Ford GoBike Data Analysis

Team members:

Baha Özşahin

Danilo Marinkovic

Mohammad Muttaqi

Goals of the project

- Trip analysis
- User demographics analysis
- Station analysis
- Bike utilization and maintenance
- Business insights and growth opportunities

Data Source

- The original dataset is from Kaggle (ref)
- A medium sized dataset (~500k rows)
- Complete ride history log of the greater San Francisco Bay area in the year of 2017.
- It also consists of various ride characteristics and riders characteristics.
  - For riders, gender and birth year
  - For the rides, duration and start, end stations.

TECHNOLOGIES USED

CODING

- Python
- PySpark

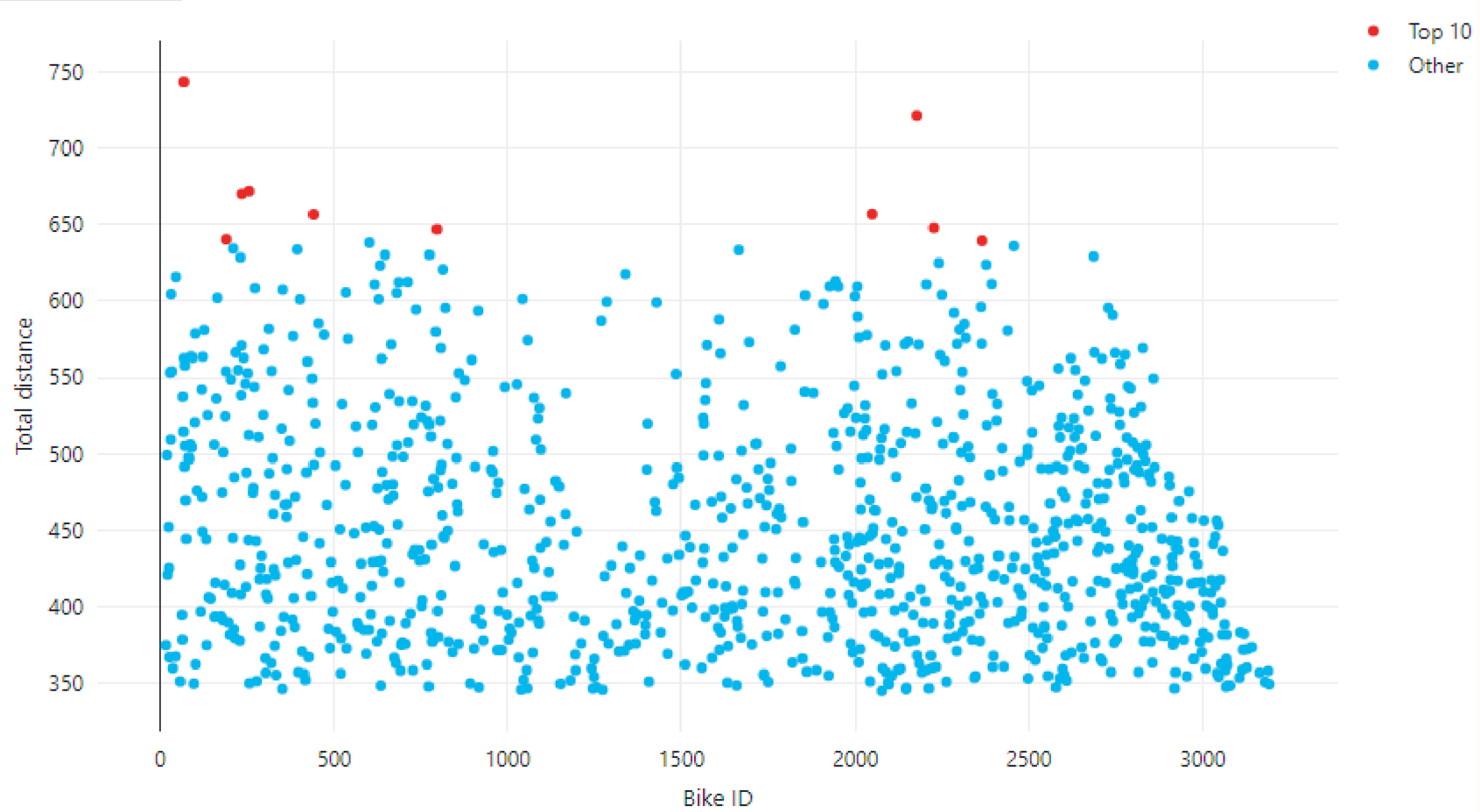
COLLABORATION

- Github
- Trello
- Discord
- Databricks
- Google Colab

VISUALIZATION

- Seaborn
- Matplotlib
- Databricks
- Canva

Calculating Distance

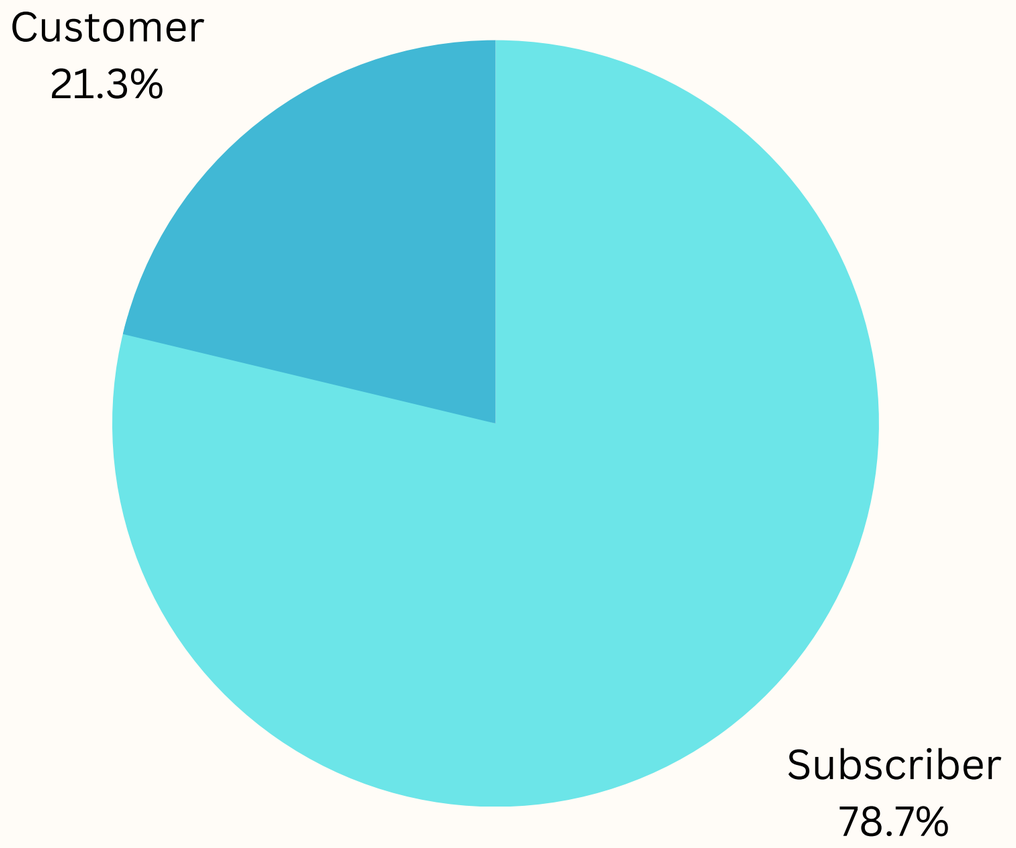


Bike Id	Distance
1842	68.2
2118	68.0
465	62.3
1210	62.3
2185	17.3
1349	17.1
780	17.1
1940	17.1
3027	17.1
27	16.0

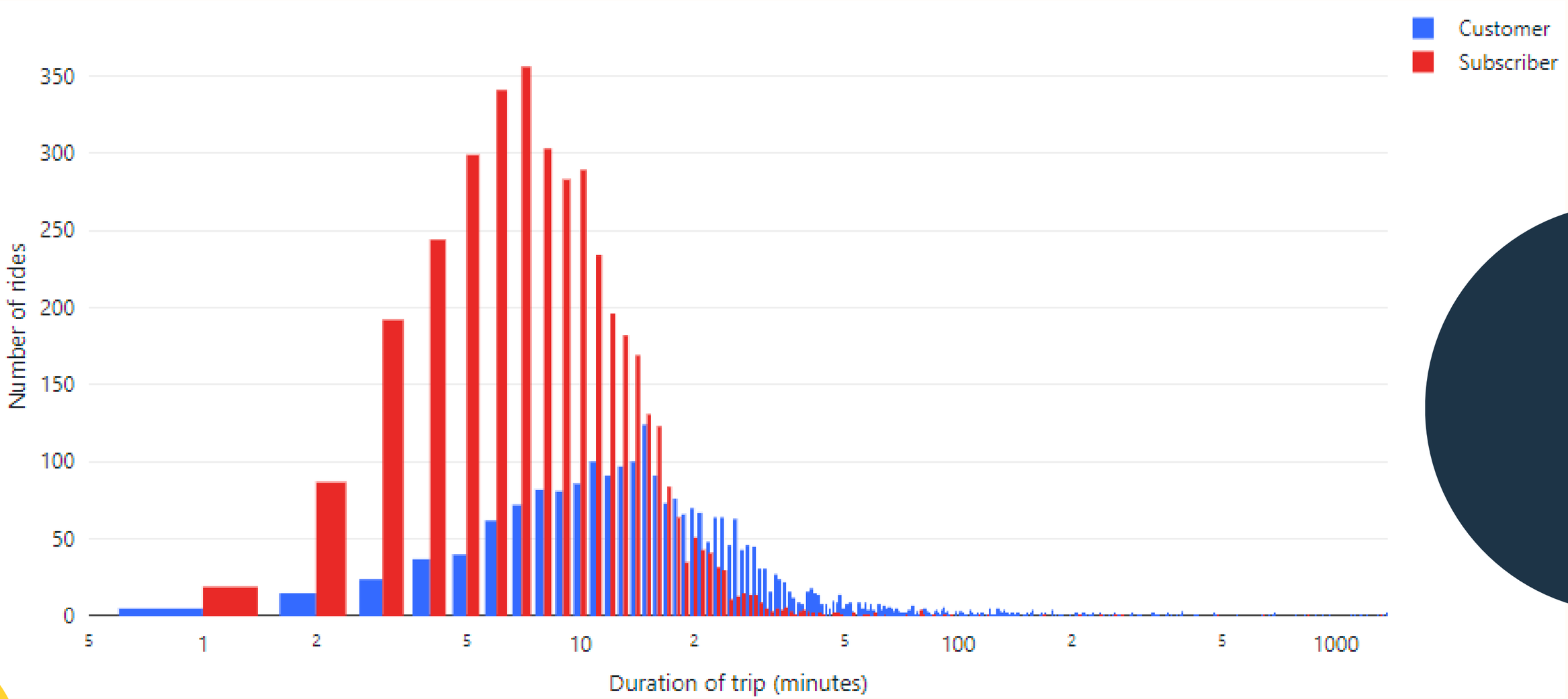
User Type Comparison

Almost 4/5 of the entire userbase consists of subscribers, though there is still a significant interest from non-subscribers.

Targeted advertising could improve the conversion rate of on-demand customers to subscribers.

