Research and Innovation: Open Data

Martí Municoy, Alba Gordó, Jan-Hendrik Niemann
and Andreas Radke

November 17, 2017

Abstract

The aim of this project is to find and visualize the events posted on `Meetup.com` in different cities in order to analyze the density of planned activities and to make a social study regarding most usual activities and interests. One aspect may concern different acitivities in several cities and countries. The data shall be obtained via the Meetup-API, Google Maps and a web scraping at `Wikipedia.org`. The most important data source for this project is the Meetup-API where we have to request the events with different Python packages like *requests*. Our second task is to visualize the locations of the events on a given city map obtained by `googlemaps.com` or to plot different stats regarding the city events in useful charts.

Contents

1 Summary and Goals

The human being is social. Every day there are millions of events. Thousands of them are posted at the webpage `Meetup.com`. It is a website where people can seek and find other people with the same interest. It helps to organize events and makes them public.

This is the reason why we chose `Meetup.com` to analyze the behavior and the preferences of cities and its inhabitants all over the world. Almost all `Meetup.com` activities contain information about their location, expressed in terms of latitude and longitude. Thus, by using the Meetup-API, we can request and extract a dataset "latitude-longitude" for all the available `Meetup.com` events. Then, we can use this dataset to plot all these located activities on a `GoogleMaps.com` map. Our aim here is to create and visualize a density map, a so-called heatmap, where one can see all the activities for a selected city.

Furthermore, the Meetup-API allows the user to get activities arranged by category. We want to use this tool to consider separately different types of events like "Food & Drink" or "Sports" and study their relative predominance in different cities. By combining this information with the geographical and/or political structure of a city, one may also identify trending districts and neighborhoods in a city. To do so we make use of GeoJSON-files which can contain the polygonal shape of a district and its name. There are several websites where one can get GeoJSON data (see Section A) or create own data. By using the information from these websites, a GeoJSON dataset is created with all the coordinates of the districts for all the studied cities. This constitutes the third data source that is used in this work.

A fourth data source is `Wikipedia.org` where a web scraping script is applied on to extract some useful population data such as the population number, the population density or the land area of each district. With all this information, one can use the following factor to evaluate the social activity per district. This factor is given as

$$\text{Social activity} = \frac{\#\text{ events}}{\text{population}}.$$

Once downloaded, we can access our data offline. This is due to the fact that we store it locally by using csv-files (comma-separated files).

This report is structured in 6 sections as follows. Firstly, in Section 2 we discuss the data life-cycle of our data. In Section 3, we present both, the third-party tools that we used and the custom methods that we developed on our own, to achieve this project. They mainly are developed by using

Python. Afterwards, we present our results in Section 4 and, in Section 5, we make a legal and ethical disquisition. Finally, in Section 6 we discuss the limitations of our work and make proposals for future work.

2 Data Life Cycle

In the following section we have a closer look at the terms *data life cycle* and *data life cycle management*.

Data life cycle is a term which tries to describe the life of data and information. Since there is no unique definition of a data life cycle, we define the folloing six stages: creation, storage, use, sharing, archiving and destruction [8], [10].

1. **Creation**: Data is created. It can be created as a structured or unstructured set of data, can be obtained by collection, measurements, generated or gathered. We create our data by using the Meetup-API. By performing a web scraping at `Wikipedia.org` we gather the number of population, the population density and the land area for different cities and their districts. In this way, we request and create the data which is going to be needed for this project.

2. **Storage**: Once the data is created, it has to be stored. Depending on where we want to apply this data for, there are different and optimal kinds of storage solutions. However, this is not subject of this report. In our case, both, the data gathered from `Meetup.com` and `Wikipedia.org`, are saved as csv-files on the hard drive. As a version control system we used Git and additionally stored our data on `github.com` for better collaboration. All our code can be viewed there [6].

3. **Use**: Data can be viewed, modified, analyzed, corrected, interpreted, visualized and joined. We use our data to visualize the social activity of cities. We join our data from different sources and visualize it on a given map which can be seen as data as well. This data visualization and its further treatment can be used to obtain new data such as the measurement of the social activity per district introduced in section 1.

4. **Sharing**: Data can be shared. There are several ways of sharing by using different file formats, applications or operating environments. In this work, we used csv-files and GeoJSON-files to share data from one method to another. For instance, the district coordinates are read from

a GeoJSON-file, the Meetup events and the population data from a csv-file. Then, these are the files created by the scripts to share data along different methods.

5. **Archiving**: When the data leaves the active use, one should archive it in a suitable way. Since archiving is very similar to storaging, one can use the same methods. One difference may be that saved data needs to be compressed to save storage or protected to prevent unwanted changed. Therefore, the access is slightly more difficult than for data in active use. The file formats used in this project does not satisfy this statement. This is due to the fact that the file formats are chosen to ease the access to their data rather than to save storage space. If we want to keep this data for a long time, we should think on developing a method to compress this data.

6. **Destruction**: There is a last stage in the data life cycle. There are several reasons why one should destroy data. On one hand, the volume of archived data increases and one needs to save storage to store new data. On the other hand, data gets old and could be not needed anymore. Or, if it is not up to date, it can be replaced by newer versions. The second case is applicable for our work. During every single run of the program new csv-files with event data and new csv-files containing populations data are created. Especially the event data underlie a rapid change.

