Interactive Data Visualization (DATA15003)

Miia Rämö
Laura Paajanen
Ville-Veikko Saari
Juha-Pekka Moilanen

Assignment 2 - Interactive visualization design

Application: https://citybike-statistics.github.io
Source code: https://github.com/citybike-statistics/citybike-statistics.github.io

Data and motivations

HSL (Helsinki region transport) provides data about the usage of city bikes from summer 2017 in json format. The data has to be visualized to get it in human readable format. Since the data was from every minute and some stations were missing some of the data, we decided to use hourly averages of how many bikes there are at a station. We showed the data on weekly, daily and hourly scale depending on the part of our visualization. The question we wanted to answer was if there are any patterns in the usage of the bike system and which factors may be the cause of that. Another interesting question is would it be possible to predict the usage before hand.

Our user base is wide since the visualization provides fun insight to regular city bike users but also a neat way for the people from HSL moving the bikes around Helsinki to learn how for example big events affect the system usage. We decided to divide our website into three subsections: system availability showing how many bikes are available, station view to see the situation at a specific station and a heatmap which gives an overall view of the system. This way we let the user decide which part of our visualization is most useful to him or her. Tasks with our visualization could be abstracted to either having fun with finding patterns and optimizing the bike transfer with the car.

Design

In the heatmap we chose to use circles to indicate the amount of bikes on a station. If a station is empty it is shown on the map with a red cross to make the empty stations pop out. We initially thought about using color scale from red to green but that being an inclusive way

of visualization due to red-green colorblindness we ended up using only green color and indicating the amount with the size of a circle. According to Tufte's data-ink principle the visualization should use as little "ink" as possible [2]. Our visualization respects this principle. In the heatmap this occurs in the background map which is presented with a low contrast styling which does not distract the observing of the markers. We also chose to use only one shade of green to increase the data-ink ratio and showing the empty stations with different shape to help color blind viewers.

We used logarithmic scale with the sizes since linear scale would be messy is certain situations (figure 1). According to Mackinlay's ranking, in this case indicating the amount of available bikes with area is the wisest since the data is quantitative. Even though area is a visual variable which perceptual judgement humans are biased [2] to we considered it to be informative enough since the circles show if there are bikes available rather than the absolute amount of bikes on a station. For a regular user it does not matter if there are 35 or 135 bikes on a station. We still did not set a maximum size for the circles so anomalies like huge events would show. The absolute amount of bikes on a station can be seen with hovering over the circles. The size difference of the circles also, according to Munzner [1], creates a popup effect for the stations with more bikes. For the HSL people driving the car which moves the bikes it might be useful to show the data some other way but our visualization at least gives an overview of the whole situation. A sample of our marker sizes can be seen in figure 2.

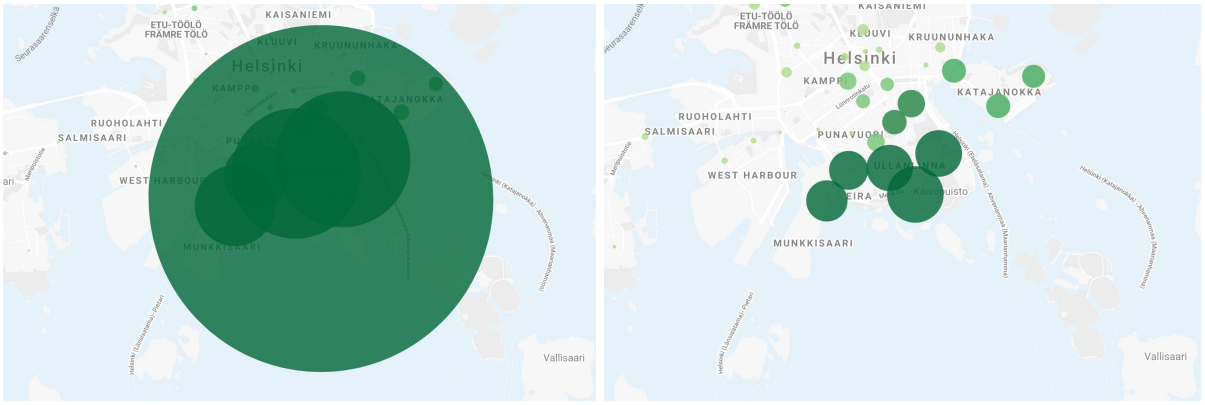

Figure 1: Linear versus logarithmic scaling.

Figure 2: Sample of marker sizes.

