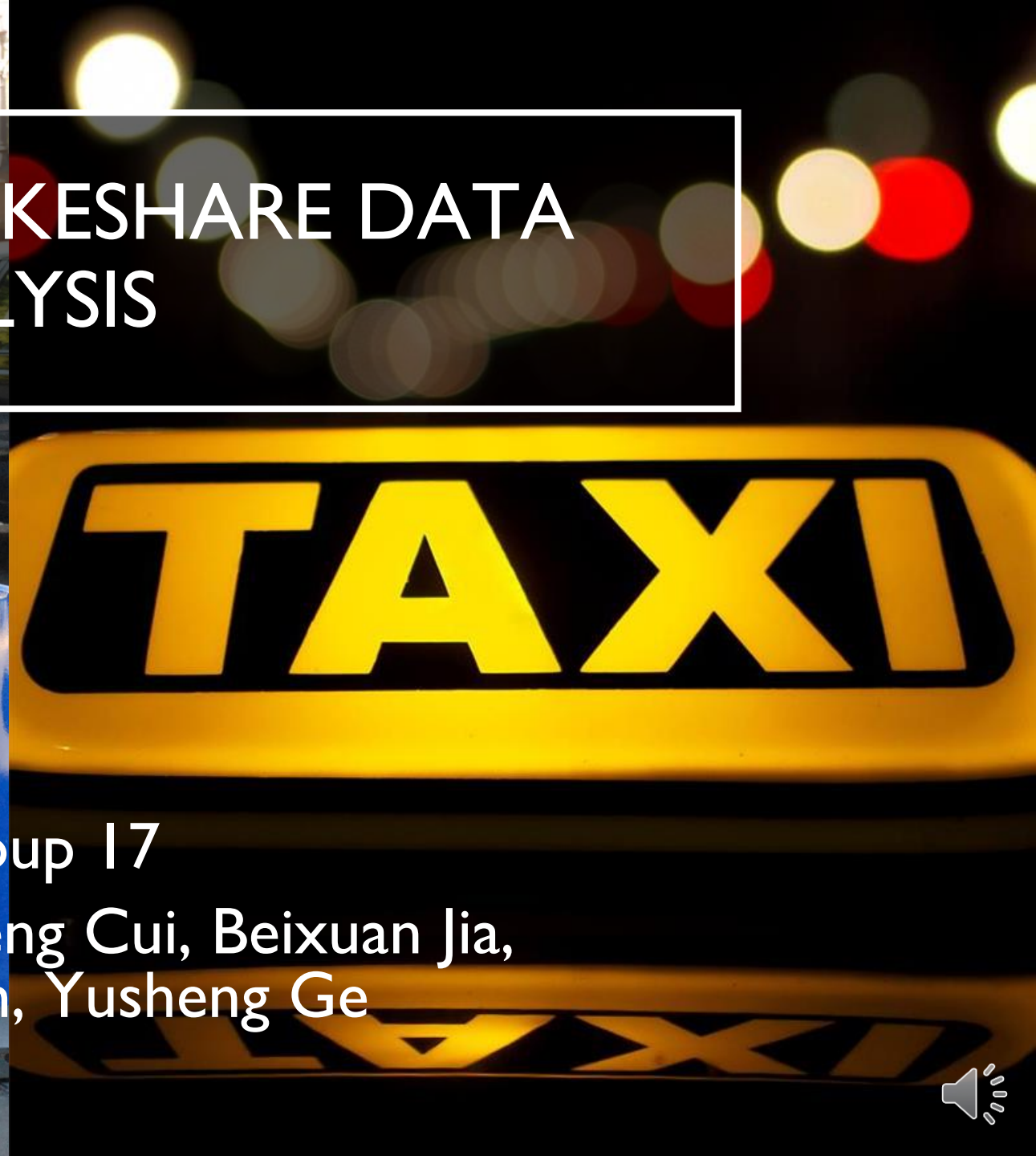
DC TAXI AND BIKESHARE DATA ANALYSIS



Group 17

Yifan Dang, Yupeng Cui, Beixuan Jia,
Huimin Han, Yusheng Ge

AGENDA

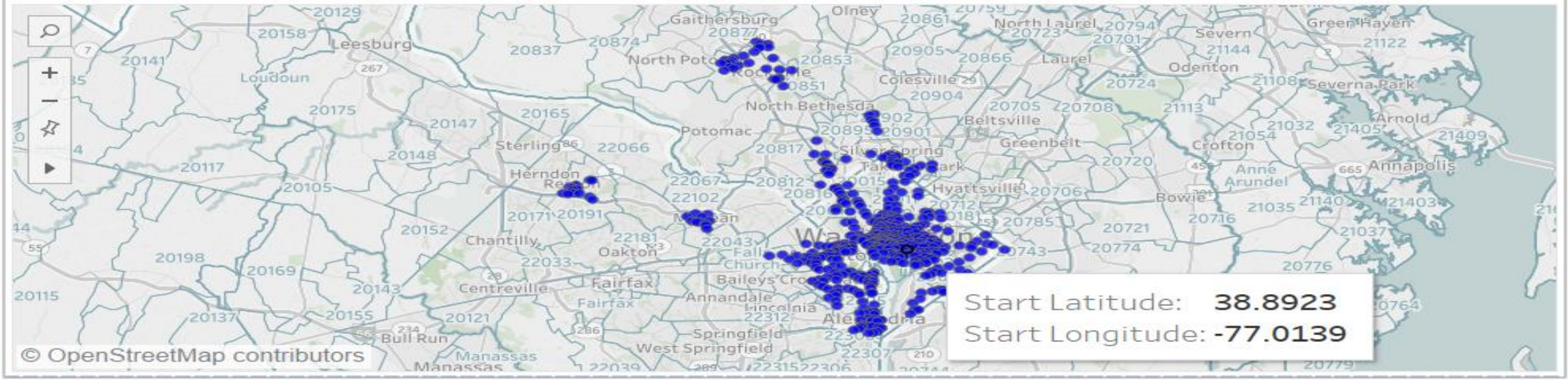
Data Preprocessing

- Used Bigquery to selected the key variables(e.g. zip codes, ID, Duration) by join the two table.
- Filtered out the zip codes within the DC areas, and kept the data only after 2017.
