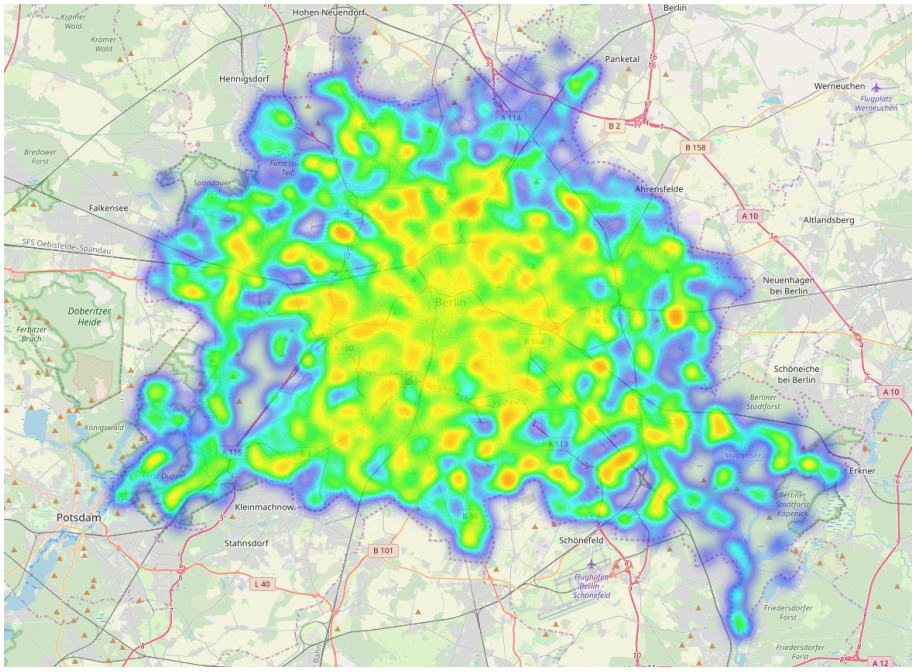

Figure 3: A total number of 9311 were found almost all over Berlin. The density of venues is notably higher in the center.

Venues were aggregated by district and by postcode for further analysis. Using the demographics data, we also calculated the venue density as a function of population.

2.1.2 K-Means Clustering of Neighborhoods

To learn more about individual neighborhoods below the district level, we use Berlin's postcode areas as proxy. There are 453 unique venue categories in the data-set. These can be used to learn more about the profile of individual neighborhoods using an unsupervised clustering approach.

We chose K-Means clustering, one of the most common and computationally cheap clustering algorithms for simplicity. K-Means clustering requires *a-priori* specification of the number of clusters. Given the high number of categories, which leads to sparseness of the data-set, and the fact that we expect neighborhood clusters to not fully separate, we set the number of clusters $K = 4$. The K-Means clustering is executed using scikit-learn [6]. We use the venue categories as features for the clustering. Because venue-category is a categorical variable, we create a dummy-variable for each venue category and subsequently normalize the features in each neighborhood to one.

2.1.3 Trending Venues

We used IBM Watson Studio to deploy a script that used the Foursquare API's trending functionality to generate a list of trending-venues. The code was run for using the

