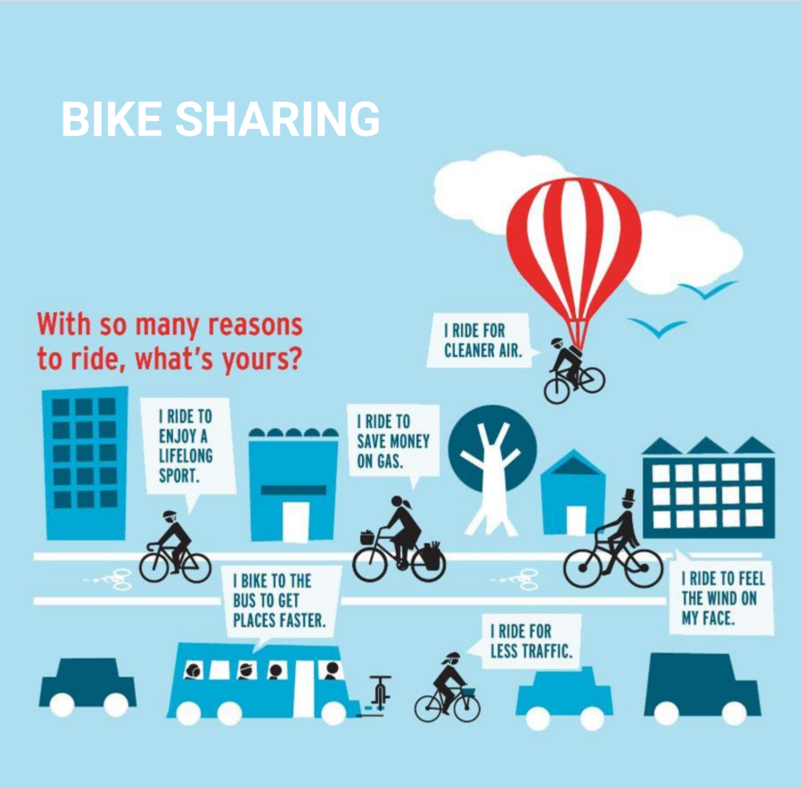
Swati Kohli
Nicholas Richmond
Kaitlyn Robinson

Analysis on Travel Behavior, and Factors predictive of its success

Introduction



1
- Background & Data
- Exploratory data analysis

2
- Purpose of the project
- Or Driving question

3
- Data Cleaning process

4
- Data Analysis

5
- Data Model
- Interpretation

6
- Conclusion
- Learning

BACKGROUND

01 BAY AREA

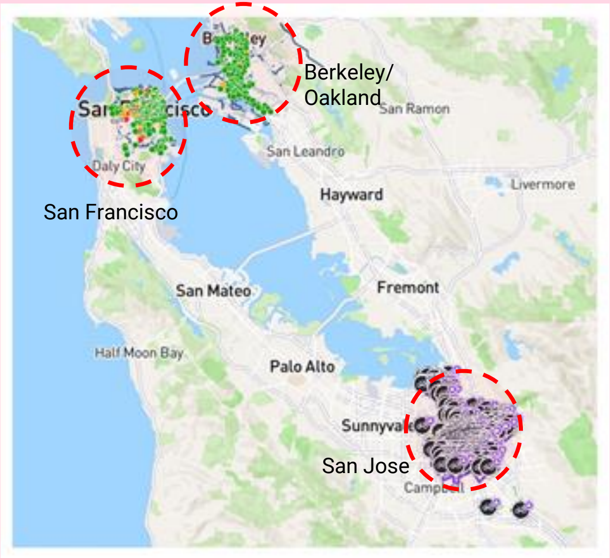
- A cultural and commercial center of California
- Ecologically important habitat
- Hustling with commuters and tourists

02 BIKE SHARING SYSTEM

- Creative initiative of bike sharing system
- Aims to reduce the congestion, noise and air pollution