The term *data life cycle management* describes the process that helps to organize the flow of data throughout its life cycle.

3 Data Tools

Python Packages

The main work for our project is done in the Python programming language in version 3.6 [9]. This programming language is used to develop different packages. Each package is in charge of managing a specific dataset and it does a certain job. In this section, we are going to introduce the most important packages that we have developed and used to retrieve the data of this project.

Meetup Package

This package has the main purpose of retrieve the information of all the available events in a city. Its main module is called *mu_requests.py*. This

module makes use of the *Python 3.6 Requests* package [3] to submit queries to the Meetup API [5]. It is able to get information about all the available categories and events. Then, it can also parse the data coming from the Meetup client, whose format is expressed in JSON, and it can store it in a local drive by writing a custom csv file. For instance, all the events found in a city they are written in this way *./csv/name_of_the_city.csv*. This custom csv-file contains all the events of a city arranged by category.

Mapping Package

This is the package responsible of plotting all the located events on a map. To do this, first, it needs to read all the events data from the previous csv-file. Then, it uses the read coordinates to display events on a *Google Maps* map. This functionality uses the *gmaps* library [7] along with a *Jupyter notebook* [4] which allow us to display data on a map.

It has some tools that allows us manage which events we want to display. For example it allows us to plot only those events that are scheduled in a certain time range or those events which belong to a specific category. We can also control the way the script colors the event marks or their opacity.

Another interesting tool that this package offers is to plot the districts of a city. In this case, *GeoJSON*-files with the delimiting points for each district are required. If you want to plot them according to their population, population data for each district is also required. The script reads these files, parses the data and it can plot and paint the districts according to the number of events per capita that each district has.

Scraping Package

This package is involved on retrieving data from `Wikipedia.org`. Particularly, it searches for the pages belonging to the districts of a city to retrieve information about their population from one of their tables. This process is known as web scraping. To check for the appropriate tables, it uses *BeautifulSoup* Package that is an *HTML* parser.

Then, the script is designed to look for these tables in an intelligent way. So it is able to check different Wikipedia addresses to get the right information. Firstly, it looks for the English articles but, in some cases, best articles are written in other languages rather than English. So it is also able to check pages written in Spanish and German.

Once it gets the right population table it comes to the hardest part. It needs to move around the table to look for the proper cells. The functions that are used in the parsing process are able to work with tables, words and

numbers expressed in different formats. For instance it supports tables with joined cells or which arrange their information in different ways.

Plotting Package

The last developed Package is used to plot data from all these retrieved events. It is useful to view and compare the amount of events for each category or city. This script makes use of *pygal package* [1] to create mostly bar charts.

Data

As already noted we used the data sources that are summarized below:

- Meetup API: To get the events in the examined cities.
- Wikipedia: For obtaining data about the population density of the cities and their districts.
- GeoJSON Files: To visualize districts onto Google Maps. Files are obtained from various resources. For the detailed sources see Appendix A.
- Google Maps: As a general back-end for our visualization [2].

4 Results

To draw meaningful conclusions we created several types of visualization of the data. First and foremost we plot the open events onto a map - either as a point map, e.g Figure 1, or as a occurrences distribution map for each city district, e.g. Figure 3. Additionally numerous bar charts are created with the number of categories (absolute and per capita) per city, compare Figure 4, and vice versa.

In this chapter we will shortly discuss several cities and their interesting characteristics.

Barcelona

In Figure 1 one can see an example of Barcelona with open events plotted as points. The different colors represent different Meetup categories. One can see that there are slightly more green points which represent the Meetup

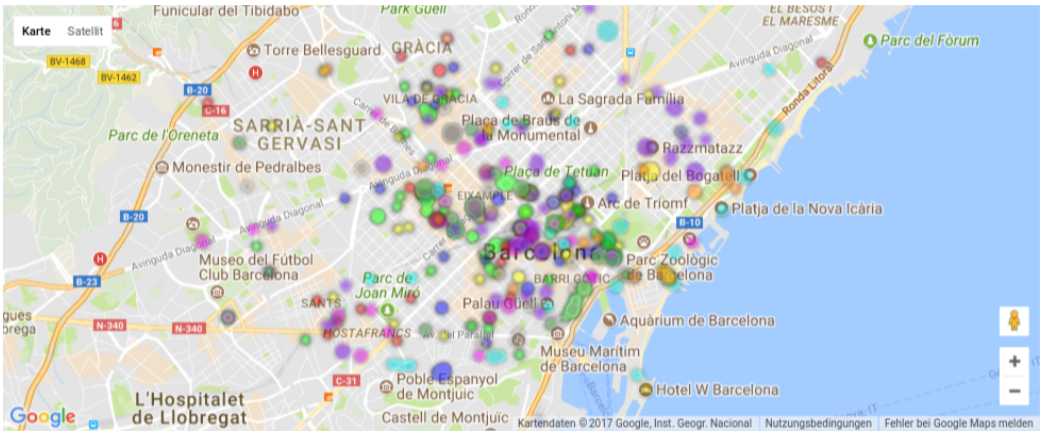
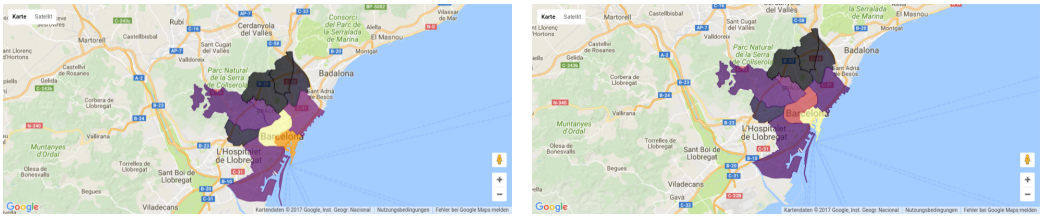


