**Data Preparation Plan**

We are using datasets of bikeshare ride data from four United States cities: New York City (NYC), NY, Washington, D.C. (DC), Chicago, Illinois, and Los Angeles (LA), California. The specific parameters of the datasets vary, but all the sets include bike ID number, trip duration (in seconds/minutes), start and stop stations (latitude and longitude data present in few), date, and time for each individual ride. We chose to use 2018 Quarter 2 data as it was available across all four cities. The bikeshare data in all cases comes directly from the bikeshare entity, listed below by city:

Chicago: Divvy

LA: Metro Bike Share

NYC: CitiBike

DC: Capital Bikeshare

**Chicago**

City of Chicago Divvy (2018). 2018 Q2 Data [comma separated values file]. Retrieved from <https://www.divvybikes.com/system-data>. Last accessed 9 November 2018.

**LA**

Metro Bike Share (2018). 2018 Q2 [comma separated values file]. Retrieved from <https://bikeshare.metro.net/about/data/>. Last accessed 9 November 2018.

**NYC**

Citi Bike (2018). 201804-citibike-tripdata [comma separated values file]. Retrieved from <https://s3.amazonaws.com/tripdata/index.html>. Last accessed 9 November 2018.

Citi Bike (2018). 201805-citibike-tripdata [comma separated values file]. Retrieved from <https://s3.amazonaws.com/tripdata/index.html>. Last accessed 9 November 2018.

Citi Bike (2018). 201806-citibike-tripdata [comma separated values file]. Retrieved from <https://s3.amazonaws.com/tripdata/index.html>. Last accessed 9 November 2018.

**DC**

Capital Bikeshare (2018). 201804-capitalbikeshare-tripdata [comma separated values file]. Retrieved from <https://s3.amazonaws.com/capitalbikeshare-data/index.html>. Last accessed 9 November 2018.

Capital Bikeshare (2018). 201805-capitalbikeshare-tripdata [comma separated values file]. Retrieved from <https://s3.amazonaws.com/capitalbikeshare-data/index.html>. Last accessed 9 November 2018.

Capital Bikeshare (2018). 201806-capitalbikeshare-tripdata [comma separated values file]. Retrieved from <https://s3.amazonaws.com/capitalbikeshare-data/index.html>. Last accessed 9 November 2018.

Chicago, NYC, and DC are all managed by an international company called Motivate Inc. In all four cities, the bikeshare programs are integrated into the municipal transportation agency. In all cases, despite somewhat varying management, the bike share data has been made publicly available. Language on the NYC website is typical of all four, stating: “We're happy to provide the data to help you discover the answers…We invite developers, engineers, statisticians, artists, academics and other interested members of the public to use the data we provide for analysis, development, visualization and whatever else moves you” (“System” 2018). Each website also provides a data license agreement. Since Chicago, DC, and NYC share the same corporate management, the license for all three is the same. The agreements grant “non-exclusive, royalty-free, limited, perpetual license to access, reproduce, analyze, copy, modify, distribute in your product or service and use the Data for any lawful purpose” (“Data License” 2018). In addition, the agreement prohibits attempts to use the data to identify individuals correlate the data or any particular person (“Data License” 2018). Our academic use of the data falls into the terms of the agreement.

The LA bikeshare program did not have a data license agreement online. The closest to a data agreement is an invitation to “Toss it, flip it, turn it inside out, combine it with your own or someone else's data…delve into the data and reveal the hidden gems of understanding that will allow us to make Metro Bike Share the best it can be” (“Data” 2018). Based on this, we feel comfortable that our use of the data falls within the limits of the agreement.

Each .csv file contains plain language column titles (tripduration, starttime, etc.) which help to identify the information in the documents. In addition, each website provides a list of all the data included in the datasets. However, the LA website is the only one that offers a traditional dictionary in which each column ID has a description. The other three websites offer a list of the data parameters without explicitly stating the corresponding column ID. All four websites also provide background information about what bikeshare programs are and some provide different reports such as member surveys and monthly operating reports.