03 LYFT BAY WHEELS

- Encourage people to cycle
- Extension of mass public transport, not replace it.
- Serves San Francisco, San Jose and Oakland/Berkeley



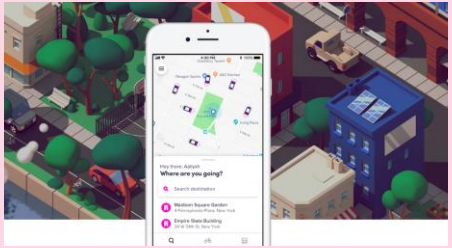
LYFT BIKE SHARING SYSTEM

1. JOIN

1. Day user (customer) or Subscriber (monthly/annually).

1. Pay through a credit card

1. Booth at few locations for cash



2. RIDE CHECK-OUT

1. Unlock- QR code scan on bike or member key

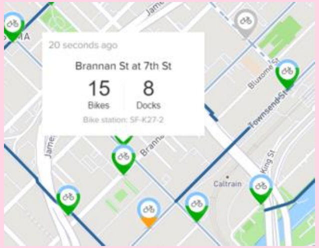
1. Ride within the time window paid

1. Extra charges exceeding time

3. RIDE CHECK-IN

1. Dock back at any docking station with empty racks.

1. System map on app shows station locations with status.



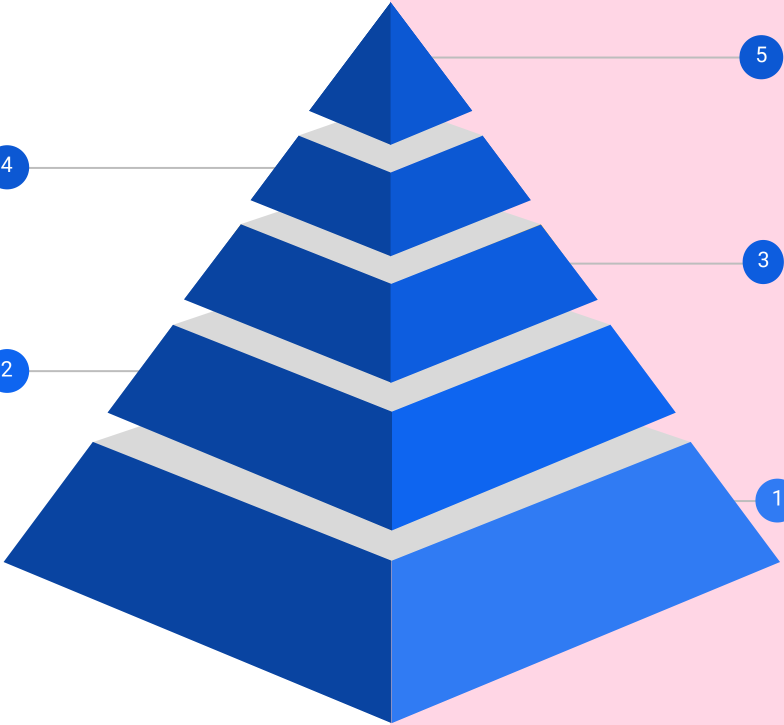
DATASET DESCRIPTION

- LYFT BAY WHEELS
  - Kaggle
    - Number of Rows = 1,053,067
    - Number of Columns = 10
    - User Information
      - User Type
      - Gender
      - Birth Year
    - Bike Information
      - Bike ID
    - Ride Information
      - Month (January to May) 2019
      - Trip Duration (Sec)
      - Start & End Station
  - Lyft
    - Location Information
      - Start & End Latitude & Longitude

DRIVING QUESTION

How has Lyft's bike sharing business fared in the first five months of 2019, in terms of travel behavior, and what factors could be predictive of its success in Bay Area?