Figure 1: Snippet of Barcelona with Open Events

Figure 2: Snippet of Barcelona with only Language & Ethnic Identity Events

category *Language and Ethnic Identity*. A more specific view offers Figure 2 with only the points corresponding to this specific category. Furthermore the already mentioned chart 4 shows that the Language and Ethnic Identity events are indeed the most frequent category which may be an indicator for the diverse and multi-cultural Catalan capital.

The most activities are unsurprisingly in the populous Eixample district but per inhabitant there are more events in Ciutat Vella. Despite having a natural concentration in more centric areas for Barcelona it seems that



(a) Activity per District (b) Activity per Capita per District

Figure 3: Barcelona

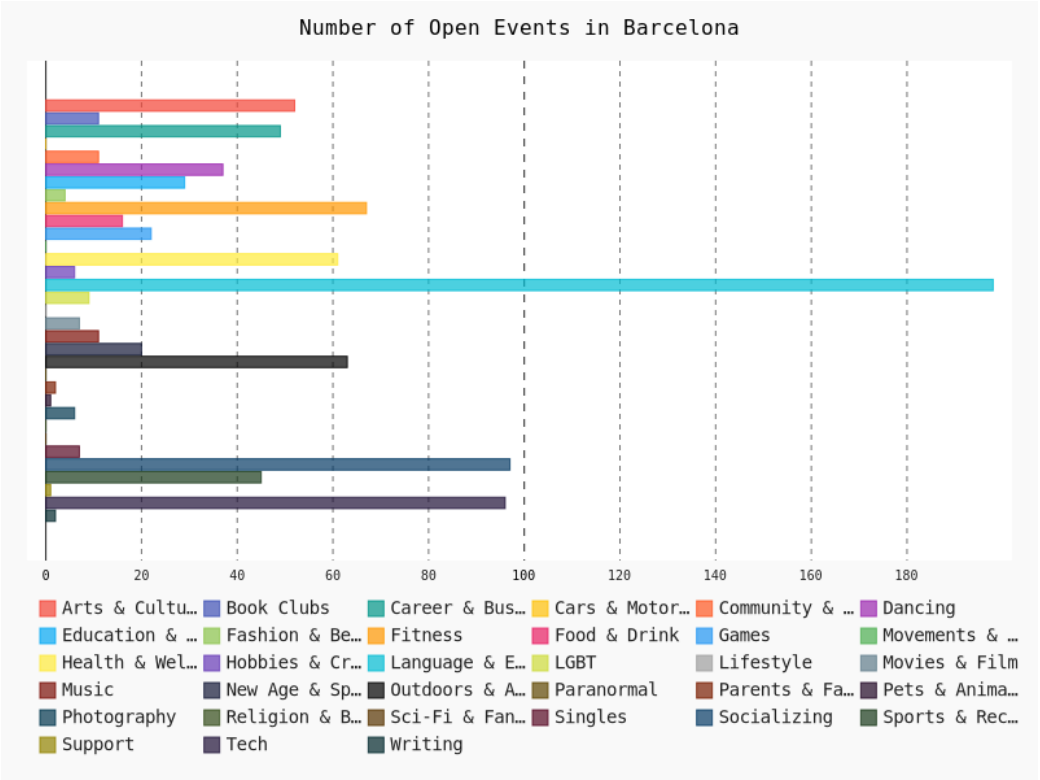


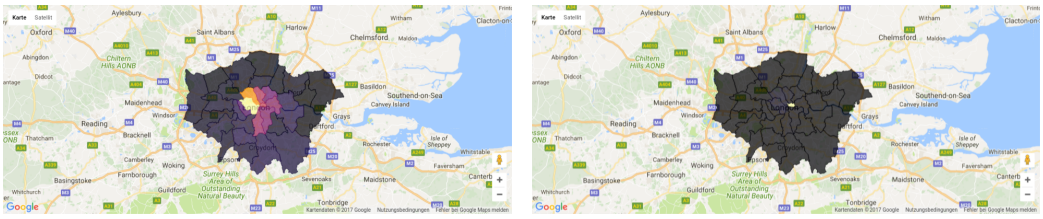
Figure 4: Activities per Category in Barcelona

the activities are a bit more spread out to the whole city. This can be due to Barcelonas small size and its high population density. In contrast there are Marid-Centro (Figure 11), New York-Manhattan (13) and Hong Kong (15) where the open events are strongly concentrated in one or a few central districts.

London

London seems rather centric with Westminster as the hot-spot district. But especially per capita everything is overshadowed by the small and little populated City of London. Later we will see that London has the highest number of events outside of Europe (in our dataset).

The most popular categories are *Socializing and Tech*.



