# Bike Sharing Data Visualization


| Konrad Piastka | Dominika Leszko | Lukas Früchtnicht | Joana Lorenz |
|---|---|---|---|
| M20180038 | M20180077 | M20180067 | M20180412 |


## 1. Introduction

Bike Sharing is becoming more and more popular with an increasing number of companies distributing their bikes over large amounts of stations and more and more users switching from their usual means of transportation to the use of shared bikes for daily commuting.[1] Extensive real-time data on these services is made readily available through the General Bikeshare Feed Specification, a uniform agreement on Bike Share data[2], and holds great potential for data-based applications that could benefit both the companies and the users of such services in the challenges that arise. While the company is in search of an optimal distribution of their limited number of bikes among the stations, the user is faced with a vast amount of stations to choose from as start- and end-destinations and is looking for a fast and easy use of the service. Thus, for our project we chose to build a prototype interactive application for a bike sharing service that supports both a user and a company log-in.

On the corporate side, we equip the company with an interactive map that should aid in the understanding of user's paths within the city and the frequency of usage in certain areas and at specific times. This could serve as a base to optimize the distribution of the bikes and thus better serve demand to ultimately maximize the utilization of the bike sharing service. Furthermore, interactive summary statistics about the customer base and their support a better understanding of the customer in order to distribute Marketing efforts most effectively and develop meaningful subscription models.

On the user-facing side, we are aiming at an improved user experience and a facilitated usage of the service by providing an interactive tool that enables the evaluation of distances between stations or to find the nearest station to a given position. The former specifically helps the user to estimate trip duration, which is the most important factor in the cost associated with a ride, while the latter is especially helpful if the user is looking for the closest station to ride to or to pick up a bike.

## 2. Dataset Description

The dataset used for the visualization project is taken from *kaggle.com/c/bike-sharing-demand* and contains bike sharing data on the Seattle Bike Sharing Service *Capital Bikeshare* between October 2014 and August 2016. Three CSV files are provided. The *Trips* dataset contains a documentation of each trip taken, including information on start and end station of the trip, the temporal details as well as information on the customer's demographics. The *Stations* dataset provides information on each station, including geospatial data and information on the number of installed docks. Lastly, the *Weather* dataset provides detailed information on Seattle weather conditions for dates within the timeframe of the *Trips* dataset.

[1] F. Richter: "The Global Rise of Bike-Sharing" (2018). Online: https://www.statista.com/chart/13483/bike-sharing-programs/

[2] Documentation of the General Bikeshare Feed Specification (2018). Online: https://github.com/NABSA/gbfs

*Table 1* summarizes the content of all three tables and specifies the linking fields through color-coding. The weather dataset contains several fields for each named observation – e.g. minimum, maximum and average temperature for a given day.

| Stations | Trips | Weather |
|---|---|---|
| **Station ID** | Trip ID | **Date** |
| Name | **Start Time** | Temperature |
| Lat | **Stop Time** | Dew Point |
| Long | Bike ID | Humidity |
| Install Date | Trip Duration | Sea Level Pressure |
| Install Dock Count | **From Station ID** | Visibility |
| Modification Date | **To Station ID** | Wind Speed |
| Current Dock Count | User Type | Precipitation |
| Decommission Date | Gender | Rain (Y/N) |
| | Birth Year | |

*Table 1: Field structure of the used tables*

## 3. Visualization and Interaction Design

Since the base of our analysis is inherently geospatial, our visualization was inspired by other visualizations that aim to show the magnitude of an attribute for individual geolocations, such as the 2015 Urban Population Visualization[3]. The interactive summary statistics were based on other Dashboard-type visualizations that aim at providing better insight into business questions[4]. All concepts and techniques were developed based on content from "Data Visualization Analysis & Design" by Tamara Munzner (2014) as well as from the Data Visualization Lecture at NOVA IMS in 2018/19.

The App is divided into two main sections, the company and the user view, which in a real-life application should be separated through a different log-in for the two parties. In our App, this division was simulated through the display of two different tabs. Then, within each tab the user can effortlessly navigate to the different sections using a drop-down menu. Additionally, we provide a *Data* tab, that allows for more detailed evaluation of the raw datasets.

For the user-facing application, the user is provided a map-view, where they can interactively choose to check the *Distances* or *Proximities.* Through the use of an reactive UI, which will adjust to the user's choices and allow further input, we are able to limit the features shown to the user, enable easier navigation and avoid confusion about the current selection.