Set a benchmark for future studies following our line of research



5 **Prediction Data Model**
Predict duration of trips based on various variables

4 **Revenue Generation**
Assessment of Contribution of Top Routes towards revenue generation

3 **Analysis of bike demand and supply**
For popular stations and popular routes

2 **Travel Behaviour**
Learn about user composition with respect to trips

1 **Usage Analysis**
Find the city most utilizes the service

OVERVIEW OF DATA CLEANING

1. month
2. trip_duration_sec
3. start_station_id
4. start_station_name
5. end_station_id
6. end_station_name
7. bike_id
8. user_type
9. member_birth_year
10. member_gender
11. start_station_latitude
12. start_station_longitude
13. end_station_latitude
14. end_station_longitude
15. start_city
16. end_city

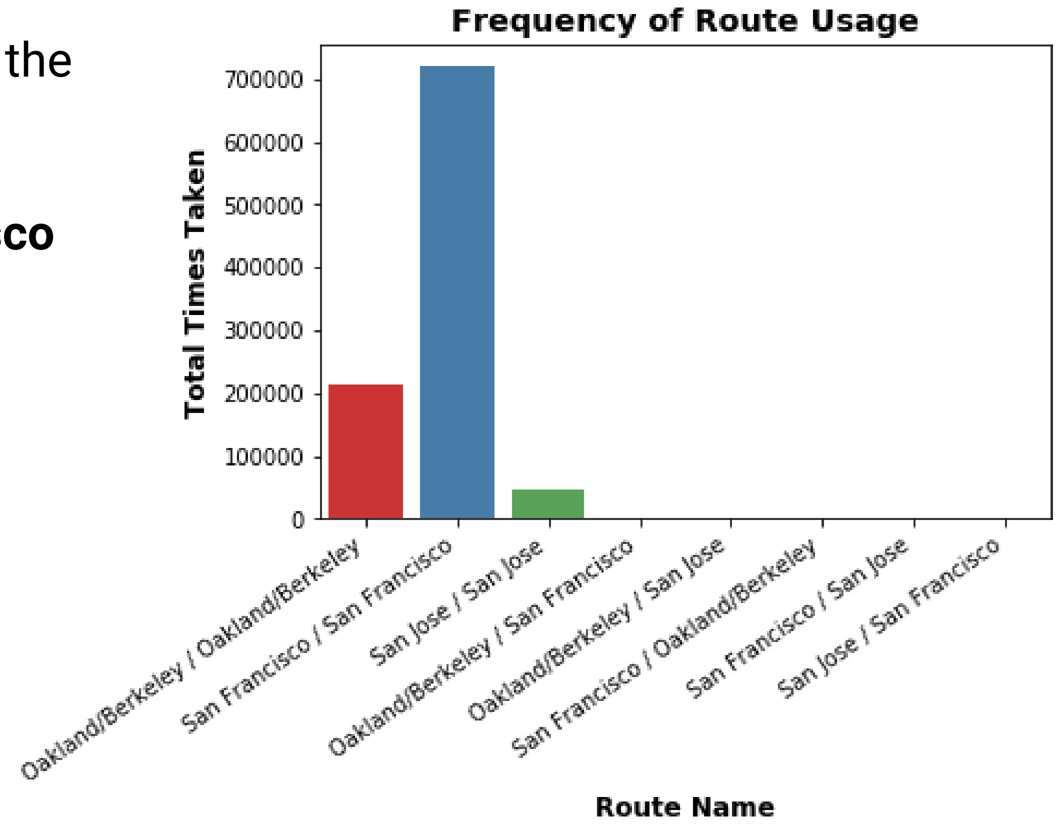
- Combined separate Kaggle CSVs
  - Located and removed NaNs and “Other” gender
- Combined separate Lyft website CSVs
  - Retained station columns and removed duplicates
- Merged Kaggle and Lyft data to include station coordinates
- Created functions to determine start and end city
- Removed outlier rides

Question #1 a)

Determine which route is used the most frequently.