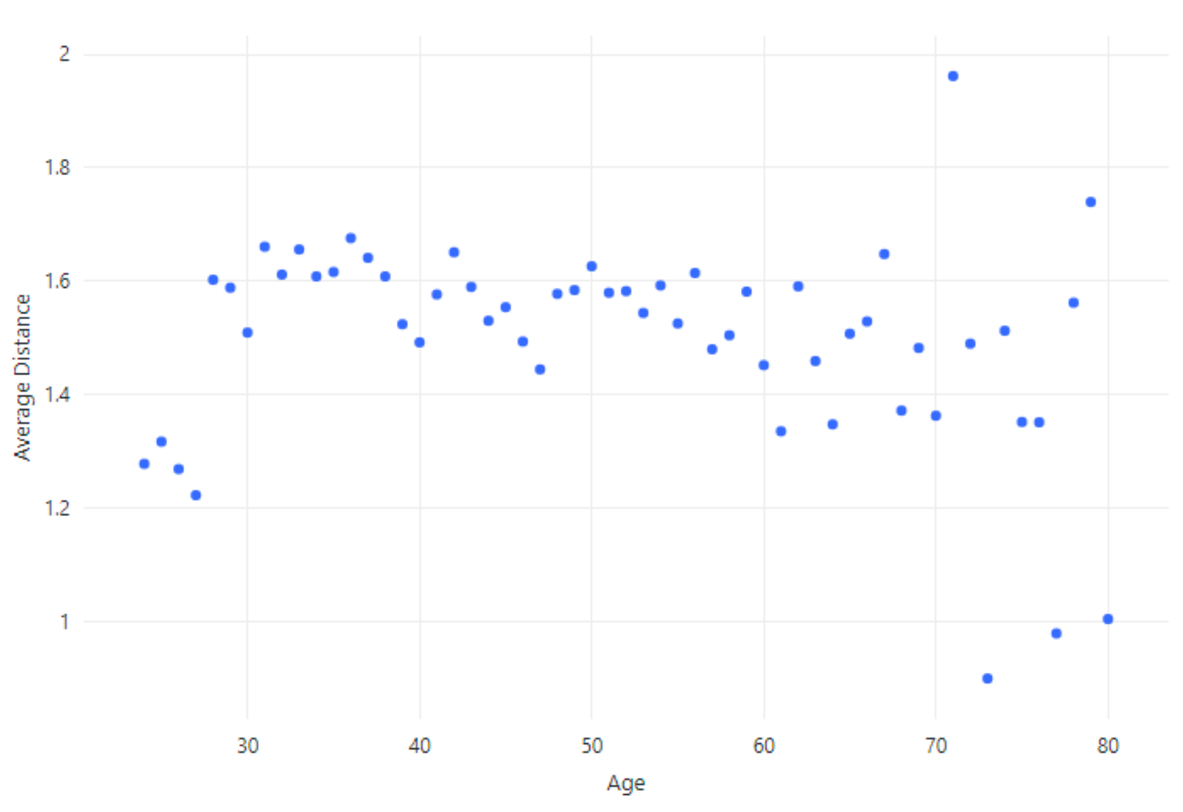


Trip duration VS user type

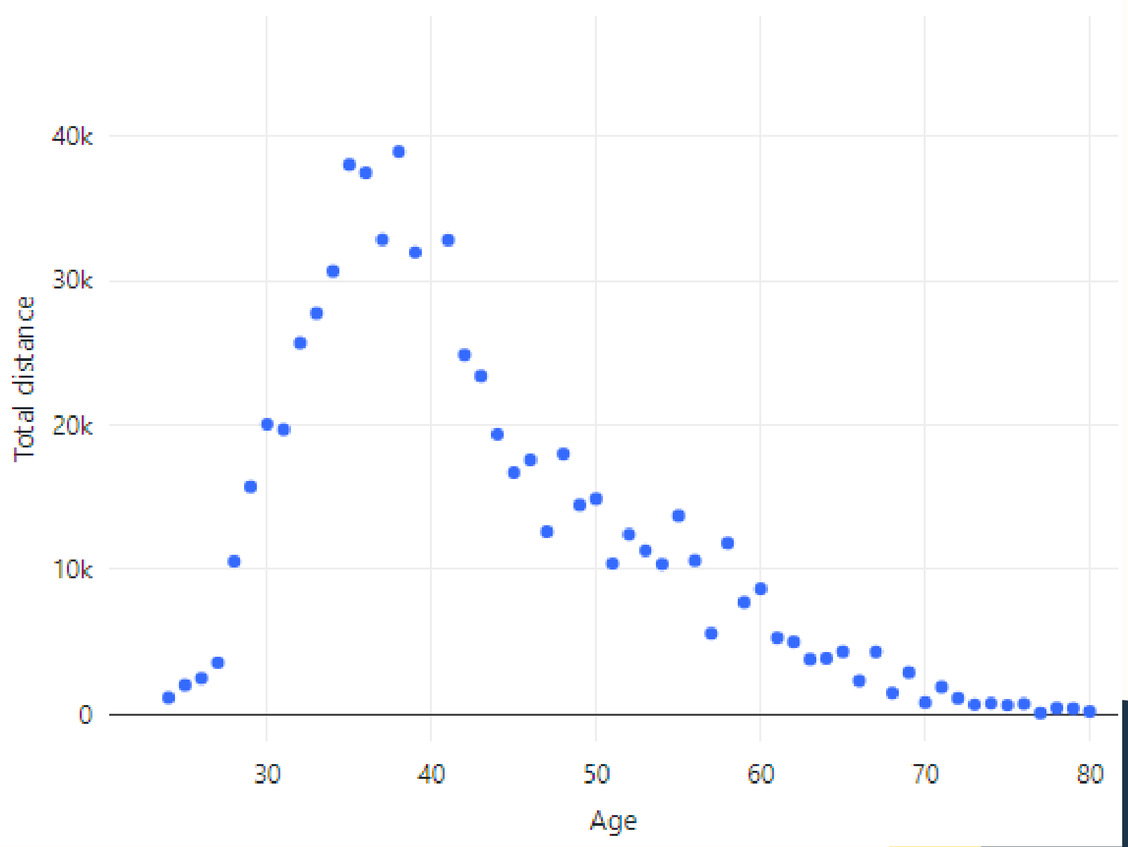