- Removed any abnormal values and numeric outliers by python (e.g. keep every input value within 3 standard deviation from mean)

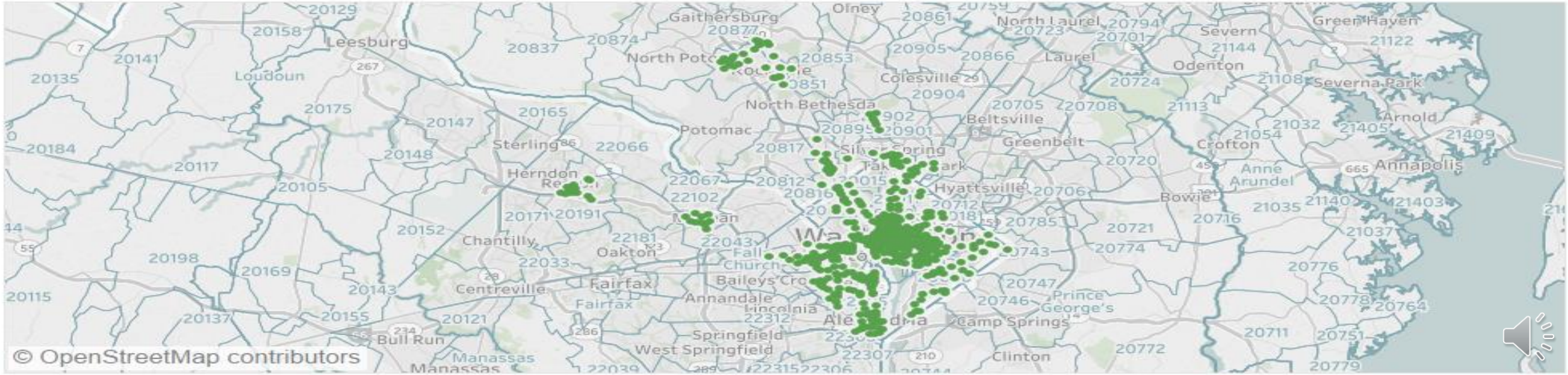
Data Visualization

- Plotted the geography distribution of both taxis and bikes in Tableau determine the characteristics of region where either one dominates.
- Find the relationships between average time spends, average distances and average money in each areas divided by zip code and generalizes interesting thoughts.