San Francisco to San Francisco

Total Rides: 719,010



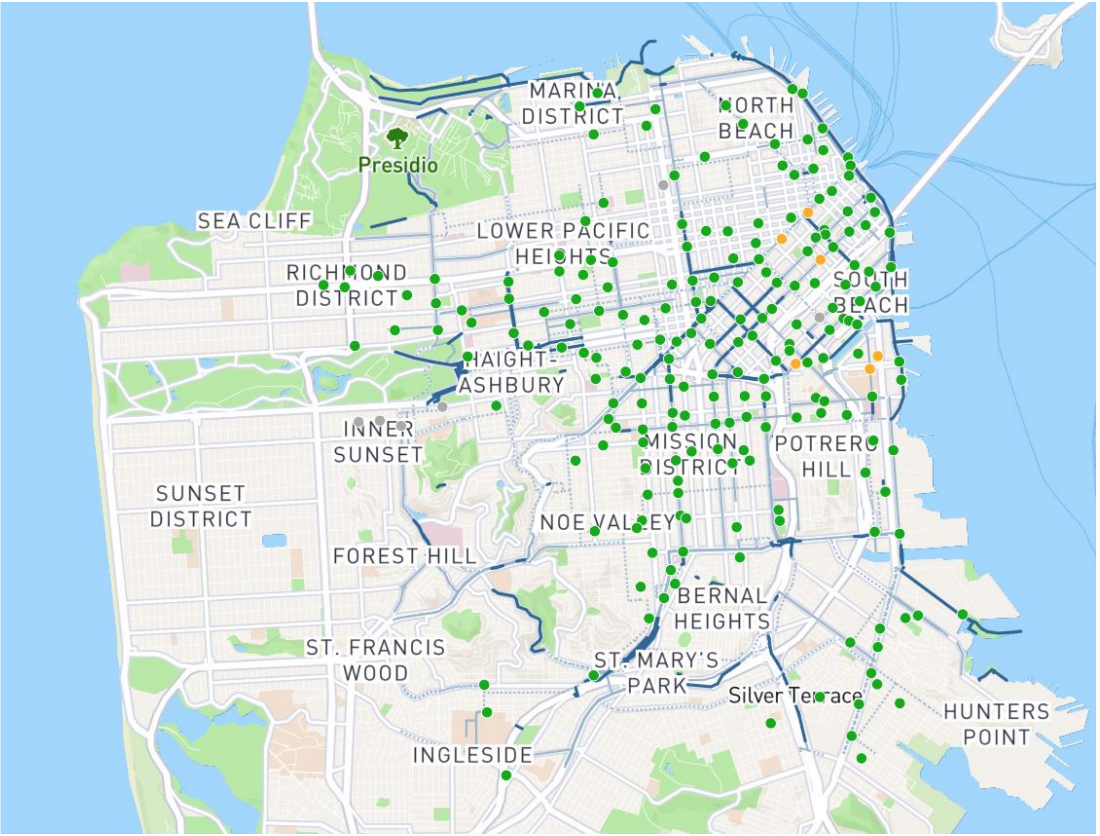
Question #1 a)

Bay Wheels in San Francisco

Unique Stations: 169
Number of Bikes: 3360

Stations spread out but clustered more in the middle which is the downtown area most likely where much of the business is done throughout the day.