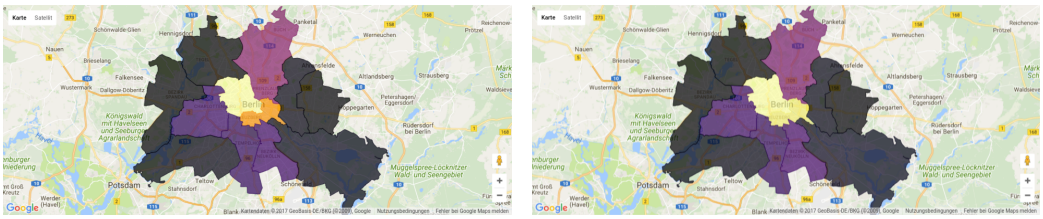
(a) Activity per District (b) Activity per Capita per District

Figure 5: London

Berlin

Because of the historical division of Berlin it does not have one big city center. One can see (Figure 7) that event concentration is slightly shifted to the east around Prenzlauer Berg (south of Pankow, East of Mitte and Friedrichshain-Kreuzberg. In contrast there is only a smaller focus at the City West (Zoologischer Garten, Charlottenburg-Wilmersdorf).

Regarding the categories (8) it is obvious that the *Tech* activities dominate in Berlin. The reason for this increased interest in technology may be due to a high number of upcoming startups. But actually the high dominance (three times larger than the second favorite; largest dominance) of technology events in contrast to other categories is a trait which all three examined German cities share. One conclusion would that Meetup seems so far to be limited to tech-interested people in Germany.



(a) Activity per District (b) Activity per Capita per District

Figure 6: Berlin

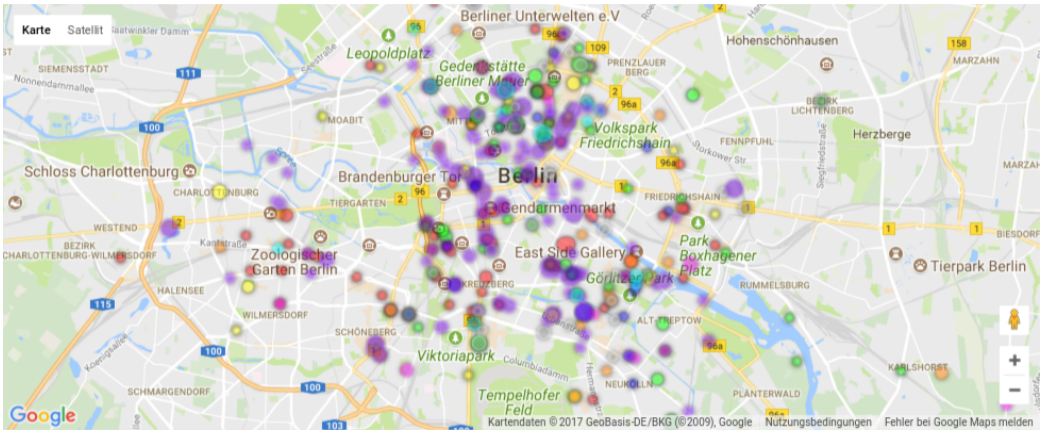


Figure 7: Snippet of Berlin with Open Events

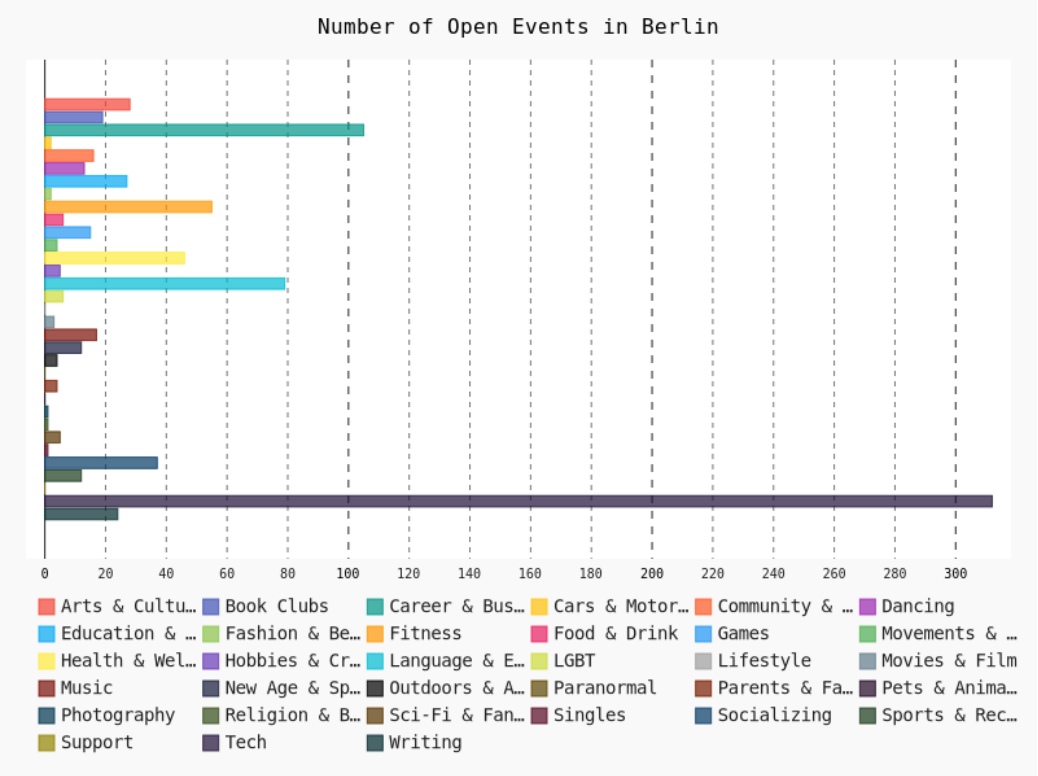
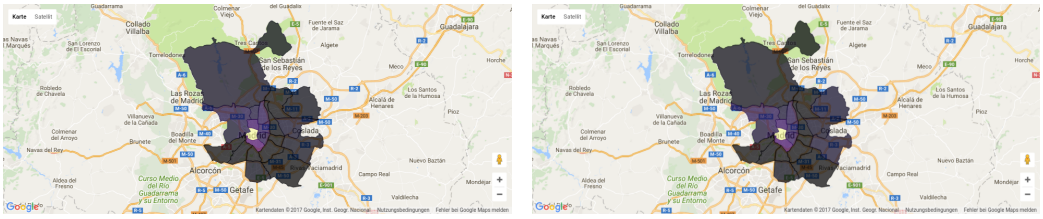