Most issues stem from the fact that four different entities have collected and distributed the datasets. For example, LA and Chicago do not include birth and gender data, but NYC and DC do. For the parameters that are the same, they appear in different orders across the four cities. For example, bikeID is the 8th column for NYC but it comes third for Chicago. None of the ride data contains blanks or errors, but some of the rider data (birth year and gender) includes blanks and unreasonable data. For example, in the NYC set, there are birth years starting from 1885. It is highly unlikely that somebody born in 1885 is still alive, and even less likely that someone of 133 years is using a bikeshare program. The blanks and unreasonable data will be excluded from analysis, but those records can still be analyzed for the automatically-generated data, such as trip duration.

Considering the remediation of the data, since all our datasets are taken from official websites of bike sharing organizations who have put up clean data, we didn’t need to allot a lot of time for data cleaning. A major chunk of time was taken by data standardization to make all the different datasets compatible for future relationship plotting and analysis. We have kept the ‘Trip\_id’ to maintain unique identifiers of all trips taken. All Datasets are standardized to have ‘Trip\_duration’, ‘Start\_station’, ‘End\_station’, ‘Trip\_Start\_time’, ‘Trip\_end\_time’, ‘Trip\_id’ etc. To maintain uniformity across, we also converted the trip duration from seconds or hours to minutes. In a few datasets, we have additional fields of birth year and age which had faulty values such as ‘Birth\_year’ had a value of ‘1800’, which is humanly impossible. To combat this issue, we found the average life expectancy of an average American (78 years) and took values greater than 70 years or ‘Birth\_year’ less than 1948 and rounded it off to 70 years and birth year to 1948 for uniformity on assumption basis.

We began the process by analyzing our database variables and different values since we had a multiple We began the process of cleaning by checking if we had null or invalid values for crucial dataset field such as ‘Trip\_duration’ and ‘Station\_from’ and ‘Station\_to’, which would dominate our analysis. Since we didn’t find any null or invalid values, we moved on to the standardization of the trip duration. In the Chicago, Washington DC and NYC dataset the trip duration was in seconds, in LA dataset it was in minutes; We converted the seconds data to minutes for better clarity and uniformity in excel by dividing the values by 60, which was later formatted in time format.

For LA specifically, station street address isn’t present in the main trip dataset and a separate dataset is available with specific station ids corresponding to the street address with other data. This data was merged with the trip data for both start station and end station using the merge function and the unnecessary columns are then eliminated using the subset function. Following are commands used:

LA\_final <- merge(LA, LA\_Stations, by.x = "start\_station", by.y = "Station\_ID")

LAFinal <-subset(LAFinal, select = -c(Go\_live\_date, Region, Status))

'Start Station' <- colnames(LAFinal)[colnames(LAFinal)=='Station\_Name']

LAData <- merge(LAFinal, LA\_Stations, by.x = 'end\_station', by.y = 'Station\_ID')

LAData <- subset(LAData, select = -c(Go\_live\_date, Region, Status))

'End Station' <- colnames(LAData)[colnames(LAData)=='Station\_Name']

Considering the invalid values for extra fields such as ‘Birth\_year’ and ‘Age’, we deleted all values that suggested that the user was over 100 years old or born in 1800 and entering 1948 as birth year and 70 years as age for all participants suggesting age to be over 70 years.

**Works Cited:**

* “Data.” 2018. Metro Bike Share. Retrieved from <https://bikeshare.metro.net/about/data/>. Last accessed 9 November 2018.
* “Data License Agreement.” 2018. Citi Bike. Retrieved from <https://www.citibikenyc.com/data-sharing-policy>. Last accessed 9 November 2018.
* “System Data.” 2018. Citi Bike. Retrieved from <https://www.citibikenyc.com/system-data>. Last accessed 9 November 2018.

Word Count: 1083