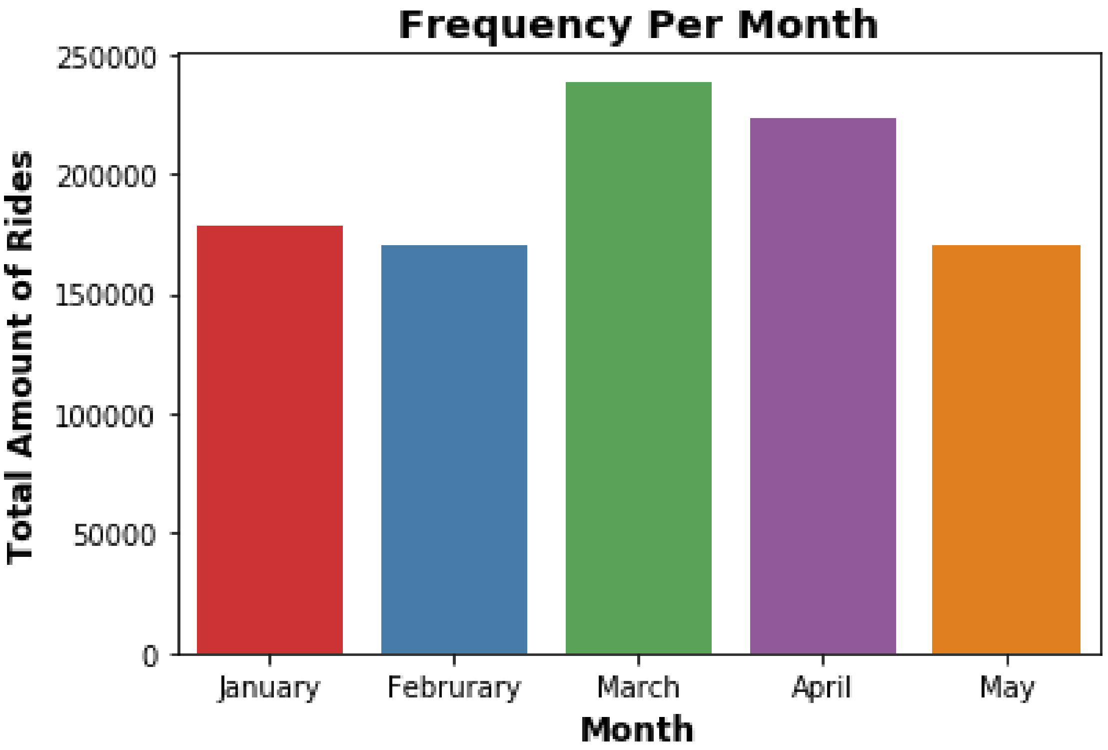
Also most likely to include many of the attractions for visitors and tourists.



Question #1 b)

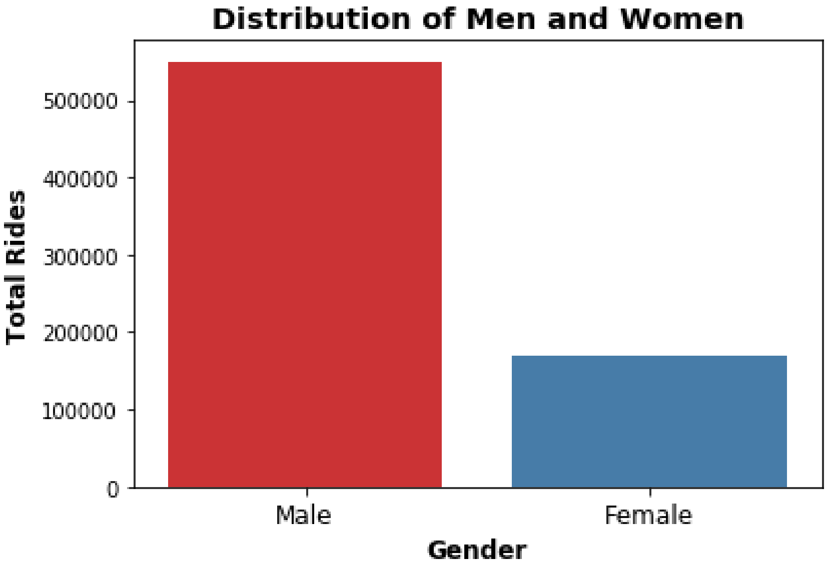
Determine which month is used the most frequently.