Figure 8: Activities per Category in Berlin

Madrid

Almost all events are concentrated in Madrid-Centro. Madrid's most favorite categories are by far *Tech and Languages* and *Ethnic Identity*.



(a) Activity per District (b) Activity per Capita per District

Figure 9: Madrid

Paris

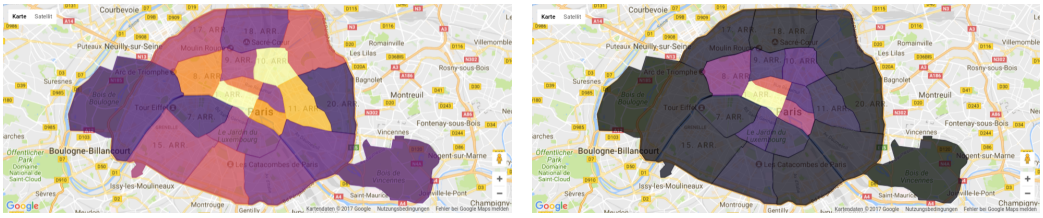
Paris is an example for a more evenly distributed city - at least in the total number of events per district. This may be due to its small size and dense population (compare Barcelona). The picture would probably be different if one considers the vast metropolitan area of Paris with its 12 Mio. people and its area of 17000 km^2 (in contrast to the actual 2.2 Mio. and 105 km^2) Again *Tech and Language* and *Ethnic Identity* are the most favored categories.

Brussels

In this case we actually did not consider the City of Brussels but rather the Brussels-Capital Region. But one can clearly see the bright yellow shape which represents the City of Brussels (this is one *district*!). Per capita some central regions like Ixelles show a higher activity too.

Hamburg

Hamburg is the second biggest city in Germany. As in Berlin mostly tech-interested people are using `Meetup.com`. Looking at the figures one can clearly identify the trending districts in Hamburg which are Hamburg-Mitte (middle), Altona, Eimsbüttel and Hamburg-Nord (from west to east). One reason for this might be that lots of young people are living in these districts. For example, in Hamburg-Mitte there is the very famous St. Pauli. In this part of the district a great part of the nightlife take place. Other districts like Altona and Eimsbüttel are trending as well since the University of Hamburg



(a) Activity per District (b) Activity per Capita per District

Figure 10: Paris

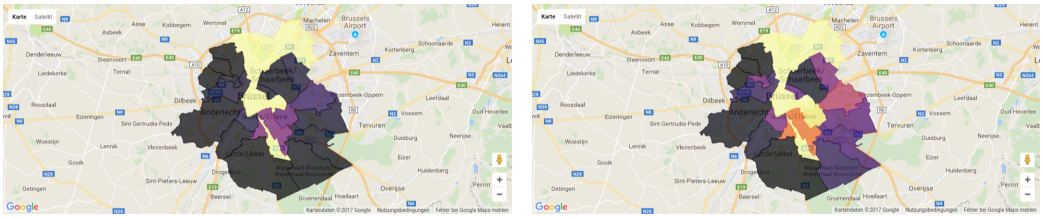
is located in Rotherbaum which is a part of Eimsbüttel and close to Altona, Hamburg-Mitte and Hamburg-Nord.

New York City

The biggest city of the United States of America has the second largest amount of open events in our dataset. Here it is necessary to keep in mind that the real number of events can vary if the Meetup communities are more organized in closed groups.

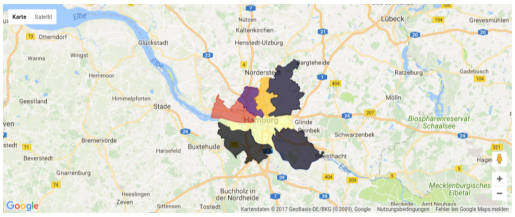
The picture New York shows is as one would expect. At least in the case for Manhattan which is by far the center of the City with over 1000 open events. It is followed by the distant second Brooklyn less than one third of activities. Staten Island and the Bronx are hardly active.