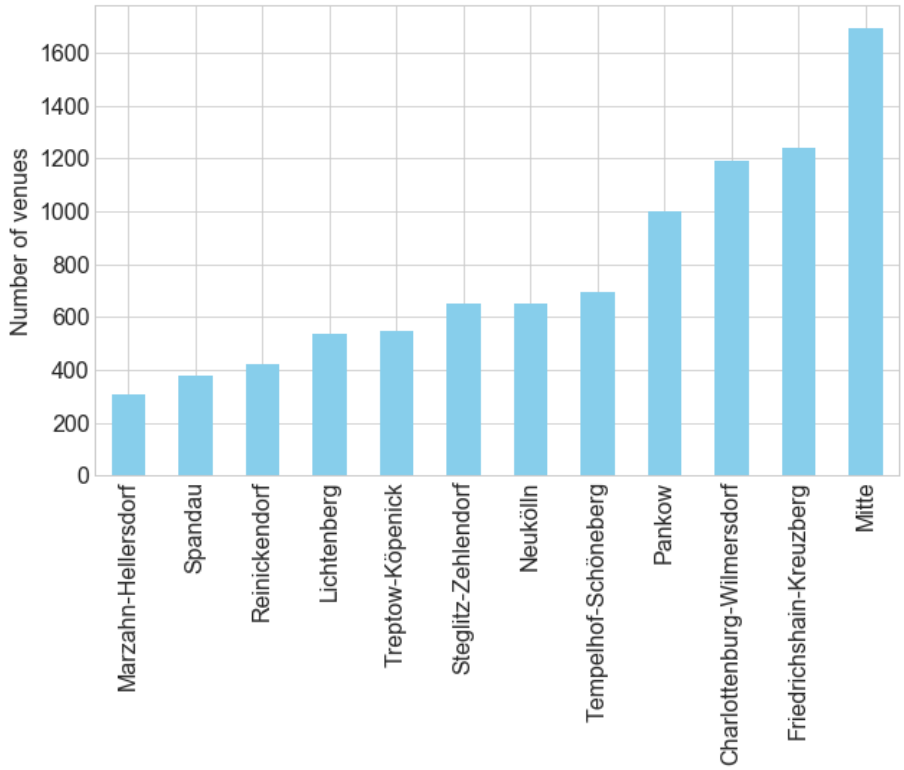

Figure 4: The highest number of venues were found in Berlin Mitte (> 1600) – the city center – and the lowest number of venues were found in Marzahn-Hellersdorf (~ 300) – a comparatively poor district.

previously noted gridded sampling approach. Returned results were saved to file for later analysis.

The code was run hourly from Thursday May–2–2019 00:00 to Monday May–7–2019 4:00 Berlin time providing information about trending venues for several workdays and a weekend.

2.1.4 Tools

Data wrangling is performed in Python using Pandas [7] and Shapely [10] (for geospatial calculations). Visualizations are done using Pyhthon's matplotlib [8], seaborn [9] and Folium [11] libraries.

3 Results

It is the goal of this report to investigate to detect to *Pulse of the City of Berlin*. To do so we first look into the relationship between demographics and venues found on

Foursquare, then investigate further what characterizes neighborhoods in Berlin, and finally follow trending venues over time.

3.1 Demographics

Figure 4 displays the total number of venues for each district. Calculating the number of venues per 1000 inhabitants for each district, we can see that the venue density (Fig. 5) closely resembles the total number of venues.