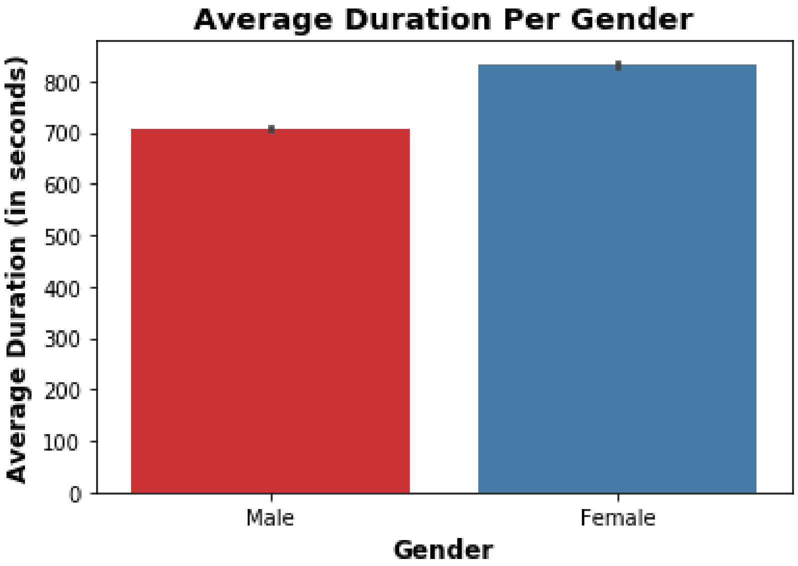
March



Question #2 a)

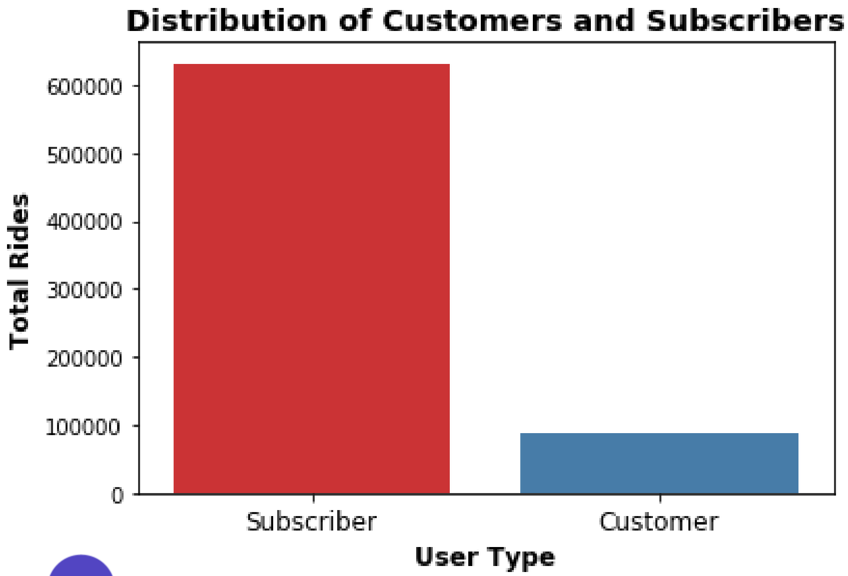
Determine the distribution of riders in San Francisco between men and women.