As *Meetup Inc.* is based in New York it shows a diverse picture in regards to categories. The leading categories are *Sports and Recreation*, *Tech*, *Career and Business* and *Socializing*.

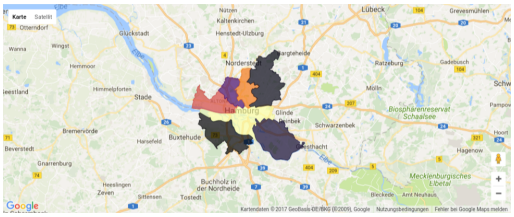


(a) Activity per District (b) Activity per Capita per District

Figure 11: Brussels

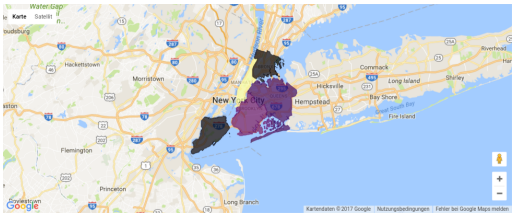


(a) Activity per District

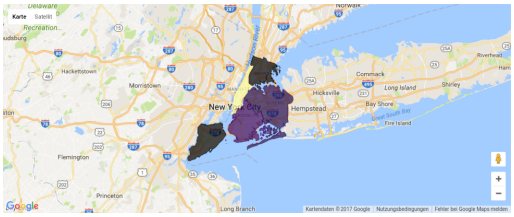


(b) Activity per Capita per District

Figure 12: Hamburg



(a) Activity per District

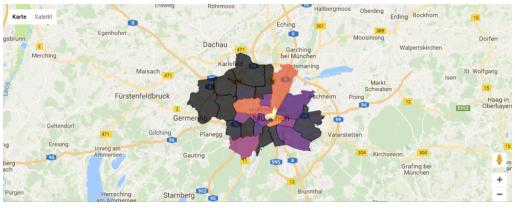


(b) Activity per Capita per District

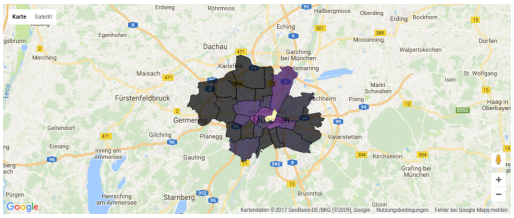
Figure 13: New York City

Munich

As the third German city in our set Munich does not seem to be much different than Berlin and Hamburg. Again most Meetup members are interested in Technology. In total number of activities Munich beats the more populated Hamburg and even Berlin if one considers activities per inhabitant.



(a) Activity per District



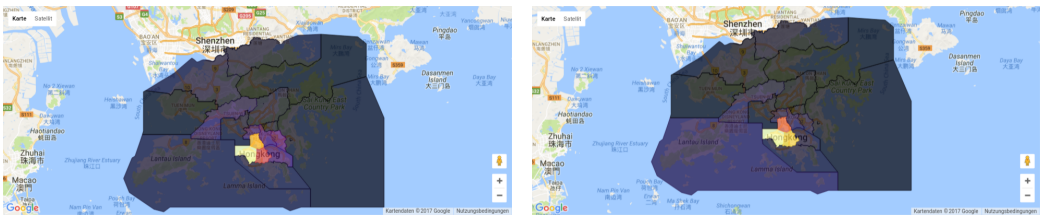
(b) Activity per Capita per District

Figure 14: Munich

Hong Kong

The first thing we notice is that Hong Kong has the lowest number of activities per capita in our dataset (compare Figure 16). It seems that `Meetup.com`

did not reach Hong Kong and maybe the whole Chinese or even Asian market. As a business center it shows the largest activity in *Socializing* and *Career and Business* followed by *Sports and Recreation*.



(a) Activity per District (b) Activity per Capita per District

Figure 15: Hong Kong

Further Observations

Because *Meetup Inc.* is based in America one can expect a wider reach in North America. As a small further comparison we considered Los Angeles, Palo Alto (California, USA), Vancouver (Canada) and Sydney (Australia). There are several immediate conclusions one can take from Figure 17:

- Palo Alto has a lot of events. Especially with its small size it surpasses all other cities in events per capita by miles. Due to its belonging to the Silicon Valley and its residing tech companies like HP, Tesla it is unsurprising that the people take most interest in technology.
- Vancouver is also rather small but has the second most events per capita. It seems also to be more diverse than others with interest in *Movement and Politics* which takes a larger part than in other cities.
- Like other North American cities Los Angeles also has a large `Meetup.com` community. Los Angeles has the most *Music* events in our dataset but surprisingly takes only the third place in *Movies and Film*.
- Despite Sydney being an English-speaking city it cannot compete with the numbers of American cities. It seems that `Meetup.com` is yet not that big in Sydney and maybe whole Australia.

5 Legal and Ethical Issues

A work must respect the protection of the information of its sources. Hence, all the data sources used along the work need to be taken into account in the study of its legal and ethical issues. But those issues related with data protection copyright and intellectual property legislation can be very tedious, specially if the source is a private company. To enhance the understanding, and even though some of the sources do not present copyright-licenses of this type, the icons used and released by the non-profit organisation Creative Commons will be used in this section. The Creative Commons licenses require attribution for copying, sharing and verbatim uses but there are other restrictions that may apply (see Table 1).