In the station view we used a familiar view and usability of the Google maps to make it as easy as possible for our users. Plotting station view graphs to the same chart would make comparison between selected dates and stations more straightforward, and to make the visualization clear it would need different colors for each line. As Munzner's model [1] suggests we used line charts instead of bar charts because the time attribute which is our key attribute is ordered and the amount of bikes attribute if quantitative. Bar chart would be a better solution if our key attribute was categorical. The minimum size for the Y axis showing the amount of bikes is 30 but it is increased if necessary. We chose to have a minimum size for the Y axis because if the size would be the maximum amount of bikes during the day, it would seem that there are many bikes at a station even if there were only five or so.

Tuftes data-ink principle [2] is respected in the station view and system availability views. The line charts are drawn in a very simplistic way so that data is easy to read. The axes have a moderate amount of ticks and the labels are short and clear. To improve readability we added lightweight horizontal and vertical lines for each tick.

Our visualization relies heavily on interaction. In the heatmap view the user is able choose date and time to show data from that specific hour. In addition to the text input we provide buttons for +/- 1 day and hour. These both serve different but logical purposes. By pressing the "+1 hour" button the user is able to see how the situation is evolving hour by hour. The +1 day button provides a way to observe how the situation varies between different days but at the same time of the day. The user can also find out the accurate amount of bikes on a specific station by hovering over a circle on the map.

Figure 3: Tooltip showing detailed information

The system availability view provide similar interaction for choosing the dates as the heat map. The user can choose a date with the text input and then alter the plot with buttons.

Figure 4: plot date input and buttons

In the station view the user can choose interesting stations and compare the usage for them. In the current implementation the comparison is made between separate plots which appear above each other in a “plot history”. The plots are deletable to allow manipulation of the “plot history” and hence different comparisions to be made.

The visualization works slowly when changing the date in the heatmap view. This is because our current implementation seeks through 105805 rows of data to find that match requested date and time. Only 148 rows are actually needed for one view of the heatmap. In a proper application the data would be fetched from a backend and a database where the fetching would be done in a much smarter manner. Drawing of the circles also seems slow because the old circles are first removed and the new ones then drawn.

Insights

Our initial question was if there are any patterns in the usage of the bike system and which factors may be the cause of that. Another thing to look at was would it be possible to predict the usage before hand. We found that during weekdays most of the bikes moved towards the city center in the morning and again out of the city center in the afternoon. This is of course due to people riding to and from work. Other areas where a lot of jobs take place can be found from the heatmap similarly. On Saturday and Sunday the most usage takes place

during noon. A spike in the usage and movement out of the city center can also be seen during the night when bars and clubs have just closed. One really interesting finding was the effect of big events to location of the bikes. For example on 9th of June most of the bikes moved from all over town towards Kaivopuisto where an air show took place.

We also noticed that the current weather can quite well be guessed from the system availability view of our visualization. If the usage of the system was low during the day it most likely rained hard. During sunny days the usage was of course high. Implementing weather data information in the system we could predict the usage for days in the future.

Limitations

Besides the lack of weather data, our visualization has some other limitations as well. The major issue at this point is the size of used dataset. Visualization uses data from June 2017, and since the current absence of backend and database, adding more data would slow the heatmap even more. HSL offers also live data, and implementing this would make more value to this visualization.

As said, the weather conditions can be easily seen from user rates, but adding real weather statistics to system availability view would verify these observations and show the correlation for users too. Also day selection with calendar view would make visualization more usable and intuitive for users. Picking days by weekday would be easier, now user needs to know exact dates to be able to compare for example weekends and week days. One way to make this selection easier would be to use a day selector in a calendar format.
At this point comparison between usage levels of different stations on station view is quite challenging since all stations and times are plotted on separate charts. This could be done more informative way if one chart could have many graphs.

In the future we have plans to implement this feature with different colored graphs, where the station number and selected time would be visible when hovering over the graphs. It would also be nice to somehow highlight the station on map which selected graph is referring since the numbers of stations we use currently are not very informative or intuitive for users. Our aim is also implement the backend and database and continue with adding live status of the stations.

Contributions

The input for this project was quite evenly balanced between all the members of the team. We all contributed to the code, presentation materials and the final report. We did a lot of pair coding to create the application and we also worked side by side when preparing the presentation and this final report. Juha-Pekka had the initial idea for the project and he did the data wrangling i.e. loaded the data from HSL, imported it into a postgres database in order to create suitable, cleaned up aggregates for this project. This extra effort was compensated with a slightly narrower contribution to the final report. Juha-Pekka and Ville-Veikko were the ones presenting our work in class. Miia and Laura were in charge of writing this final report with help of Juha-Pekka.

References

[1] Tamara Munzner. Visualization Analysis and Design. CRC Press, 2014.

[2] Edward Tufte. The Visual Display of Quantitative Information, 2nd Edition. Graphics Press, 2001.