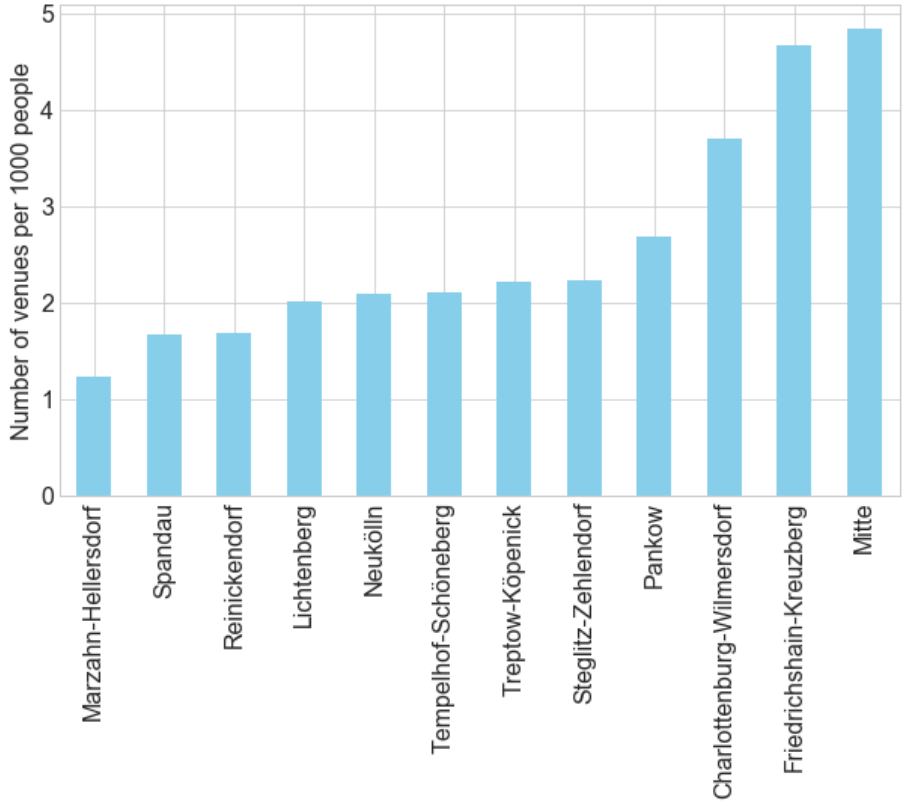

Figure 5: The highest number of venues per population were found in Berlin Mitte ($\sim$ 5 venues per 1000 people) – the city center – and the lowest number of venues were found in Marzahn-Hellersdorf ($\sim$ 1 venue per 1000 people) – a comparatively poor district.

Interestingly, there is a clear relationship between demography and venue density. We can see that districts that have a higher share of young and middle age adults, have a higher venue density, while districts with a higher share of people older than 55, have fewer venues per people (Fig. 6). This may either indicate that younger people actively seek more dynamic living environments or that commercial venues are are more likely to target a younger, wealthier, and more active demographic.

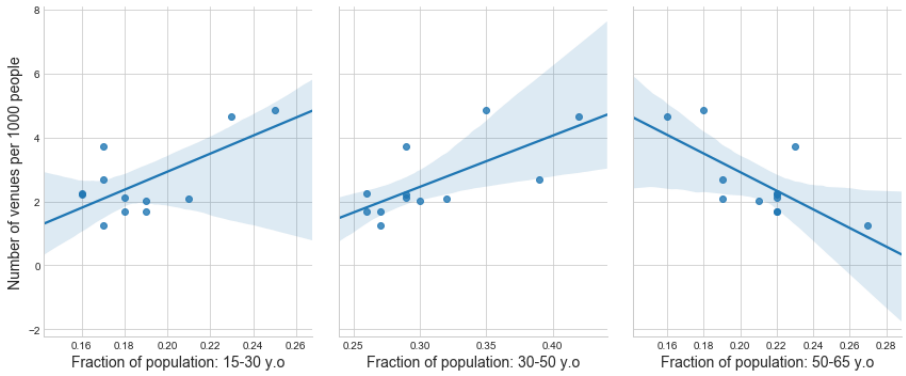

Figure 6: There is a clear relationship between demographics and venue density.

3.2 Clustering of Neighborhoods

To further investigate the make-up of Berlin, we analyze the venue-data on the neighborhood level. We find that centrally located postcodes have more commercial venues than peripheral postcodes (Fig. 7).

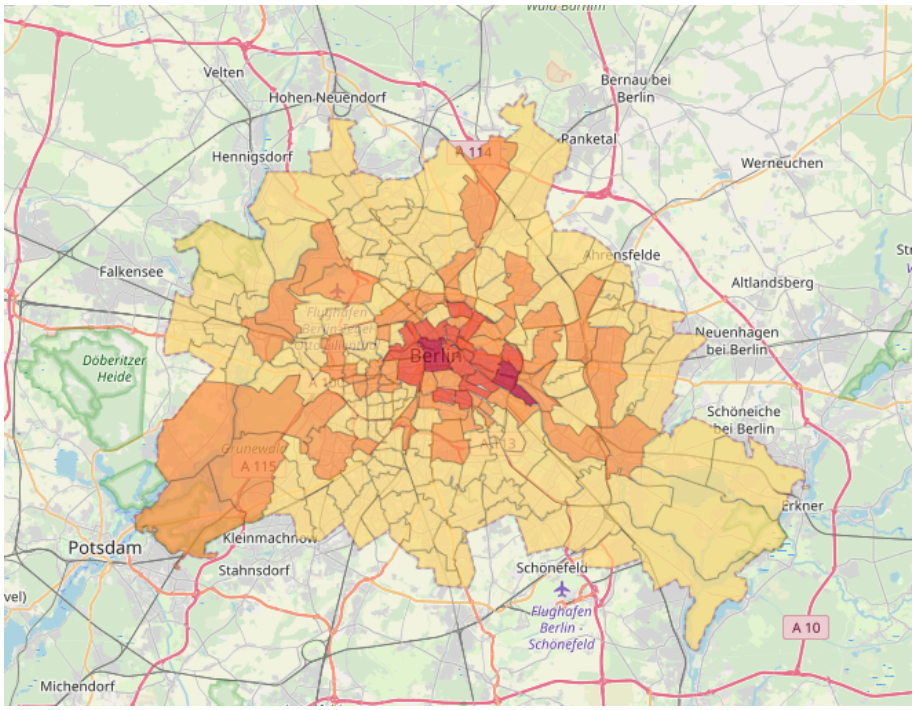

Figure 7: The highest number of venues is concentrated near the city center.