Meetup API

The information regarding the use that can be done with any information extracted from the Meetup API is detailed in section 5.8 *API License* of the

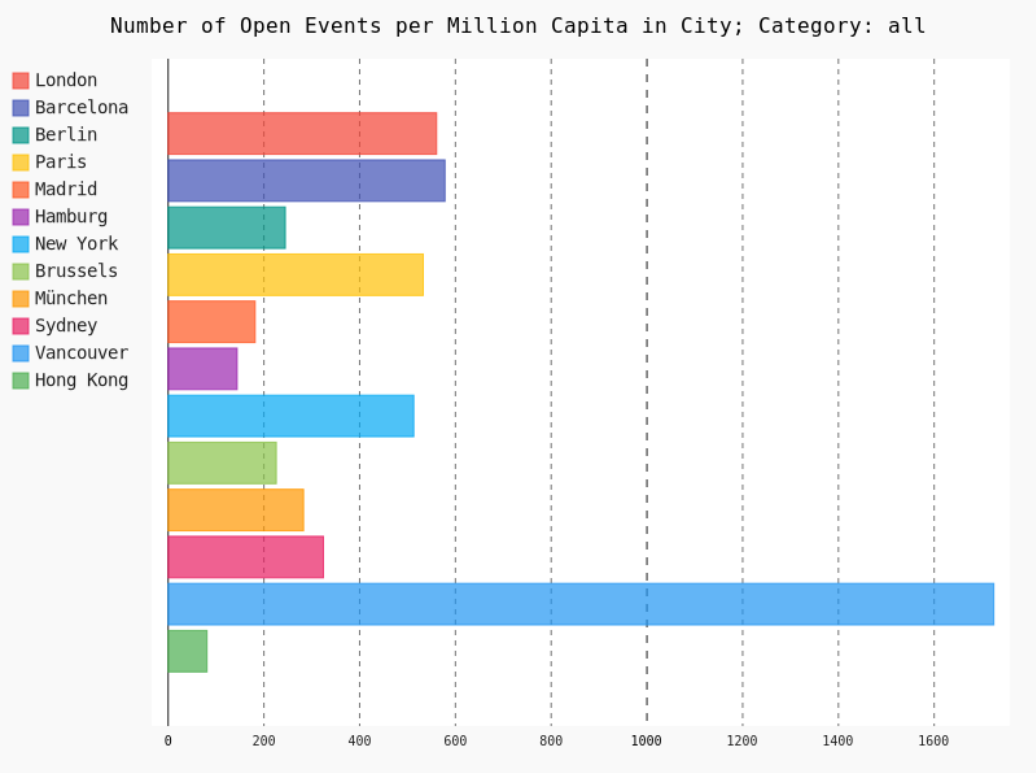
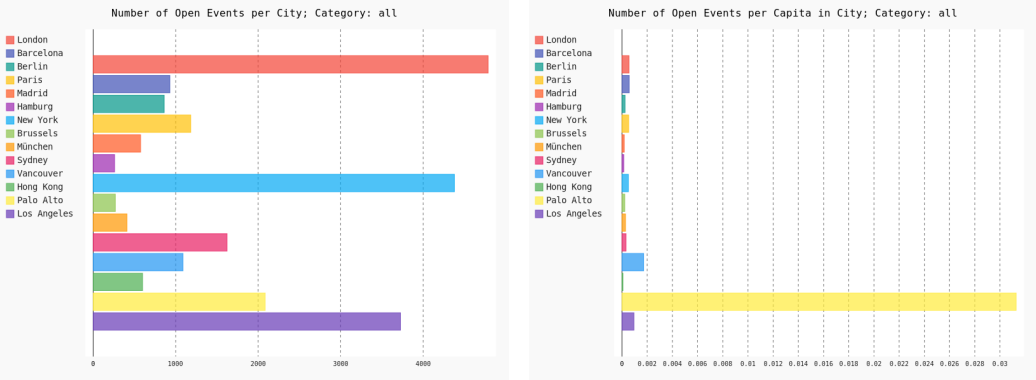


Figure 16: Number of Open Events in all Categories per Capita



(a) Open Events

(b) Open Events per Capita

Figure 17: Comparison with selected US Cities

Table 1: The five types of restrictions that may apply to a Creative Common license.

	**BY**	**Attribution** - credit to the creator is required
	**SA**	**Share Alike** - distribution of the work needs to be under the same licence of the work (e.g. a copy made from a Creative Commons work cannot be copyrighted)
	**ND**	**No Derivatives** - cannot make changes to or remix the work
	**NC**	**Non-commercial** - commercial use of the work is not permitted
	**NC**	**Public domain** - free for use by anyone for any purpose without restriction under copyright law

Terms of Service of Meetup, though some other relevant information is linked to other subsections. In this section it is explained that the restrictions **BY** and **ND** apply for any further application of the API. Notice however that commercial use of the work is not said to be prohibited. Section *Meetup API License Guidelines* details what kind of applications and conditions under which the commercial use of the information provided by the platform is allowed: "(...) enhance the Meetup experience or create specialized versions of our platform (...)" This conditions of "Reasonable commercial uses" already restrict a lot the possibility of making any commercial use of the data, so restriction **NC** could almost be added to the restrictions list too.