On the company-facing side, four sections are provided. In the *Trip Analysis*, an interactive map enables the company to dive deeper into the details of bike usage throughout the city, by showing the counts of incoming or outgoing trips per station. Following the principle of

[3] "Visualize 2015 Urban Populations with Proportional Symbols". Online: https://carto.com/blog/proportional-symbol-maps/

[4] "Dashboard Visualization Options". Online: https://www.kyubit.com/Dashboard-Visualization-Options

'Overview first, detail on demand', the default view shows the average count of starting trips per day per station. The interactive UI allows the user to either look for general day types or observe particular dates. The latter could specifically be useful in understanding how events on a particular day influence the usage behavior. In order to show differences in movement direction, both start and end stations can be analyzed, based on the user's choice. Furthermore, differing times of day are provided. In order to guide the user in his understanding, the top five stations will always be showcased, providing the user with the option to narrow down the shown stations on the map as well.

The *Trip Stats tab* aims to give the user an interactive tool to explore the trip data in more detail. Two graphs allow the user to see more specifics about the trip. The first graph offers a time series analysis of the trip minutes per week. As this data is rather extensive, the user is given the feature to zoom into the graph either by directly selecting the specific area in the graph or by selecting a specific time period from a range selector displayed underneath the graph. The second graph shows dependency of the number of trips per day on given weather conditions. A scatter plot is used to explore the influence of the mean temperature in Fahrenheit of a day on the number of trips per day. Moreover, the user is given an option to draw a regression line on the scatter plot to better capture the dependency.

The *Customer Analysis tab* is devoted to provide insights about the tendencies across different types of customers. As the user opens the tab, an overview is shown through the number of trips for both male and female. Further, the user can select an option to show 'More details', where additional insights about customers are provided with the ability to filter data according to the gender. As the customer filters data as needed, an interactive box plot and pie chart appear, providing statistics about trip duration distribution and customer age proportions.

Lastly, in the *Dataset* section, the stakeholder is able to view the full data set based on the given inputs, in case a deeper understanding of the specific data structure is needed.

## 4. Data Abstraction and Encoding Choices

In order to make the visualization most useful and facilitate the understanding of connections, we performed some adjustments on the dataset prior to the analysis. First, for the *Trip Analysis*, we calculated averages for particular day types, such as weekends. This analysis provides both a broader and a more accurate picture than looking at specific dates, which will most probably include random variance. Furthermore, since the total count of trips per station is generally relatively low in our dataset and an hourly analysis, for example through an interactive slider, would not bring useful results, we decided to aggregate the daytime values to four time-slots: Morning, Midday, Evening and Night. This way, we are limiting the user in their choices on one hand, but are able to guide the user in their decisions and output more usable results on the other hand. Furthermore, in order to smooth extreme values in our channels, we took the logarithmic value of our variable of interest, the trip count, as a cue for the circle radius channel. Moreover, the logarithmic scale also allows to focus on the relative changes in the number of trips than absolute change.

Regarding data encoding, it was generally clear, given the explicit geospatial positions provided in our data, that a geometric visualization would be appropriate for our analysis. For the *Trip Analysis*, beside the map-view, other considered options included a more abstract geometric

visualization, such as geometric fields or a grid structure. However, since the specific location of a bike sharing station is probably the main influencing factor of its usage, we wanted to showcase each station embedded in the overall infrastructure of Seattle, in order to identify trends within certain regions of the city, such as downtown, seaside or outskirts. Circles were then used as marks for each station's specific geospatial coordinates.

In order to encode the variable of interest, the count of incoming or outgoing trips per station, we used the circle radius as a channel. Thus, the larger a stations circle marker, the higher the amount of trips at this station. Alternatively, a heat map could have been used. However, this did not seem appropriate since we were specifically interested in the unique stations, rather than areas or aggregates of stations. Also, the number of trips could have been encoded through a luminance/saturation instead of circle radius. But since space is the most valuable and recognizable channel, we chose to use area as the main channel in order to reach maximum effectiveness. We refrained from double-encoding in order to not overload the map.

For the *Stations Analysis*, we were equally dealing with geospatial data, thus similar considerations were made as for the *Trip Analysis*. Additionally, for the *Distance* analysis between two stations, the distance in meters, i.e. the link between them, was encoded as a line. This should specifically make clear to the user that the calculated is an aerial distance, not a specific route between the two stations.

In the *Trip Analysis,* in the scatter plot each data point, i.e. single day, is encoded as a mark, with its spatial position determined by the number of trips (Y-axis) and the mean temperature of the day (X-axis). Encoding the two attributes through the most effective channel type, the spatial channel, allows for maximum interpretability of the correlation. Readability is improved, and cognitive load reduced through the addition of a rug on both axes. Furthermore, a pattern on the Y-axis, expressing the density of data points in a specific region, and the possibility to display a regression line support the user in the interpretation of the relationship between the two variables. The user is furthermore enabled to interactively choose between weather events [Rain, Sun], which leads to a sampling of only days with rain or sun for the scatter plot, respectively.