Relation between trip distance and age

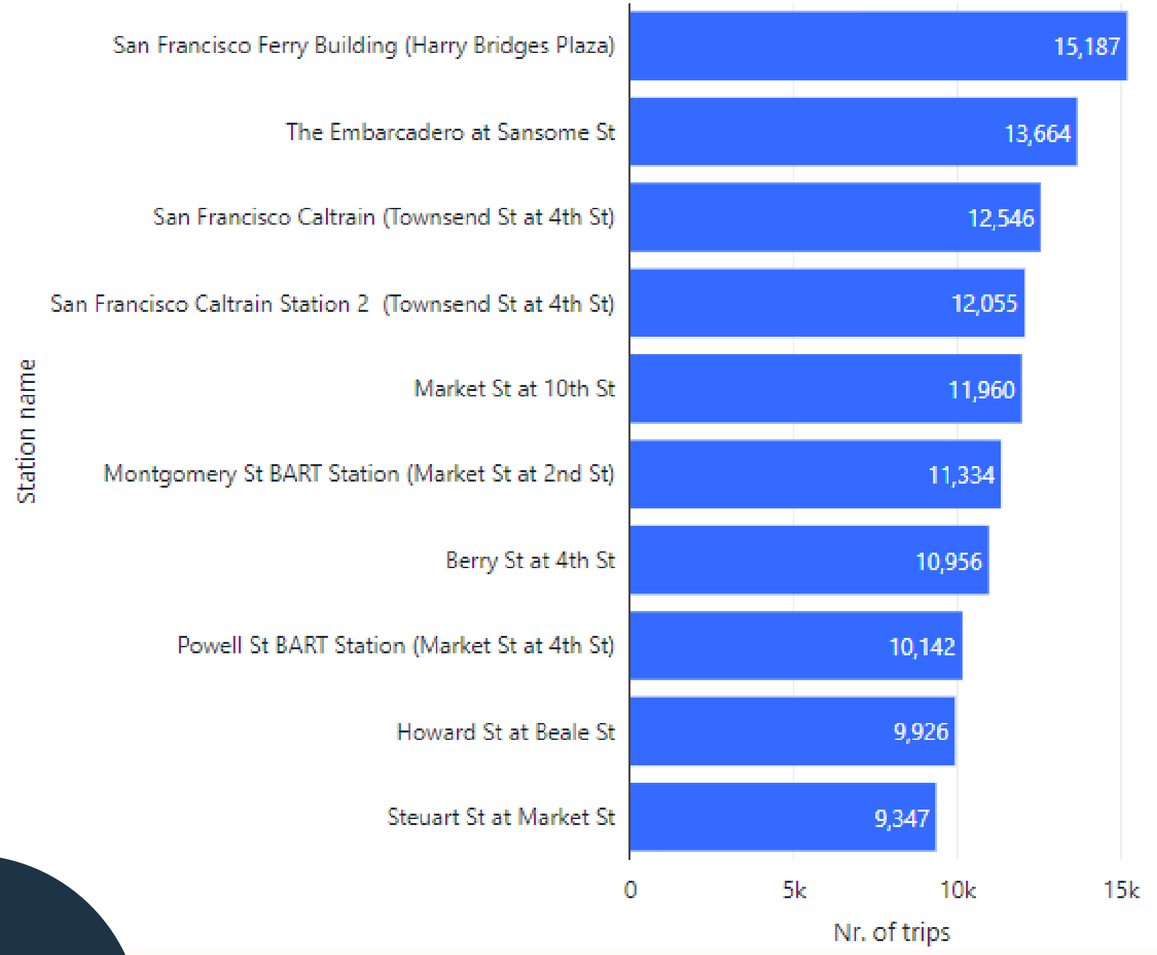
Age VS average trip distance

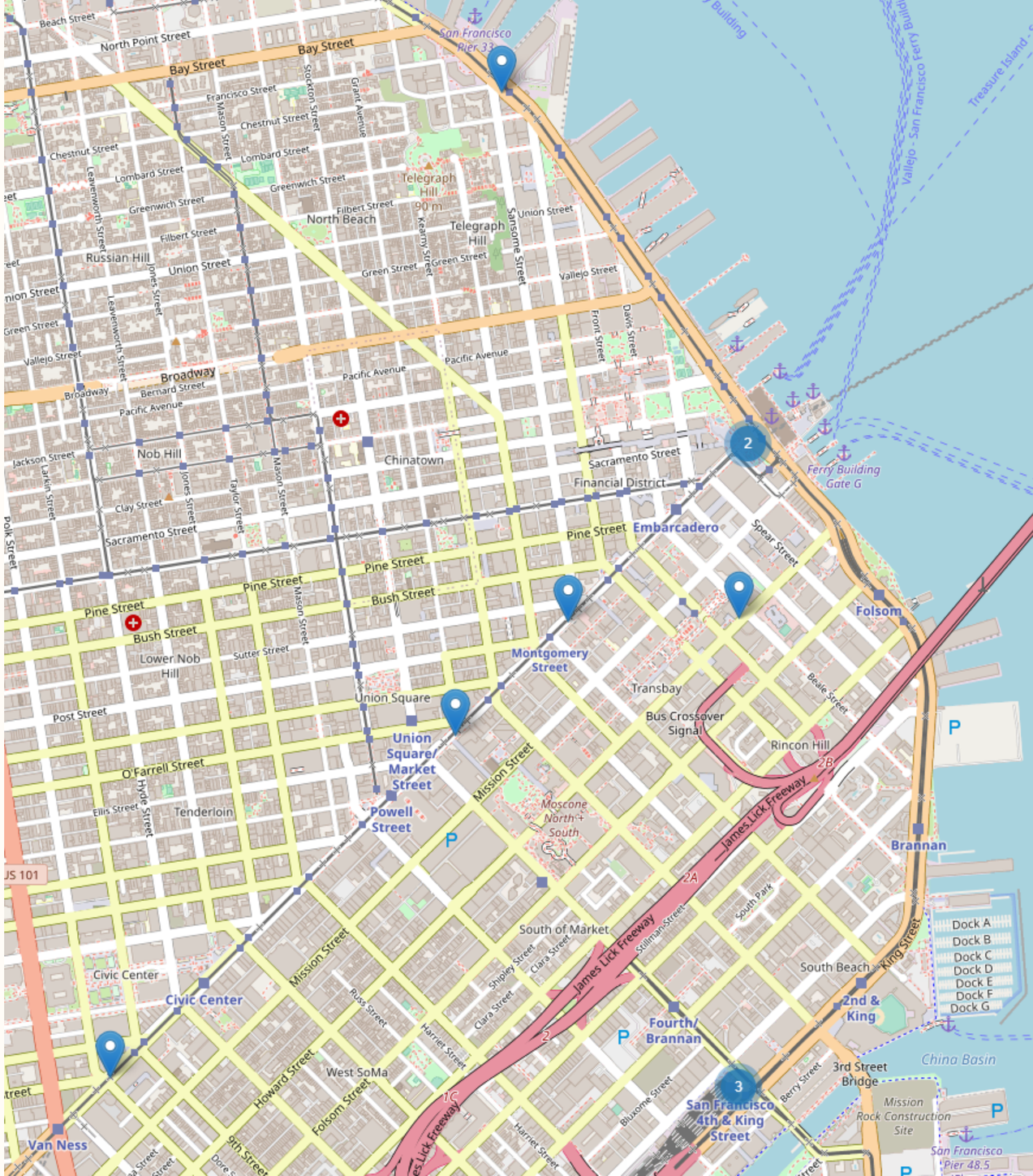