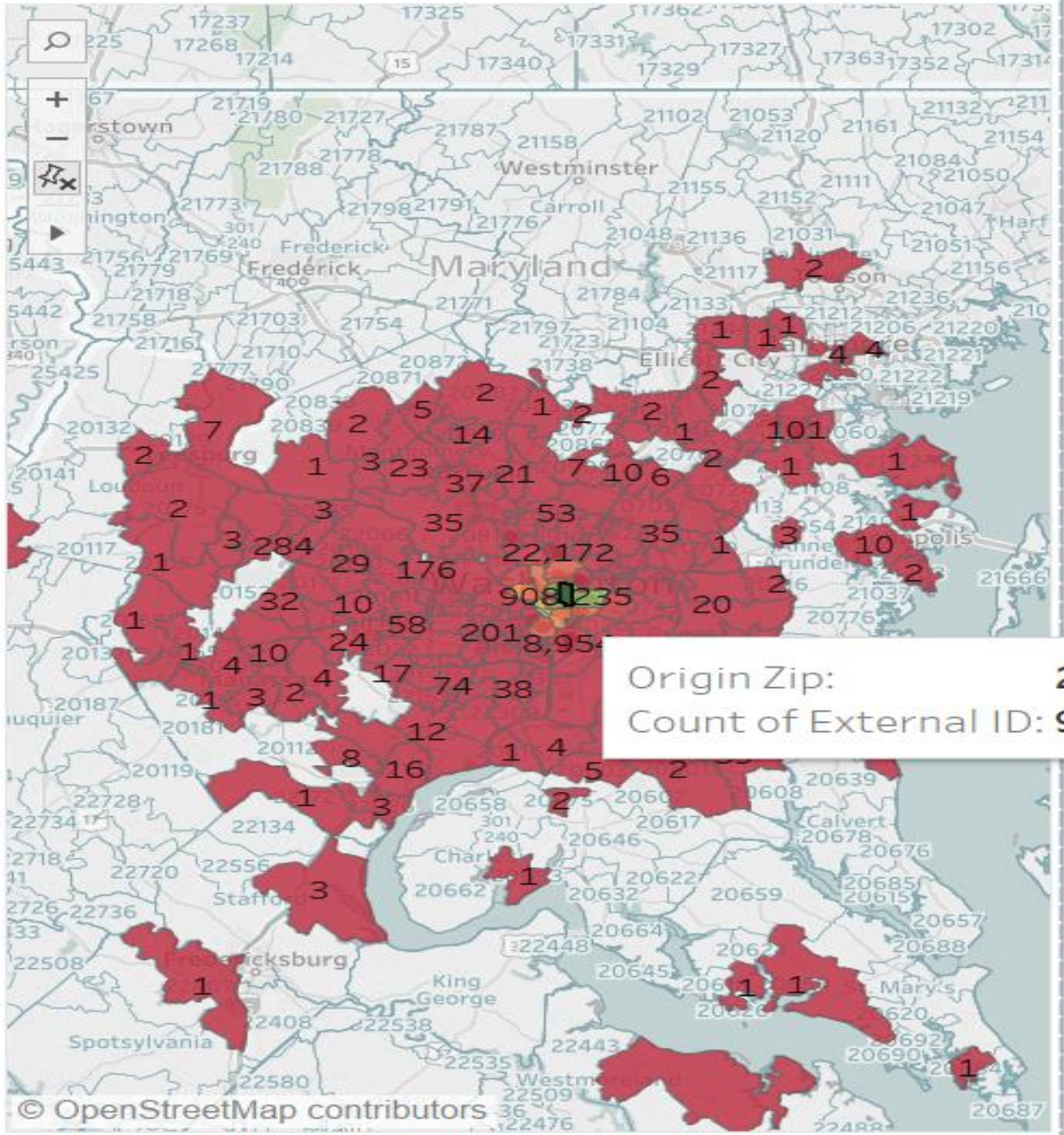
Bike Start Station



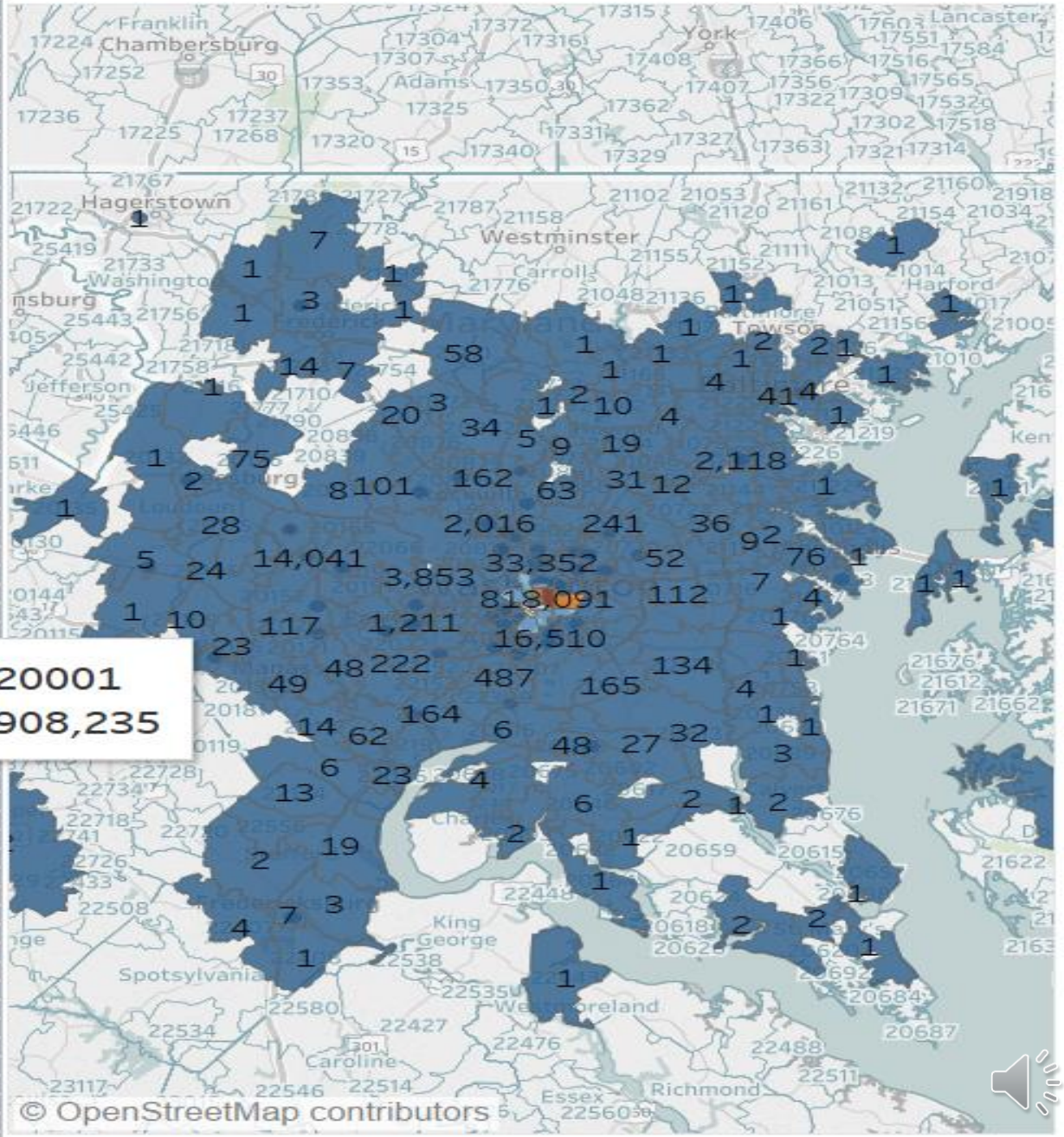
Bike End Station



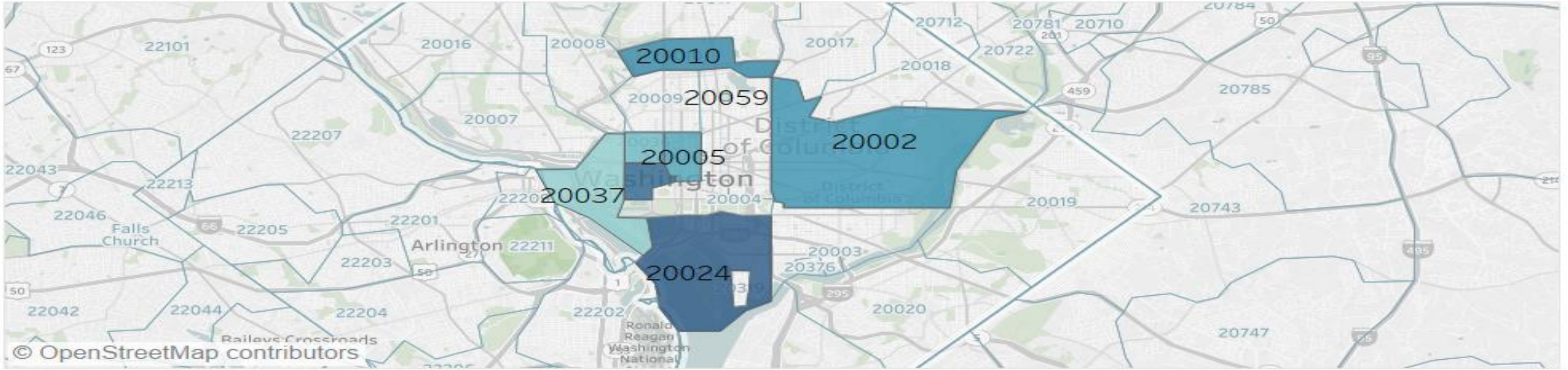
TaxiStartZip



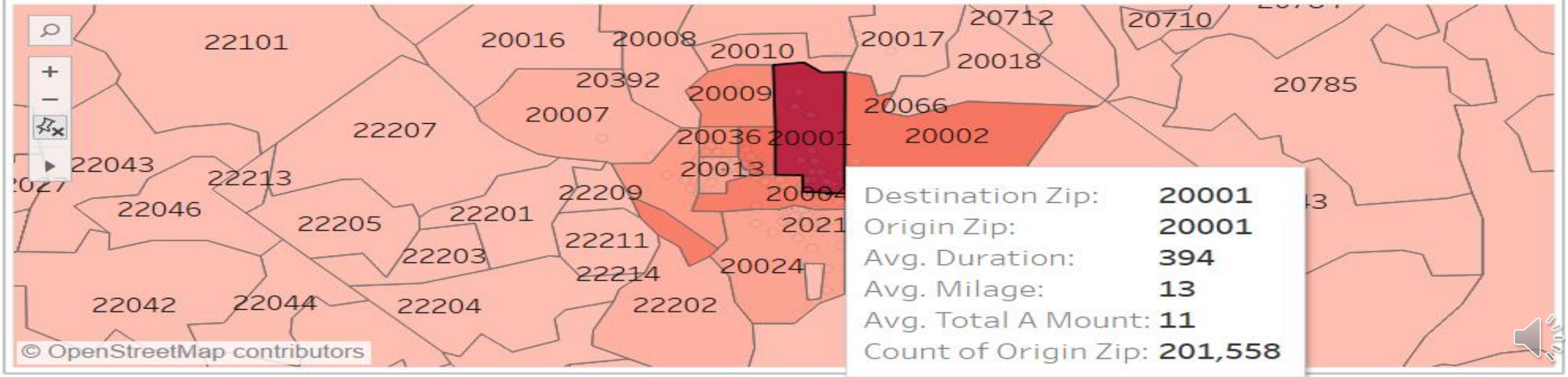
TaxiEndZip



