

2. Data preparation

Description of all data sources used

The fictional company Cyclistic's historical trip data is used.

Years 2019 and 2020 is compared, using the zip files *Divvy_2019_Q1.csv* and *Divvy_2020_Q1.csv*.

The files are separated into

Documentation of any cleaning or manipulation of data

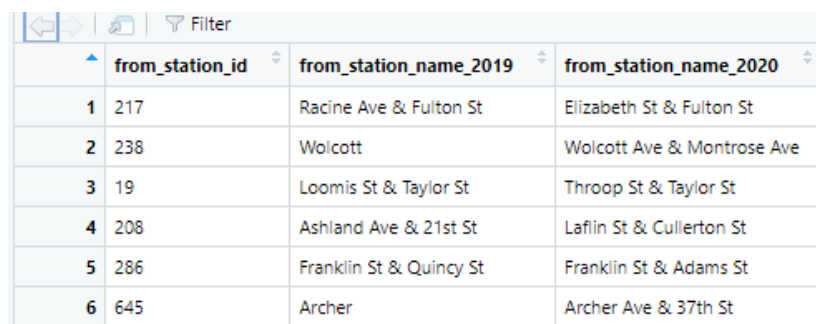
1. Columns that are useful and comparable are picked, and the column names are unified.

- Several columns are removed due to their lack of usefulness:
- 'rideable_type' column contains only a single unique value "docked_bike" and is deemed unnecessary.
- Unique identifier columns 'trip_id', 'bikeid', and 'ride_id' are incompatible and serve no purpose.
- 'gender' and 'birthyear' columns appear useful but are removed because they are not present in *Divvy_Trips_2020_Q1*.
- 'start_lat', 'start_lng', 'end_lat', 'end_lng' columns appear useful but are removed because they are not present in *Divvy_Trips_2019_Q1*.
- The column names differ between files. They are standardized for consistency.
- 'tripduration' doesn't exist for *Divvy_Trips_2020_Q1*. Therefore, this column is calculated using the difference between 'end_time' and 'start_time'.
- 'tripduration' does exist for *Divvy_Trips_2019_Q1*. However, we do not know the unit of the values in the column. Therefore, it is calculated again using 'end_time' and 'start_time'.
- Data types of 'from_station_id' and 'to_station_id' are changed from INT to CHR, since the values are nominal.

Divvy_Trips_2019_Q1	Divvy_Trips_2020_Q1	Divvy_Trips_2019Q1_2020Q1	Data Types	Data Format
trip_id	ride_id	start_time	POSIXct	ymd_hms
start_time	rideable_type	end_time	POSIXct	ymd_hms
end_time	started_at	tripduration	num	seconds
bikeid	ended_at	from_station_id	chr	
tripduration	start_station_name	from_station_name	chr	
from_station_id	start_station_id	to_station_id	chr	
from_station_name	end_station_name	to_station_name	chr	

Divvy_Trips_2019_Q1	Divvy_Trips_2020_Q1	Divvy_Trips_2019Q1_2020Q1	Data Types	Data Format
to_station_id	end_station_id	usertype	chr	Member, Casual
to_station_name	start_lat			
usertype	start_lng			
gender	end_lat			
birthyear	end_lng			
	member_casual			

- Additional data wrangling operations are performed, with detailed steps provided in the R Markdown file *ANALYZE_casestudy_01_cyclisticbikeshare_2019to2020*.
- This point is explicitly mentioned because these changes were made at the analyst's discretion without consulting the data owner. (As this is a capstone project, there's no one to contact for clarification on the metadata's meaning.)
- When checking for station_id and station_name consistency across and within the datasets, discrepancies remain even after removing station names containing parentheses (e.g., "(*)" and "(temp)"). Except for the changes below, other values remain as-is pending further investigation into "to_station_id" and "to_station_name":
 - Wolcott → Wolcott Ave & Montrose Ave
 - Archer → Archer Ave & 37th St



	from_station_id	from_station_name_2019	from_station_name_2020
1	217	Racine Ave & Fulton St	Elizabeth St & Fulton St
2	238	Wolcott	Wolcott Ave & Montrose Ave
3	19	Loomis St & Taylor St	Throop St & Taylor St
4	208	Ashland Ave & 21st St	Lafin St & Cullerton St
5	286	Franklin St & Quincy St	Franklin St & Adams St
6	645	Archer	Archer Ave & 37th St

Image: Before changes to Wolcott and Archer are made



	from_station_id	from_station_name_2019	from_station_name_2020
2	19	Loomis St & Taylor St	Throop St & Taylor St
3	208	Ashland Ave & 21st St	Lafin St & Cullerton St
1	217	Racine Ave & Fulton St	Elizabeth St & Fulton St
4	286	Franklin St & Quincy St	Franklin St & Adams St

Image: Differences remaining after the change

- The same steps are performed on "from_station_id" and "from_station_name" for 2019 and 2020 datasets. The results are consistent with the "to" cases, as shown below. Hence, these

changes are made:

- a. Wolcott → Wolcott Ave & Montrose Ave
- b. Archer → Archer Ave & 37th St

	to_station_id	to_station_name_2019	to_station_name_2020
1	19	Loomis St & Taylor St	Throop St & Taylor St
2	217	Racine Ave & Fulton St	Elizabeth St & Fulton St
3	238	Wolcott	Wolcott Ave & Montrose Ave
4	286	Franklin St & Quincy St	Franklin St & Adams St
5	208	Ashland Ave & 21st St	Lafin St & Cullerton St
6	645	Archer	Archer Ave & 37th St

Image: Before changes to Wolcott and Archer are made

	to_station_id	to_station_name_2019	to_station_name_2020
1	19	Loomis St & Taylor St	Throop St & Taylor St
4	208	Ashland Ave & 21st St	Lafin St & Cullerton St
2	217	Racine Ave & Fulton St	Elizabeth St & Fulton St
3	286	Franklin St & Quincy St	Franklin St & Adams St

Image: Differences remaining after the change

6. For the purpose of this analysis, 2019 station names for station_id values 19, 208, 217, and 286 are maintained for 2020 station names.