# Dataset

The dataset chosen was the set provided by Udacity titled '201902-fordgobike-tripdata.csv'. This dataset contained 183412 unique entries for Ford GoBike trip data that occurred during February 2019, with 16 features. During the cleaning process, I dropped all the null rows as well as corrected some data types, leaving the dataset with 174952 unique entries.

I also thought about what a typical company would want from the data. I concluded that a company like Ford GoBike would be interested in launching promotions so I decided to investigate 3 qualities: peak usage, target demographics and geographical usage.

The results I chose to illustrate in this analysis are:

- Bike Station Bikes Surplus and Deficits
- Bike Service Usage Per Hour and Weekday
- Target Demographic
- Geographical Usage

During the data exploration, I first opted to analyze the activity per station using the start_station_id. This was largely to get a feel for the data and see perhaps if there were extremes in the dataset. The visualization showed activity at all the stations and this led me to further exploring trends in start_station and end_station. I ended up subtracting the two metrics to determine the difference and visualizing the stations with the biggest differences. This created a new dataframe with still 174952 unique entries but with 4 new features: station_id, start_count, end_count and the difference. Given that there are over 300 stations, I opted to only use the top5 and bottom5 differences.

During the exploration for frequency, I looked to measure both hour and weekday. Using a method dt.hour and dt.weekday, I created 2 new features in the dataset to indicate the hour of day as well as the weekday, given the start_time.

For the target demographic, I opted to use member_birth_year in conjunction with member_gender as a bivariate analysis. I noticed that there were birth years from the 1800's which is obviously incorrect (user error probably). I dropped rows that had an age greater than 79 (member_birth_year < 1940) as the data was either user error or insignificant.

Lastly, for geographical location, I looked into the latitude and longitude as 2 features and also measured the frequency on a scatterplot for a multivariate analysis. Immediately after plotting on a map, I noticed 3 clusters of points corresponding to San Francisco, San Jose and Oakland. I divided the dataset into 3 separate datasets for each region and plotted them independently on their own maps.

# Summary of Findings

For bike surplus and deficit, the main finding was that given the service allows users to end their ride at a different station than they had originally rented from, it created an imbalance of bikes at certain stations, while creating a potential deficit of bikes at others. The biggest culprit, station 67, had a surplus of over 1200 rides in the month of February 2019 while the biggest deficit was station 243 of around 650. For peak usage, it was discovered that the majority of the users are commuters, using bikes as a mode of transportation to commute to work and home. This was interpreted from the peak usage at 8am and 5pm (coinciding with a typical 9-5 job) as well as a significant drop in usage on Saturdays and Sundays compared to the other 5 days of the week. For the target demographic, 25-34 age group was the highest group, while males were generally more likely to use the bike service. This coincides with the general age group in tech companies given the bay area is currently a tech hotspot. Lastly, the geographical usage overlaid on maps showed general hotspots for bike usage. Illustrating the hotspots allows for better allocation of resources should on-site service be required, but also allows for a better idea of expansion. One key discovery was a sizeable interest in between Oakland and Berkely, which could be a potential region for future expansions.

# Key Insights

I opted to use bar charts to illustrate the univariate data sets Bike Station Surplus and Deficit. To keep the visualization clean, I opted to plot only top 5 and bottom 5 differences. The key insight in this plot is to demonstrate that in only a month, certain stations have a significant surplus of bikes (hotspot for ride end location) while others have a huge deficit. This will create an imbalance of bikes at stations and could cause customer quality issues (if there are either no places to park the bike, or no bikes at all).

I also opted to use a bar chart for Bike Service Usage Per Hour and Weekday. As the hour and weekday is interval data, it is easy to spot the trend moving with time. The key insight from these plots demonstrates the main user groups are commuters, as the peaks of the hourly data coincide with peak commute times while the weekday data drops on the weekends (Monday - Friday 9-5 jobs).

For bivariate data, I opted for a heatmap, measuring gender and birth_year. This was to show a broader visualization of the cluster of ages that uses the bike service. As the individual age does not matter at such granular level (ie. someone age 25 and someone age 27 has no real significant difference in terms of categorization), I wanted to show the general trend. The key insight from this analysis was that 1) males are more frequent users of the service than females and 2) the 25-34 age group is the biggest user of our service. This can be used to target promotions and/or look for market expansions.

Lastly, for multi-variate data, I used a scatter plot overlaid on an image taken from Google maps, with color and size depicting the higher data counts. I wanted to map the plots of the highest activity of the service. Overlaying on a map provides a real-world depiction of the neighbourhoods with highest activity. The key insights show a cluster of activity near downtown SF and northeast of it close to the piers, clusters in Berkeley and Oakland (potentially coinciding with the colleges), and a small cluster in central downtown San Jose. This data can be used to allocate resources and staffing efficiently.

## Notes

I used Google Colab as an alternative to Jupyter Notebooks as I currently work for Google and would like to apply my new data analyst knowledge to my current role. I took the opportunity for this assignment to further familiarize myself with the Google tools at work.

For libraries, I used the standard numpy, pandas, matplotlib.pyplot, seaborn for analysis and visualizations. I further used colab_pdf (https://github.com/brpy/colab-pdf?fbclid=IwAR1Q9HO4PRYf68RrNp1W6ce478j0MPH2U-QFEFdBoqUZHWRY9jrEnReUN9c) to convert my ipynb to a pdf (feature not included in Google Colab)

For the maps used in the map visualizations, I measured using Google Maps (https://www.google.com/maps/d/u/0/edit?mid=1bfZUBsY30oUQQKUDDJhhxPvU3uJkl0hZ&ll=37.58612734137075%2C-122.15000000000002&z=10) and downloaded the images as a PNG file to be imported using matplotlib.image

Lastly, I referenced stackoverflow numerous times to aid throughout this entire analysis.