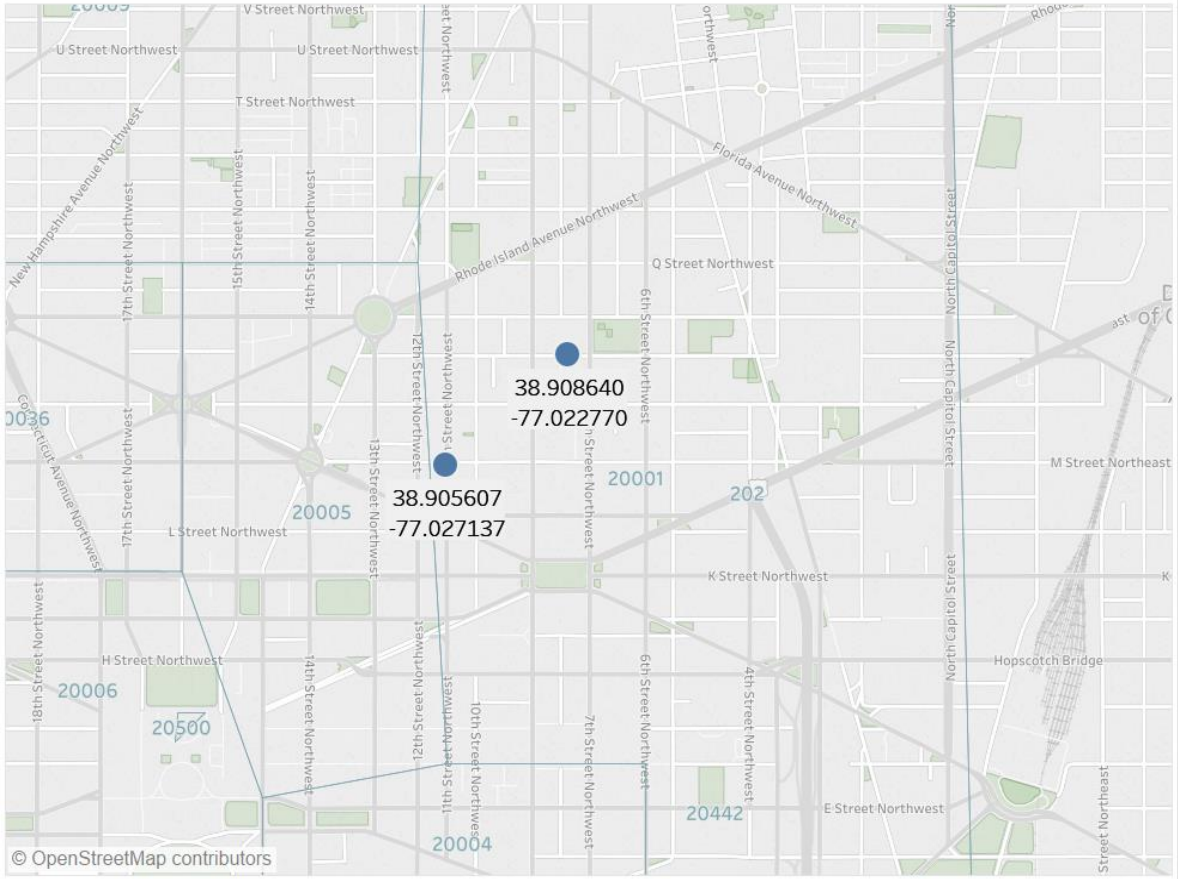
DC Bike Density



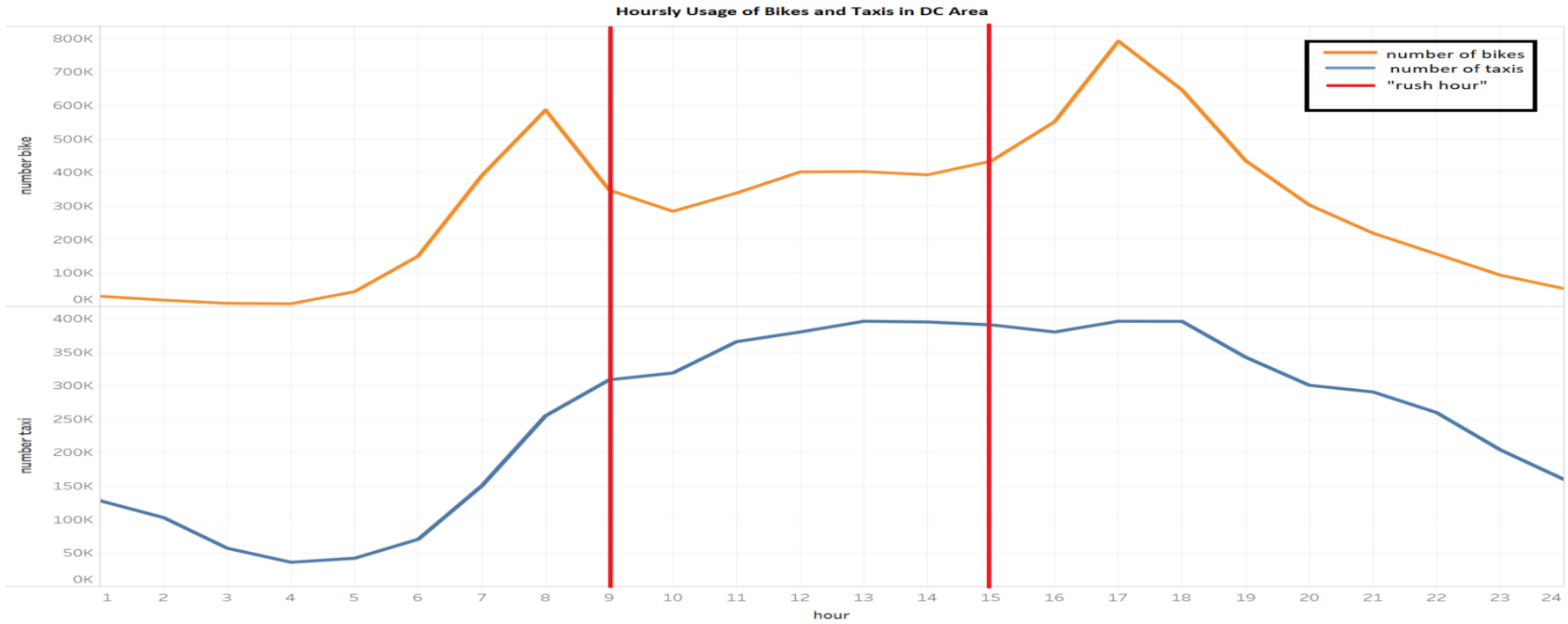
DC Taxi Density



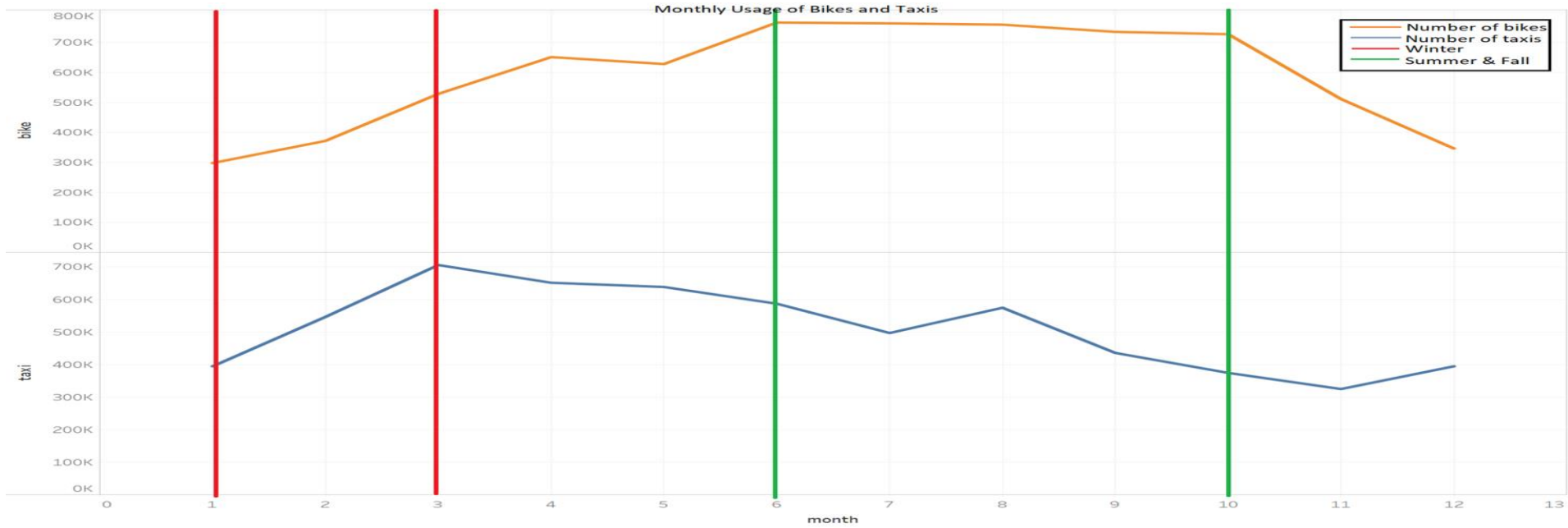
Bike Round Trip Stations (ONLY TWO STATION!)



Bike number	nd station numbe	tart station numb	Duration
W00009	31281	31266	377
W00009	31266	31281	336



The trends of number bike as an attribute and number taxi as an attribute for hour. The view is filtered on hour, which ranges from 1 to 24.



SUGGESTIONS

- **The Taxi Company should avoid the zip code areas which has higher bikers, such as the area 20024, focus on the areas where the bikes appear less.**
- **Avoiding the areas which mileage in the unit time is blow 90 percentages of distribution. Because those areas high not driver favor or not cost-efficiency.**
- **Pay attentions to the bike stations where the Universities and shopping center is nearby, Because people in those areas tend to use bike for round trips.**