Age VS total distance

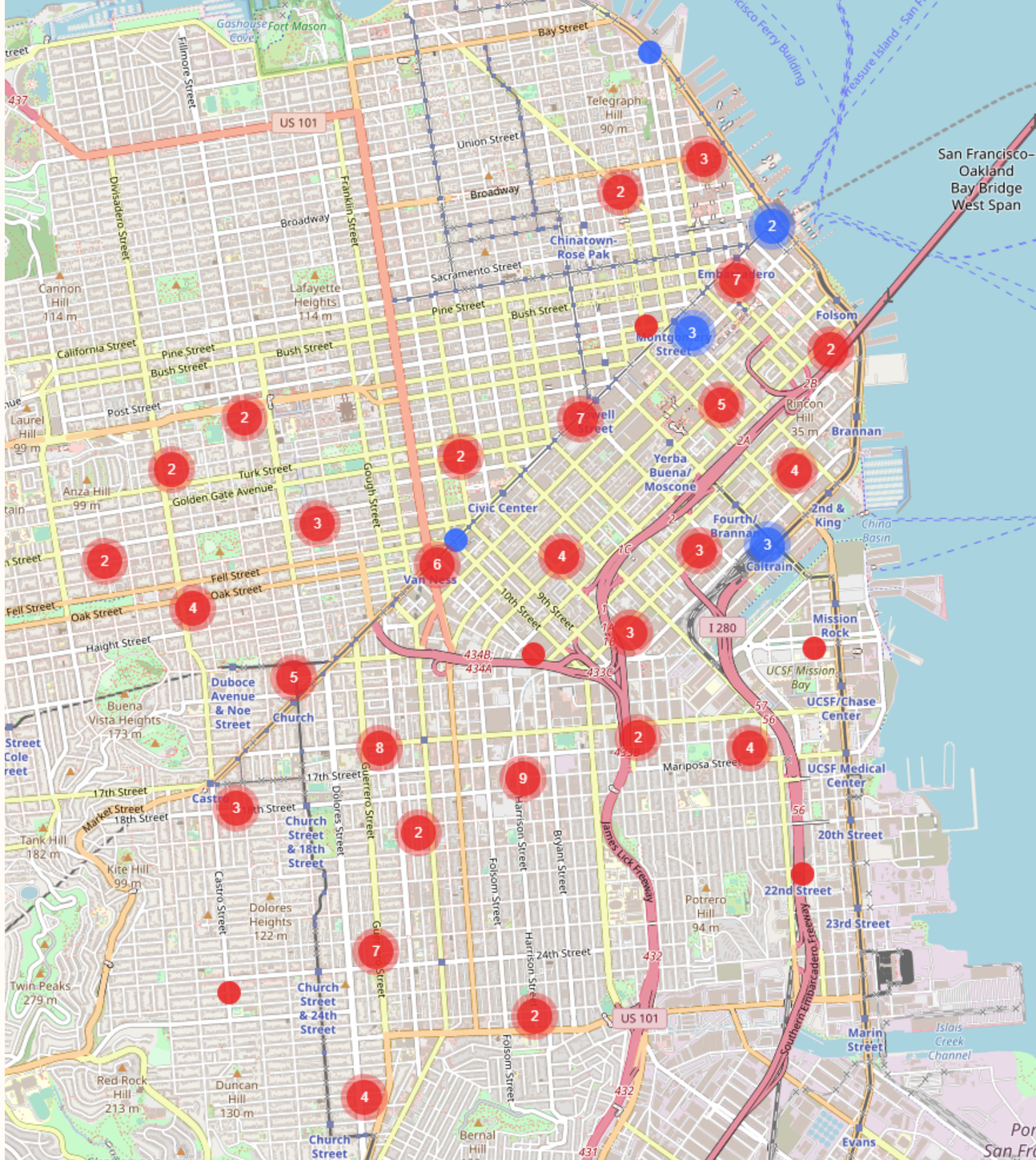


Most Used Stations