The clustering results in a 4 clusters. Based on the dominant venues in each cluster, we can investigate the unique features of each cluster.

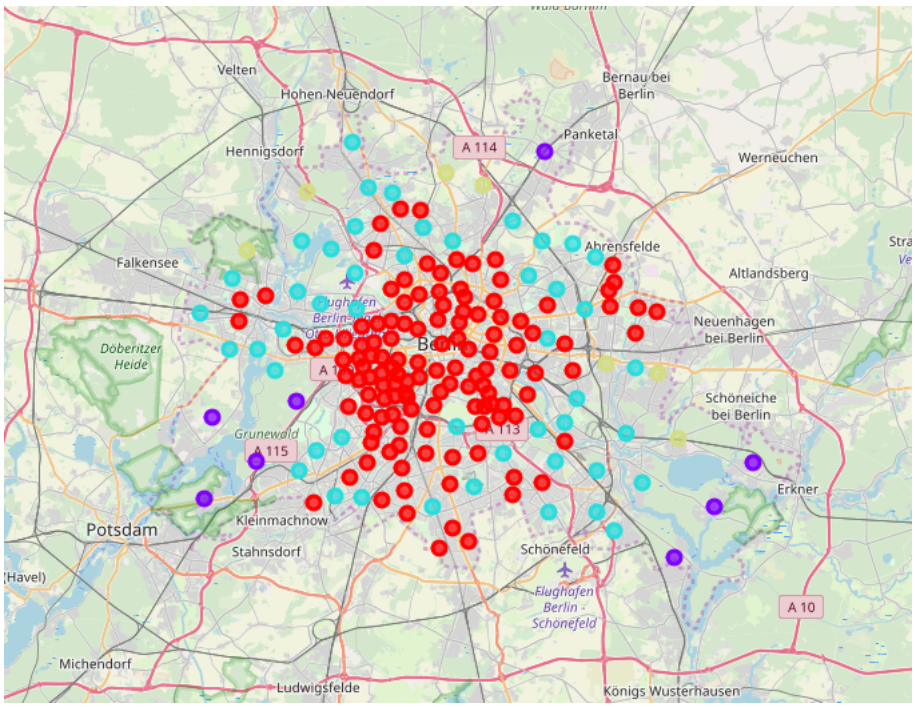

Figure 8: The clustering shows a central cluster which is surrounded by several peripheral clusters.

First cluster - The bustling center: This cluster (indicated in red in Fig. 8) is located in the center of Berlin. It is dominated by hotels, restaurants, cafes, indicating lots of commercial activity. Hotels indicate tourism for business and pleasure, while cafe's and restaurants contribute to a vibrant atmosphere.

Second cluster -People live here . . . : This second urban cluster (cyan) is more residential compared to the first urban center. Supermarkes and transit venues are mixed with other commercial venues. These neighborhoods are also located further away from the city center, but still densely built up.

Third cluster - Respite from the city: This peripheral cluster (purple) is also residential with supermarkets and transit venues. At the same time it also features hotels, parks, lakes, marinas, and even a nudist-beach, indicating a recreational quality to be enjoyed by tourists and locals alike. Looking at the map we can also see that lakes, rivers, and forests are nearby confirming our analysis.

Fourth cluster - The city's edge: This peripheral cluster (green) lacks the attractions of the third cluster. Venues such as parks, fairs, farms indicate lower density. Hotels, forests and lakes indicating recreational activities are largely missing.

3.3 Trending Venues

A total of 350 trending venues were identified between Thursday May 2, 2019 00:00 and Tuesday May 7, 2019 06:00 Berlin time (Fig. 9). Most venues were located in the city center, but there are also a few locations outside the city center that trended during this time.

It can clearly be seen that the number of trending venues follows a clear diurnal patter with, most activity taking place during business hours and the evenings, indicating that Foursquare API's trending feature is indeed capable of detecting the ebb and flow of activity that makes up the Pulse of Berlin (Fig. 10).

