Scientific Visualization
Second Information Visualization Project
CitiBike Usage in NYC

Andrea Mastropietro

Deadline: January 21, 2018

The aim of this project consists into creating a dashboard-like visualization that provides information about the usage of CitiBike, the public bike rental service in NYC. This service is widely used by NYC citizens since is not common to own a car in New York, and people prefer to move using taxis or bikes. To convey more insights about the data, the visualization combines a map, bar plots and small multiples as coordinated views.

1 Data processing

The CitiBike service regularly publishes on its website data about the usage of the bikes. For the project, I downloaded data regarding one month of usage (October 2018) but I concentrated on the average daily usage of the service. The dataset provides detailed information about each trip, such as the starting station, the trip time, the age and the gender of the user and additional information such as for example the kind of user (if he/she is a subscriber to the CitiBike service or not). The dataset I downloaded did not present particular problems such as missing information except for one station that had a NULL value as name; I was able to edit the dataset setting the correct name only after having plotted the station on the map to see where the station was located (Fordham University in the Bronx). Moreover, I discovered some inconsistencies regarding one station using the map visualization and the clustering, but I will describe this situation later. From the original dataset I created several additional files containing only the information needed for the specific view.

2 Visualization design and usage

As aforementioned, the visualization provides different ways to explore the data. First of all, a map of NYC shows the positions of the bike stations (properly divided into clusters), provides details on demand and works as a filter on the other views. Secondly, small multiples allow the user to inspect each element of the cluster and finally, two bar plots allow for a comparison of the usage of the bikes according to several filters on some properties (station, age and gender).

2.1 NYC Stations Map

For the map visualization I used the images by Openmaps, a free service that provides map visualizations showing different details of the area according to the zoom level; as I will say, this will come pretty useful while studying the stations behavior. Each station is represented in the map as a circle which color depends on the cluster the station belongs to.

Clustering In order to let the visualization be more responsive, I precomputed a k-means clustering (k = 5) in Python using the methods provided by the `sklearn` library. I run the clustering algorithm twice on the data, once using as feature vector for each station only information about the number of trips and the average trip time and the second time using all the data available, such as for example the number of female/male users, the user type and so on.
The algorithm yields two different clustering results according to the feature vector used. Since both of them are worth to be visualized, the user can select the kind of clustering to be shown from a drop-down menu.

Anyway, regardless the clustering features used, by looking at the map we can immediately have an overview of the stations distribution and behavior. First of all, despite of the service being meant for all NYC, it is straight to see that we have a high concentration of stations only in Manhattan and on the marine areas of Queens and Brooklyn; the Bronx, Staten Island and the internal areas of Brooklyn and Queens have none or a scarce number of stations. Therefore, a person living in those areas has no possibility to use the service and if a manager of the CitiBike system used the visualization, they could understand in a quick way where to build a new station.
Moreover, we can see an enough clear division of the usage of the stations

between Uptown and Downtown Manhattan, meaning that people living or working in the same area have similar habits that are different from the ones in the other area.
By hovering on one of the circles in the map it is highlighted and details on demand about that station are shown (such as for example average trip time, number of female and male users and so on).
Combining clustering and details on demand we can discover outliers and make hypothesis on why they are actually outliers. One of the clusters, in fact, has only one station (in the both clustering outputs). Analyzing the details we can see that its average trip time is more than 6 hours; this is probably an error in the data that let the station fall in a cluster on its own. Thus, this fact shows how a visualization can also be used to find errors in the data that were not seen before.
The details provided by the Openmaps views, as I introduced, are quite useful. For example, by exploring the map we notice that there are several stations on a line that belong to the same cluster; by zooming we see that those are the stations on Broadway, one of the most important and crowded streets in NYC and indeed by looking at the details we can see that those stations all have a very high number of trips per day.
The map gives also the opportunity to switch from standard colors to colorblind-safe colors for the clusters using a checkbox; doing so, also the legend of the map is updated accordingly.

2.2 Cluster Explorer - Small Multiples

The clusters can be explored using small multiples. This visualization provides a maximum of six bar charts showing the usage of the bikes throughout the day (each bar chart is referred to a station), showing the bikes usage in the morning, the afternoon, the evening and the night. Since small multiples are meant to show information about many samples, the granularity of the time division on the x-axis is coarse since giving details about each hour would result into a too cluttered view.
Using a drop-down menu it is possible to decide which cluster to show in the view and since the clusters usually have a rather high number of elements, using two buttons it is possible to move among the stations in the cluster showing a maximum of six stations at a time.
Moreover, the view provides details for each clusters, about the daily number of trips and the numbers of stations in the cluster.
Since the views are coordinated, when changing the clustering method in the map, also the stations in the clusters in the small multiples view will be

updated. Thanks to the small multiples, it is also easy to identify outliers and study them, since an outlier forms a cluster on its own.
To allow for comparisons, all the charts shown in the view have the same scaling on the y-axis.

2.3 Bikes Usage Bar Charts

The bottom half of the view is dedicated to two bar charts, that are more detailed than the ones of the small multiples. Those charts let us study the evolution of the usage of the bikes during the day seeing the average number of trips at a granularity of 2 hours, conveying thus more details than the small multiples.
Initially, the plots show the usage in all NYC and we can immediately notice that the highest number of bikes is used from 8 to 10 and from 16 to 20, that are the hours in which usually people go to work and come back home.
The plots can be filtered according to the gender and the age, letting the user study which kind of people prefer to use the service. Using the filters, we can see that the CitiBike service is mostly used by men rather than women, and that 40-year old people use it more than 20-year old people.
As in the small multiples, to allow for a confrontation, the charts have the same scaling on the y-axis.
Moreover, when clicking on a station in the map, it will act as a filter on the plots that will show the usage of bikes in the selected station.
When we hover on a bar in the plot, on the other bars we will see the difference of the number of trips at that hours with respect to the hovered one and since the plots are coordinated, the same happens in the other plot.

3 Conclusions and Future Work

This visualizations provides a compact and complete way to analyze how people use the CitiBike service in NYC and which kind of people prefer to use it, providing both an overview and details on the data, being able to filter them according to several properties. It can be used to see what are the zones in which the service is mostly used and which ones need an improvement of the service, for example by building additional stations in areas in which there are not enough. The ones illustrated above are only example on how the visualization can be used; by further exploration of the data one could discover new and different insights.
Moreover, as future work the visualization could be improved for example by showing which are the most common pick-up/drop-off paths done by

people and by linking the small multiples to the map in a way such that selecting a station in the small multiples centers the map to the position of such station.