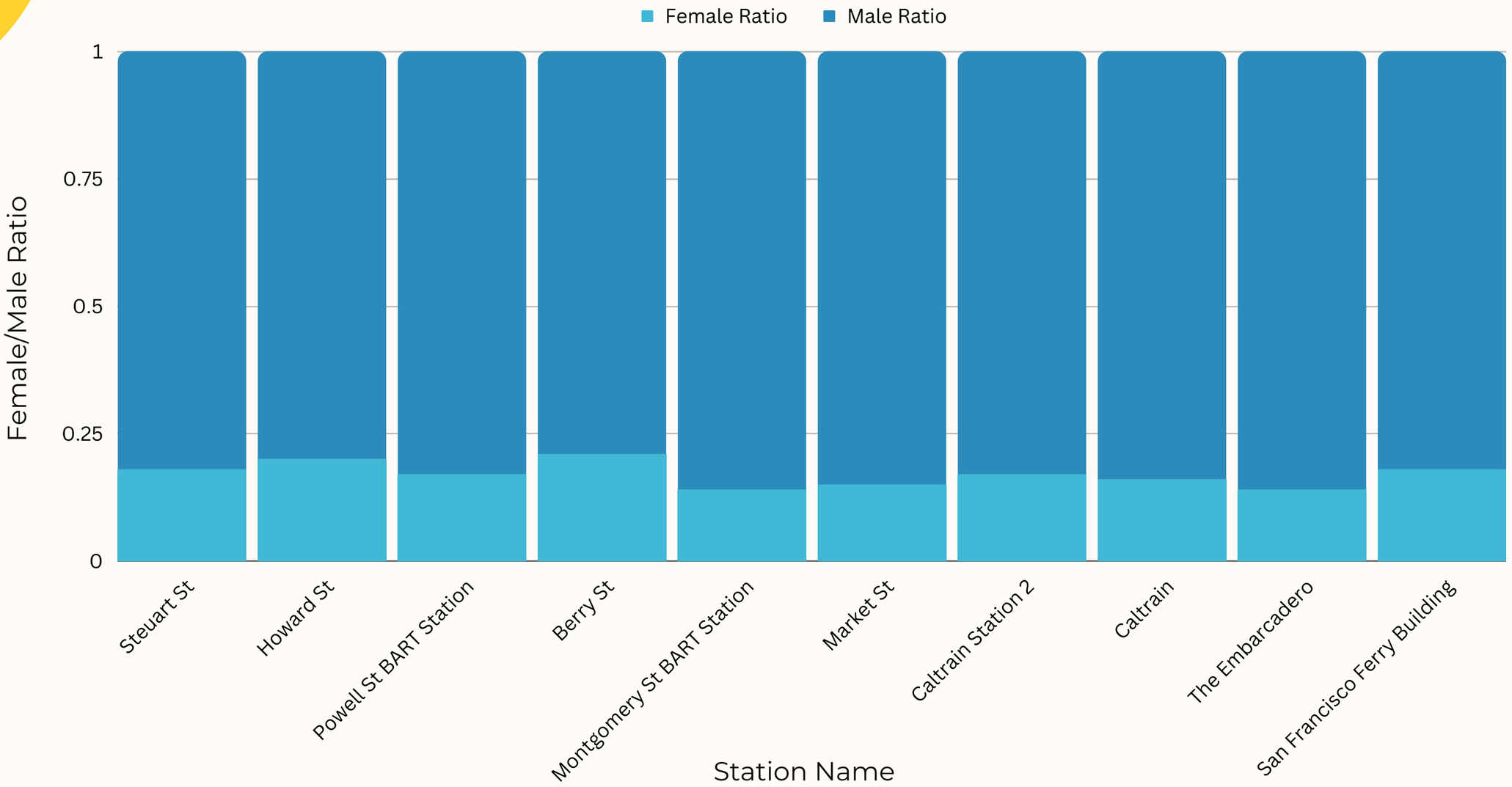




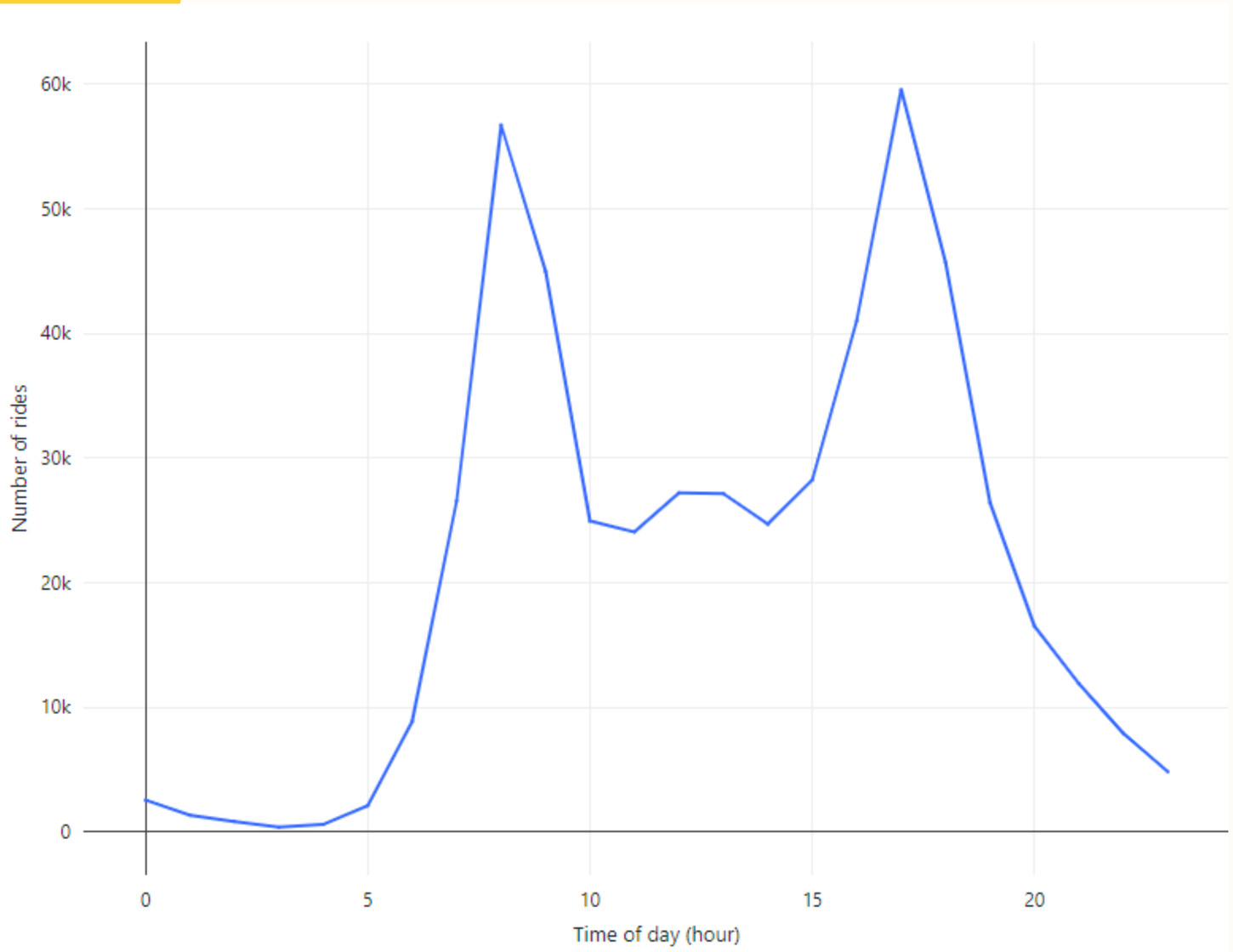
Most used stations VS the rest



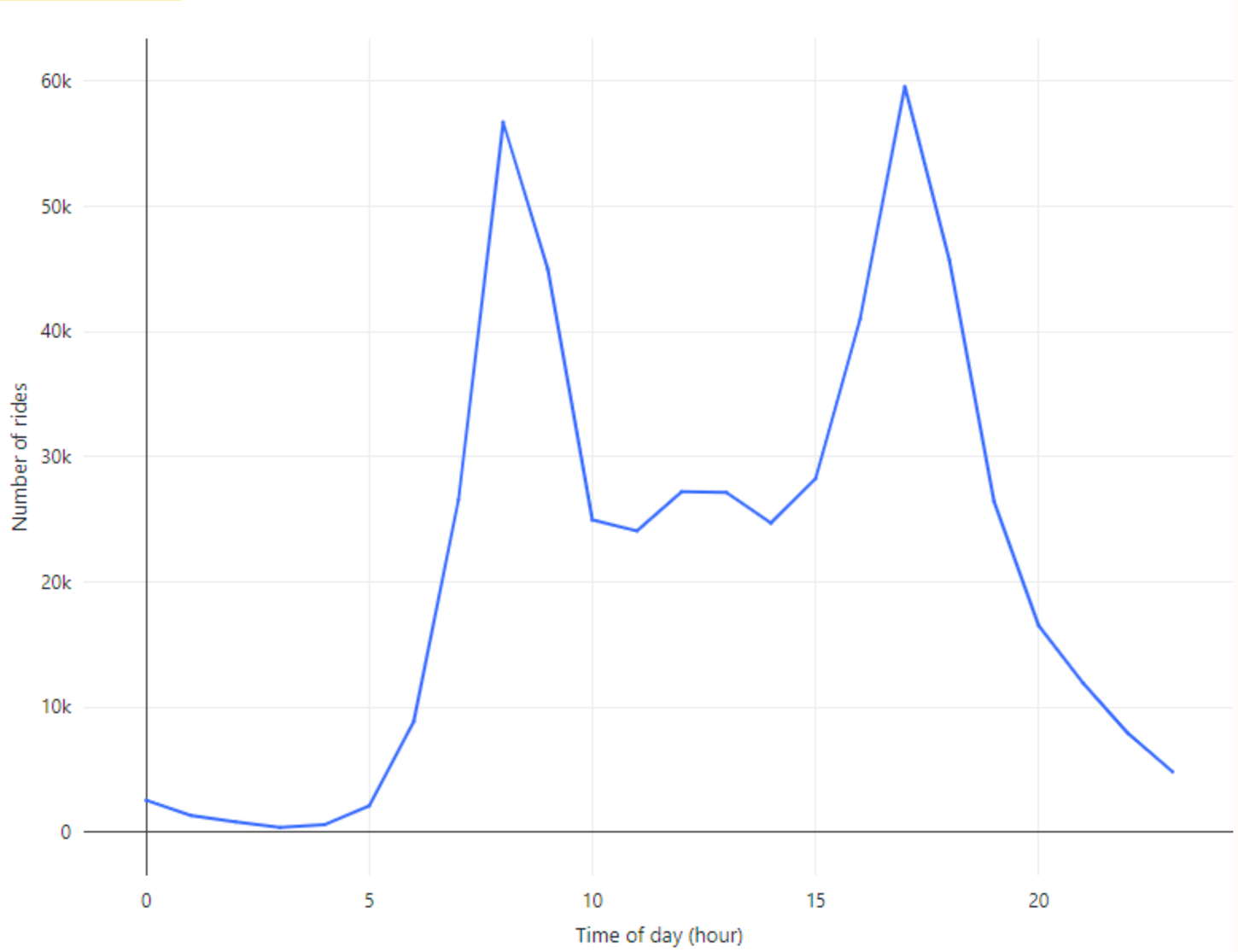