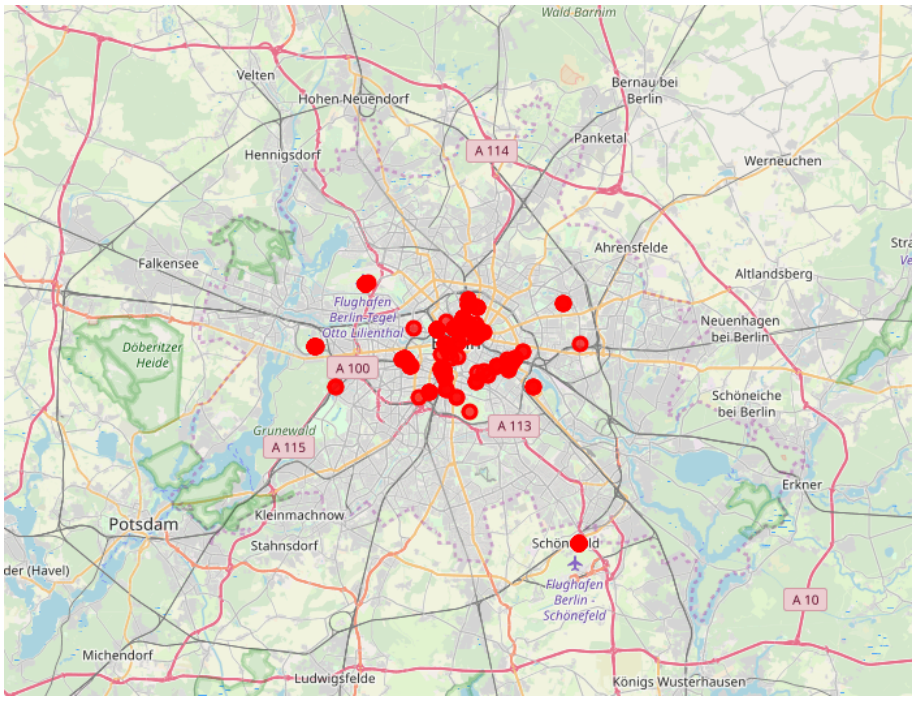

Figure 9: A total of 350 trending venues were identified between Thursday May 2, 2019 00:00 and Tuesday May 7, 2019 06:00 Berlin time.

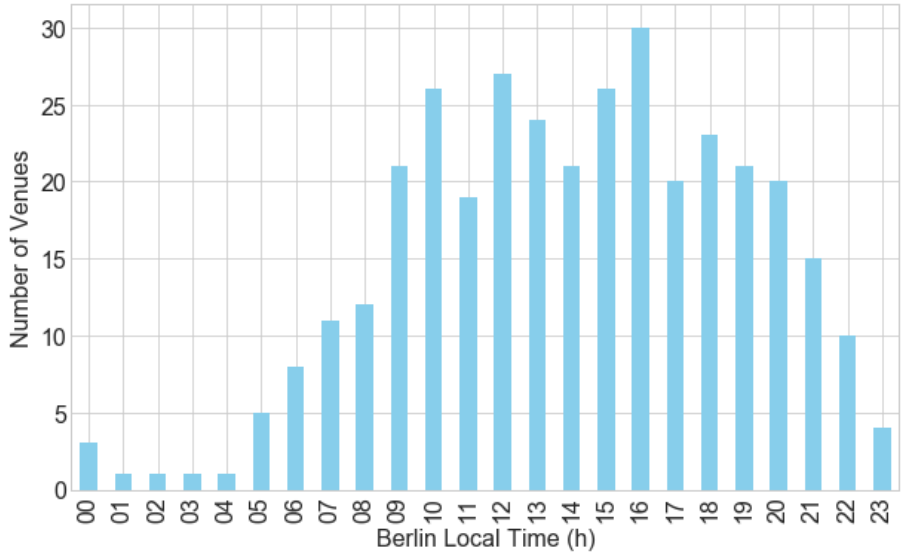

Figure 10: The number of trending venues follows a clear diurnal pattern.