The time series plot enables the user to thoroughly analyze bicycle trips over time. The user has an overview of the sum of trips per week but can also take a closer look into the data by zooming in on specific date ranges. The time series uses a sequence of time data, which was discretized to a week level to increase readability and effectiveness of the visualization. For encoding, we chose the most common, thus familiar, approach by using a line chart to connect the individual data point marks, with a colored area underneath the line representing the attribute size at each position. The data point marks that are encoded as a star shape stand out against the line chart very clearly. The ease of use of the interactive navigation enables the user to discover trends or any anomalies in a fast and effortless way.

In the *Customer Analysis* tab, the *Number of trips per month* chart is an example of a side-by-side bar chart, where month and gender are key attributes, while the continuous variable of number of trips is expressed as the height of a bar. The second part of the tab content is interactive and filtered by the values of the *gender* attribute. The *Trip duration by month* box plot shows the distribution of trip duration per month (by gender when filter is in use). Finally, the pie chart is using a radial layout to express the proportions of different age groups of customers, as it is

common practice for part-of-whole analyses and thus a familiar design for the user. The number of customers falling into each category is encoded by means of area marks and an angle channel. Additionally, age categories are encoded with a color hue, to clearly separate groups.

## 5. Implementation

The App was fully developed using R in the Shiny environment. First, several different date fields were calculated, since a lot of analysis depend on different levels of date time. Next, we created the UI, selecting the most appropriate widget for each interactive feature. Within the server, the reactive expressions were used in order to constantly update the output based on the user input. For an easy data frame manipulation, we used the dplyr library and the magrittr library, specifically the piping functionality. For all interactive map presentations, we used the Leaflet library as well as the mapview package to extend the Leaflet functionalities. For the *Analysis* tabs, we used the GGPlot and the Dygraph package as well as the Grid Extra package that allows to define GGPlot themes.

## 6. Conclusions

During the project we performed data abstraction to optimally adjust the data to the tasks at hand, we chose a visualization layout that enables a quick and effortless navigation through all provided sections and evaluated encoding options with regards to effectiveness and efficiency.

The resulting Data Visualization application not only supports interactive data exploration for stakeholders of Bike Sharing providers, but can also facilitate the usage of such services for the end-user. Through multiple interactive elements, the user is able to thoroughly navigate through the data in an intuitive map-environment. Additionally, multiple reactive statistical analyses of trip and customer information allow for an improved understanding of the customer base and their usage behavior for an improved targeting of customer groups.

Challenges we encountered during this project were specific limitations of the standard R libraries, such as the limitations to certain widget types, with more advanced UI elements usually requiring a much more advanced integration of externally created packages. Furthermore, a user log-in for both user groups could only be simulated in our project.

Further work building on this project could integrate real-time bike sharing data using web-scraping techniques, in order to access information on current bike availability and station capacity, which would be particularly helpful for the user-facing application. Furthermore, the current prototype could be transformed into a real web-application, enabling a log-in option for both customers and company stakeholders. In production, the user's current position could be estimated through GPS information provided by the PC or smartphone and distances could be provided in a more realistic way, e.g. integrating with Google Maps and calculating the most efficient route between two stations

# Bibliography


1. Munzner, Tamara: “Visualization Analysis & Design” (2014)
2. “Bike Sharing Demand”, *Kaggle* (2014).
   Online at: kaggle.com/c/bike-sharing-demand
3. Dygraphs for R Documentation.
   Online at: https://rstudio.github.io/dygraphs/shiny.html
4. Shiny Bootstrap Themes.
   Online at: https://rstudio.github.io/shinythemes/
5. Richter, Felix: “The Global Rise of Bike-Sharing”, *statista* (2018).
   Online at: https://www.statista.com/chart/13483/bike-sharing-programs/
6. “Documentation of the General Bikeshare Feed Specification”, *github* (2018).
   Online at: https://github.com/NABSA/gbfs
7. Akella, Mamata: “Visualize 2015 Urban Populations with Proportional Symbols”, *carto* (2016).
   Online at: https://carto.com/blog/proportional-symbol-maps/
8. “Dashboard Visualization Options”, *Kyubit BusinessIntelligence.*
   Online at: https://www.kyubit.com/Dashboard-Visualization-Options


**GitHub repository:** https://github.com/joanalorenz/DV-App-Group-7