Google Maps API

The terms in which the use of this API is restricted are spread out along sections 6 to 12 of the *Google Maps/Google Earth APIs Terms of Service*. The summary of restrictions is found in section 8.2, *Service License*, where it is made clear that the license to use the Service is non-sublicensable. Section 9.1 *.1 Free, Public Accessibility to Your Maps API Implementation.* specifies the terms under which a commercial use of the Service would be allowed, while attribution and non derivatives restrictions are also specified in sections 9.4 and 10.5 respectively. All in all, the set of restrictions is finally **BY**, **ND** and **NC** in some cases.

Different sources of *GeoJSON* files

Since the GeoJSON documents where extracted from different sources, the restrictions of each map (i.e., of each city) vary. The Licenses of the files can be seen in the Appendix A.

Wikipedia

As explained in the *Terms* section of Wikimedia, the contents of this platform are only restricted by **BY** and **SA**, though exceptions exist for content contributed under "fair use".

From all the sources of data, we observe that the one providing from the Meetup and Google Maps APIs are the most protected. The data obtained from this sources needs to be treated very carefully, since legal issues may arise easily. On the other hand, the data extracted from Airbnb is completely

free to use, even though Airbnb itself is a very powerful company. We found this very interesting.

6 Limitations and Future Work

One major issue is that this project is depending strongly on the quality of the data available at `Meetup.com`. We have to rely on the users of `Meetup.com` so this issue can hardly be improved. One idea would be to integrate other social services like `Facebook.com` to extrend our data source and to get more independent. In this way one could also fix the problem that `Meetup.com` might be less known in non-English speaking countries. Having a wide base as data source one gets more independent of local and regional preferences regarding the social networks.

We should also keep in mind that we only retrieved data for *open* events. So far it is not possible to access all events in a city because for most of them you have to be a member of specific groups.

Another problem is that this project has the potential to be extended in the following way: So far our program is restricted to predefined cities. It is not possible to select new cities or regions. This is due to the fact that we work with GeoJSON-files which are saved locally on the hard drive. One improvement could be to create a user interface and the possibility to select new cities and regions. By selecting single categories of events one could get special information, e.g. where to go running. Additional, in an application, one could automatically create suitable graphics, charts and tables.
Another limitation is as well related to `Meetup.com`. Since we only have one data source for events, we only can give sufficient criterions for districts. A district having lots of events seems to be trending. However, a district having less events is not automatically less interesting and popular. It just means that its inhabitants are not using `Meetup.com`. Again, one should take more data sources into account.
One further aspect could be to look for interest of twitter users to make recommendations to activities nearby. The geotag of the tweets provides the central information for the search.

Our tool could help social scientists to prove their hypothesis with the provided data. Beside literature, studies and surveys our tool seems to be a good method for analyzing the behaviour and preferences of populations.

A Sources and Licenses of GeoJSON Files

Table 2: Sources and Legal Protection of the GeoJSON Files

City	Source of data	License
Barcelona	https://github.com/martgnz/bcn-geodata/blob/master/	CC-BY
London	https://joshuaboyd1.carto.com/	BY
Berlin	https://github.com/m-hoerz/berlin-shapes	CC-BY
Madrid	https://github.com/codeforamerica/click$\backslash_$that$\backslash$_hood/tree/master/public/data	-
Paris	https://github.com/blackmad/neighborhoods/blob/master/	-
Brussels	http://insideairbnb.com	CC0
Hamburg	https://matthiassuessen1975.carto.com/	BY
New York	https://github.com/blackmad/neighborhoods/blob/master/	-
Munich	https://mucx.carto.com/	BY
Hong Kong	http://insideairbnb.com	CC0

References

[1] Florian Mounier. Pygal. http://pygal.org, 2016. Python plotting Library.

[2] Google Inc. Google Maps. https://maps.google.com, 2017. Interactive online Map Service.

[3] Kenneth Reitz. requests - Python HTTP for Humans, version 2.14.2. https://pypi.python.org/pypi/requests, 2017. Python Library for sending HTTP requests.

[4] Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, Carol Willing, and Jupyter Development Team. Jupyter notebooks – a publishing format for reproducible computational workflows. pages 87 – 90, 2016.

[5] Meetup Inc. Meetup API. https://www.meetup.com/meetup_api/, 2017. [Online; accessed 11-November-2017].

[6] Martí Municoy, Alba Gordó, Jan-Hendrik Niemann, and Andreas Radke. Research and Innovation Project. https://github.com/schakalakka/uab-ri, 2017. GitHub Repository.

[7] Pascal Buignon. gmaps, version 0.70. https://github.com/pbugnion/gmaps, 2017. Plugin for including interactive Google maps in the Jupyter Notebook.

[8] Philipps Universität Marburg. Was versteht man unter dem Data Life Cycle / Daten-Lebenszyklus? https://www.uni-marburg.de/projekte/forschungsdaten/faq/datalifecycle, 2014. [Online; accessed 15-November-2017].

[9] Python Software Foundation. Python Language reference, version 3.6. https://python.org, 2017.

[10] Spirion LCC. What is Data Lifecycle Management? https://www.spirion.com/data-lifecycle-management/, 2017. [Online; accessed 15-November-2017].