In total, 51 distinct venue categories as defined by Fousquare were returned as trending. To facilitate interpretation, these venue types were manually grouped into similar categories. This was also necessary, since some venue categories were nearly identical to each other (e.g. *Airports* and *Airport Terminals*). Notably, transit hubs such as airports and train stations were among the most popular venues 11). Similarly, events and outdoor spaces were also very popular. Surprisingly, bars and nightclub venues featured less heavily in the data-set, despite the data-set covering a weekend period.

The Foursquare API returned a few "surprising" venues. For example, two advertising agencies were featured prominently in the trending data-set. Similarly, the API identified a prison and an assisted living facility as trending during the time period. However, closer inspection revealed that these are due to mis-classification. The prison Hohenshönhausen is a museum and memorial dedicated to victims of the East-German secret police and the assisted living facility was identified as *rp19wg*, which was the hash-tag of the re:publica 2019, a conference/event about digital media and society held in Berlin on May 6-9, 2019.

The re:publica 2019 event can also clearly be detected in the time-series of trending venues (Fig. 12). While some venue categories, such as transporation hubs were popular during all times, except the very early morning hours, other venue types, such as outdoor and event venues were mostly popular during the weekend.

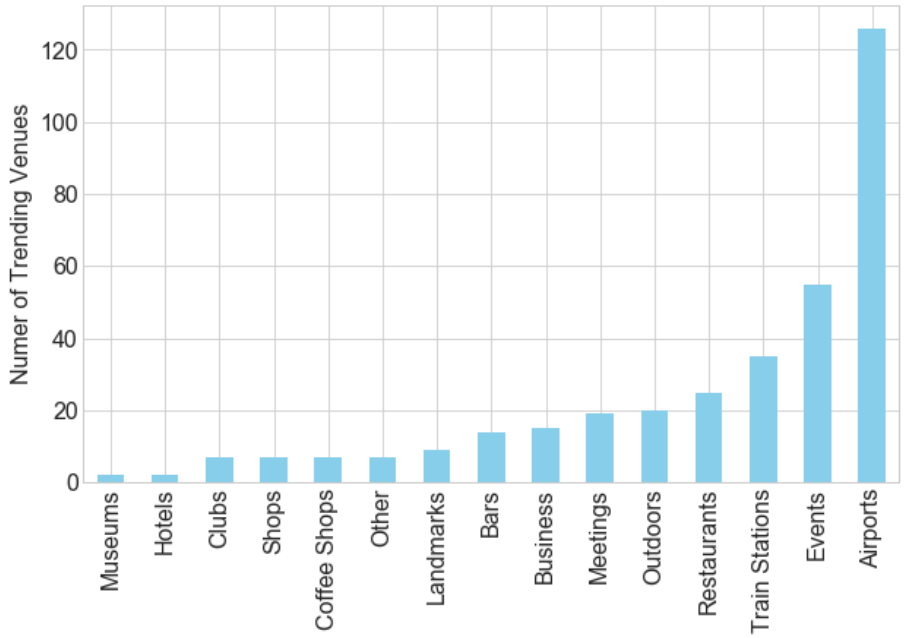

Figure 11: Transit hubs, event spaces were the most popular venues.

