Final Report Geovisualisation Assignment

Lukas Graf
University of Salzburg, Department of Geoinformatics Z_GIS
Winter-Term 2019/2020

Motivation

BLUEBikes[1] is a provider of rental bikes in Boston and surroundings, Massachusetts, USA. The data on the geographical location of the rental stations as well as historical data on rides have been made publicly available for research purposes and provide valuable information on user behaviour and bicycle use in Boston.

In addition to the time and route of the trips (start and end stations), the data provided by BLUEBikes includes anonymous information about the users. The gender, age, time of the trip, and type of use (customer or subscriber) can be viewed and are available in tabular form. However, in order to generate information from the data, further processing and visualization is indispensable. The methods of geovisualisation, which combine explorative statistical analysis with cartographic methods, are therefore particularly suitable for further analysis of the data.

In the present visualization such an attempt was made and graphs were combined with a cartographic representation. The resulting dashboard allows the data to be analysed interactively on a daily basis and can, for example, help those responsible at BLUEBikes to better understand the day-specific user characteristics. The dashboard uses exclusively open-source technologies and has a modular structure. This means that individual functions can be decoupled, expanded, or supplemented with missing analysis options and thus adapted to user needs. To allow for full reproducibility of the workflow and dashboard, the required scripts and style sheets are available on Github[2].

In the following, the methods, i.e. the technologies used, and the results of a demo use case are presented and the advantages and disadvantages of the dashboard solution are discussed.

Methods

The data is displayed using a web-based dashboard, which is written in JavaScript and HTML. The following technologies are used:

- **Leaflet** – a library for creating interactive web maps - is used for visualising the cartographic content of the dashboard
- **D3** – a library for creating data-driven contents of web pages such as graphs and diagrams – is used for visualising some statistical properties of the data
- **Ajax** – an API for making asynchronous calls – is used to load the required BLUEBikes data into the dashboard and refresh the content when required without reloading the entire dashboard
- **Python3** and especially the **pandas** library are used to pre-process the BLUEBikes

1 https://www.bluebikes.com/
2 https://github.com/lukasValentin/BLUEBikes_Visualization

data and convert the tabular data into JSON files that can be loaded into the dashboard
- **CSS** and **HTML** for styling and structuring the dashboard and its contents

The dashboard shows six different things:

- temporal distribution of trips on an hourly basis
- frequency of user-types
- frequency of different age-segments
- frequency of different gender
- a heat-map showing the spatial pattern of bike usage
- the overall number of trips

The reason for selecting these variables is to provide a quick overview of the main user characteristics in terms of time and space and demographics. The data are aggregated in such a way that a quick assessment can be made, as the number of categories is relatively low, but on the other hand there is still enough information left for analysis.

The sequence for creating the dashboard is divided into two parts:
First, the provided Python script is used to convert the data into the required format:

Data is downloaded from BLUEBikes hosted on Amazon Web Services covering a desired period of time (e.g. October 2019) using *wget* and its Python3 interface. Next, the downloaded data is unzipped and loaded into a *pandas* data frame. Incomplete records are remove from the data and only the desired columns (in this case: age-segment, gender, user-type and geographic location as well as temporal information) are kept. Since the data provided by BLUEBikes is very large even it usually only covers one month, the user has to specify a particular date which will be exported to a record-oriented JSON file. This JSON file can then be hosted on a web server where it is accessible through the internet (e.g on Github). Please note that a restriction to single days is necessary as browsers will most likely fail to load the whole BLUEBikes trips of an entire month.

Second, the JSON data once hosted on a web server is called and loaded using ajax. The cached JSON data is subsequently parsed over to D3 to make the graphs and Leaflet to provide a map visualisation. In more detail, the temporal characteristics of the bike trips are visualised on a hourly basis to allow users of the dashboard identifying temporal peaks. Moreover, the user-type (customer or subscriber), the distribution of age-segments (grouped in age cohorts at 10 year intervals) and the distribution of gender (unfortunately, the gender is encoded as numbers of 0 to 2 without any explanation) as well as the overall number of trips for the depicted date are shown using D3 row and bar charts. This allows a quick overview of the user characteristics.
The leaflet map shows a scaled heat-map, which uses the heat map plug-in for leaflet and thus shows where a particularly large number of trips start.

Two days in October 2019 were selected for the demonstration of the tool, representing a weekday (October 1) and a weekend day (October 13). Thus, differences in user behaviour between weekday and weekend can be shown. It is of course possible to include and analyse further dates.

Results and Discussion

In the following paragraphs the main characteristics of the shown two dates are briefly outlined to demonstrate the capacity of the dashboard for analysing the BLUEBikes dataset:

On the depicted weekday two temporal peaks around 9 am 6 pm are clearly visible which represent most likely the rush hours when most people are heading to their offices and home again. Overall around 11 000 trips were made. A clear majority of the trips was made by subscribers which clearly outnumbered the amount of trips made by customers (9726 to 1133, respectively). Most trips were made by users aged between 20 and 40 years with only 30 out of 11 000 trips made by users older than 70 years. The heat-map clearly revealed that most trips were made in the inner city area of Boston with peaks around the MIT campus and train stations.

On the selected weekend day, no rush hours were visible but the number of trips increased towards the evening and 5 pm and slightly declined afterwards. The overall amount of trips was smaller (6874) than on the weekday and the difference between subscribers and customers was smaller. (4692 to 2182, respectively). The gender and age segments did not reveal any significant differences but slight changes were observed in the geographic distribution of trips.

This means that the dashboard presented can be used to quickly and effectively highlight differences in user behaviour and their characteristics. It could be shown that there is no rush hour at the weekend but that the afternoon has significantly more trips than the morning and late evening. During the weekend, there are also more occasional drivers (customers) on the road than during the week, when subscribers in particular use the service for their way to work or university. The offer is mainly used by younger people (possibly students because of the many trips to MIT) and is mainly concentrated in the inner city areas.

The dashboard is a simple and visual way to analyse and quantify data. It helps to perform simple analyses (such as which times are particularly busy) but reaches its limits for more complex analyses. In addition, the limitation to a few thousand trips is not necessarily satisfactory for all applications due to browser capacities. For more complex analyses or investigations with more data, desktop solutions such as Jupyter notebooks are therefore at an advantage because they are more powerful.

Conclusions and Outlook

A dashboard for geovisualisation, which is fully reproducible, was presented, and the possibilities of the tool were illustrated using two sample data sets obtained from BLUEBikes. The dashboard enables quick and simple analyses, but reaches its limits when it comes to more complex tasks and larger data volumes.

Further developments may include the creation of additional statistics, as well as an improvement in performance so that larger data volumes can also be analysed.