Gender Ratio for Stations

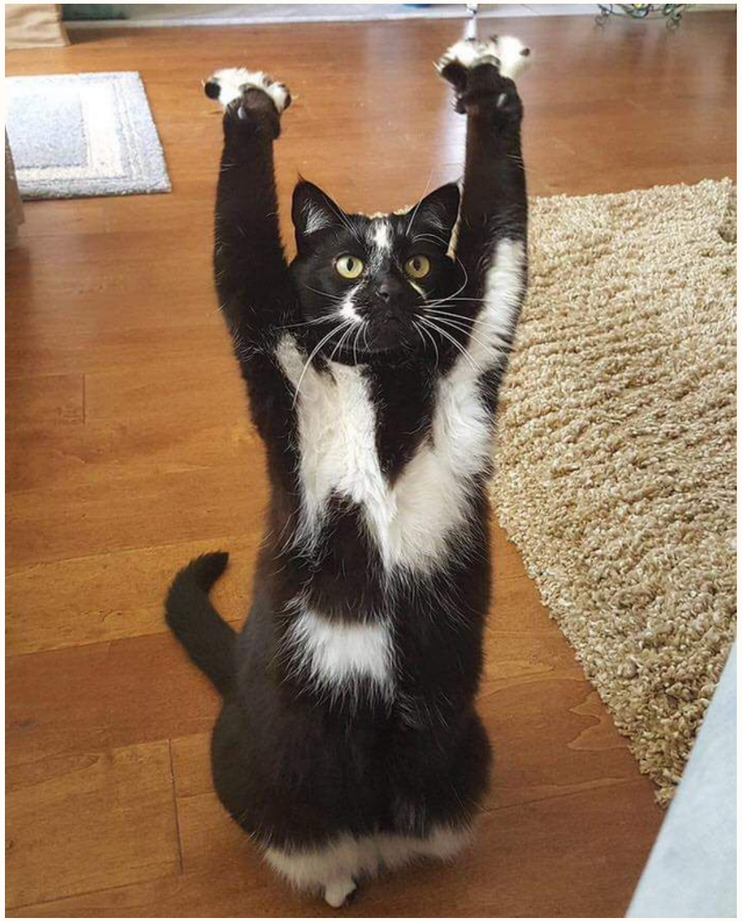


Hourly trip distribution