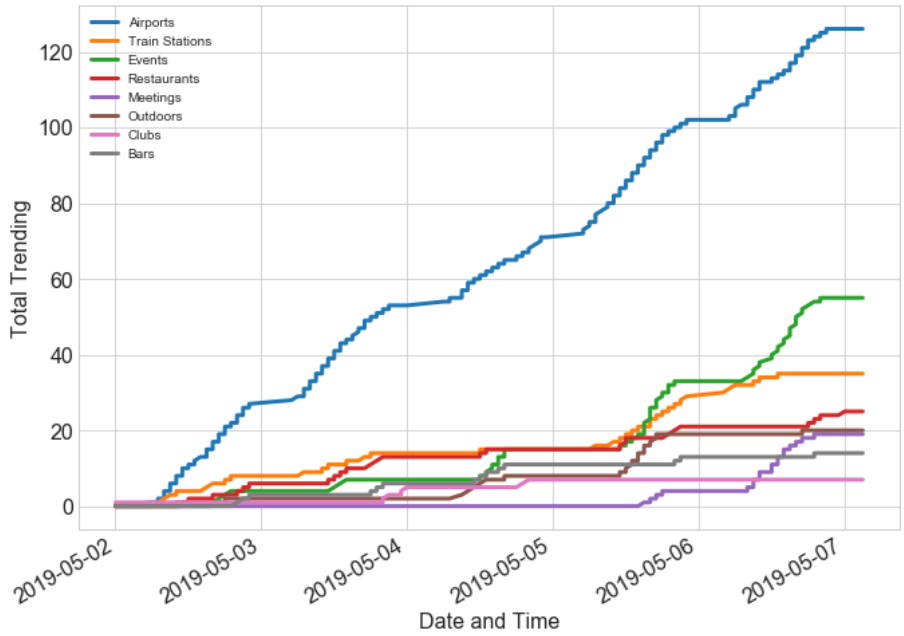

Figure 12: The popularity of venues varies with time.

4 Discussion

This work applied three data-driven approaches to explore the *Pulse of Berlin*.

First, we showed that district level demographics and business activity were closely related, with younger and possibly more affluent people living near centers of commercial activity. At the same time, is is unclear whether this is a causal relation, or whether for example both businesses and younger people respond to similar attractors, such as urban centers or spaces of cultural and touristic value.

Second, aggregated to the neighborhood level, Foursquare venue data was successfully used to categorize Berlin's neighborhoods into different clusters. The results showed that Berlin could be separated into urban core and peripheral neighborhoods, which likely serve different roles as living and recreation spaces.

Lastly, we explored Berlin as a dynamic space in time with its unique ebb and flow of activity. Using Foursquare's data it was possible to identify a large public event that took place in Berlin - the re:publica 2019 digital conference. Similarly, we could observe that transit and airports never sleep and are the beating heart of the city.

At the same time, we identified several limitations in this work, which are necessary to be discussed. While there is an association of younger demographics with the density of business venues, it is unwise to infer causation from this without further in-depth analysis. Despite the wealth of data available from Foursquare and other open data sources, not all necessary data to infer causation may be available. Similarly, Foursquare's data may not always be reliable as seen by the mis-attribution of the re:publica hashtag to a nursing facility. In general Foursquare's data may skew towards business venues that do not adequately capture all portions of life in a city. Without knowing how such data is aggregated caution should be taken to assume that the data-set is inclusive of all populations. For example, the re:publica event, frequented by digital natives featured heavily in the data-set of trending venues. Similarly, two advertising agencies – presumably with media-savvy employees – were also featured in the data-set, which seems like a stark over-representation compared to the total number of businesses and the role of advertising in the economy.

5 Conclusions

This analysis present a first exploration to use a dynamic data-set to capture the hustle and bustle of a dynamic city – Berlin. Based on the data used, we clearly demonstrated the utility of location data in combination with other data-sources to learn more about commercial activity at a surprisingly high temporal and spatial granularity, which could be further exploited to aid in business decisions. For example activity data of popular venues could be used for decision on marketing of opening new venues in popular spots.

References

[1] U.S. Small Business Association: Do economic or industry factors affect business survival? 2012
https://www.sba.gov/sites/default/files/Business-Survival.pdf,

[2] Amt für Statistik Berlin-Brandenburg: Einwohnerinnen und Einwohner in den Ortsteilen Berlins am 30.06.2016, 2016
https://daten.berlin.de/datensaetze/einwohnerinnen-und-einwohner-den-ortsteilen-berlins-am-30062016,

[3] Technologiestiftung Berlin: Die Bezirksgrenzen der 12 Berliner Bezirke, 2017
https://data.technologiestiftung-berlin.de/dataset/bezirksgrenzen

[4] Technologiestiftung Berlin: PLZ – Postleitzahlgebiete Berlins, 2015
https://data.technologiestiftung-berlin.de/dataset/plz

[5] Foursquare Labs Inc: Places API, https://https://developer.foursquare.com/docs, Last Access: 4/28/2019

[6] scikit-learn, https://scikit-learn.org/stable/, Last Access: 5/11/2019

[7] Python Data Analysis Library, https://pandas.pydata.org/, Last Access: 5/11/2019

[8] matplotlib, https://matplotlib.org/, Last Access: 5/11/2019

[9] seaborn, https://seaborn.pydata.org/, Last Access: 5/11/2019

[10] Shapely, https://pypi.org/project/Shapely/, Last Access: 5/11/2019

[11] Folium 0.8.3, https://python-visualization.github.io/folium/, Last Access: 5/11/2019