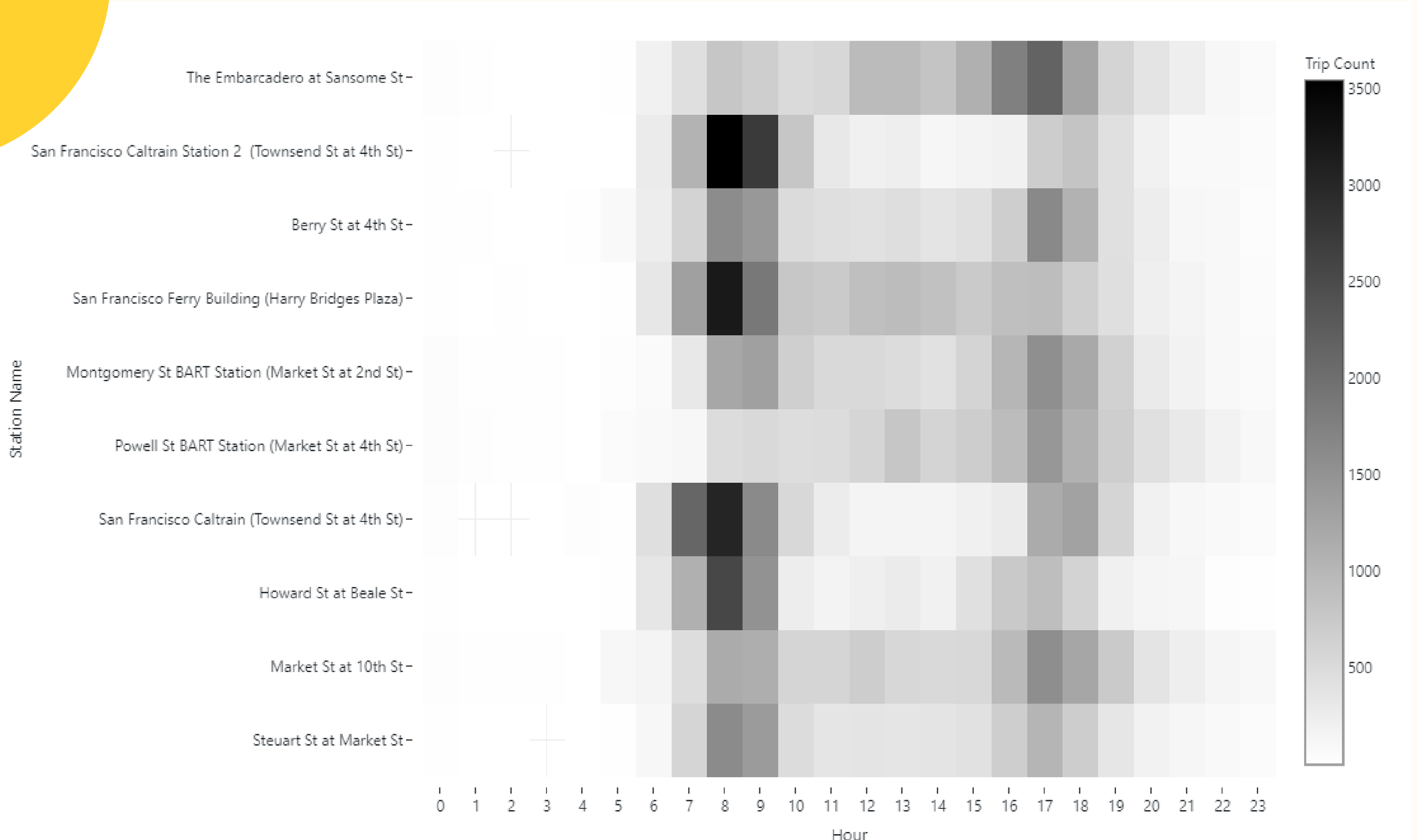


Hourly trip distribution





Top 10 station hourly distribution



Thank you for listening!