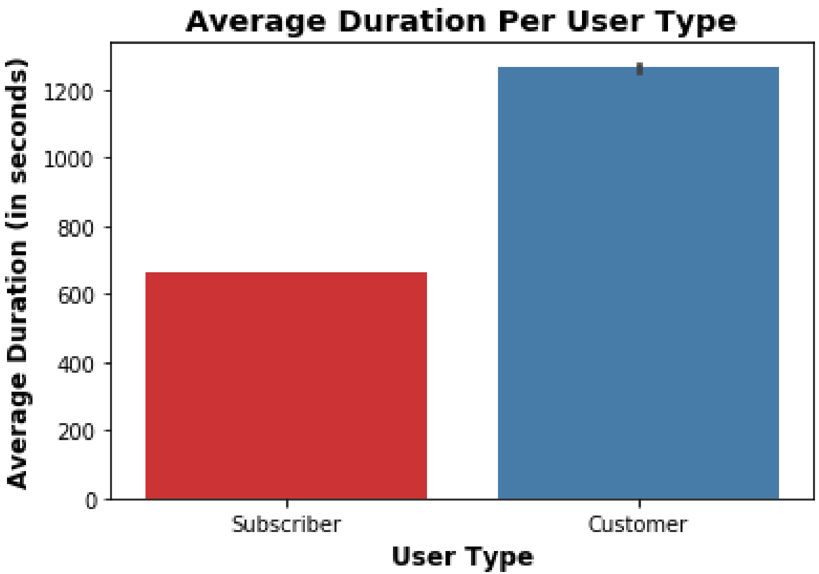




Question #2 b)

Determine the distribution of riders in San Francisco between customers and subscribers.



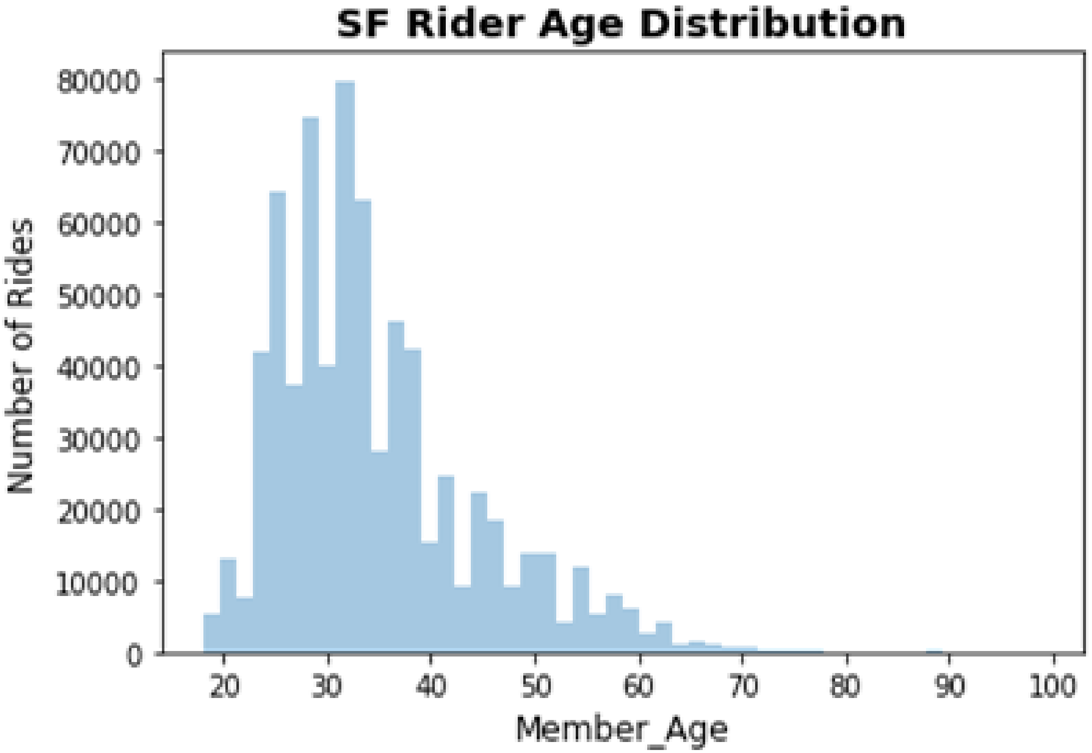


Question #2 c)

Determine the distribution of riders in San Francisco based on age.

The largest spikes represent the ages of 29 - 32.

- Overall most active users are from age group 26-35.
- Could be office goers or tourists



Question #3 - Analysis of Bike Supply and Demand

STEP 1

APPROACH AND WHY

System has a vast reach

Bike rotate within the system

Strategize bike availability as per analysis

STEP 2

CHALLENGE AND STRATEGY

Analysing popular stations based on usage

Analysing popular routes based on frequency of check ins & check outs

STEP 3

STEPS AND DATA VISUALIZATION

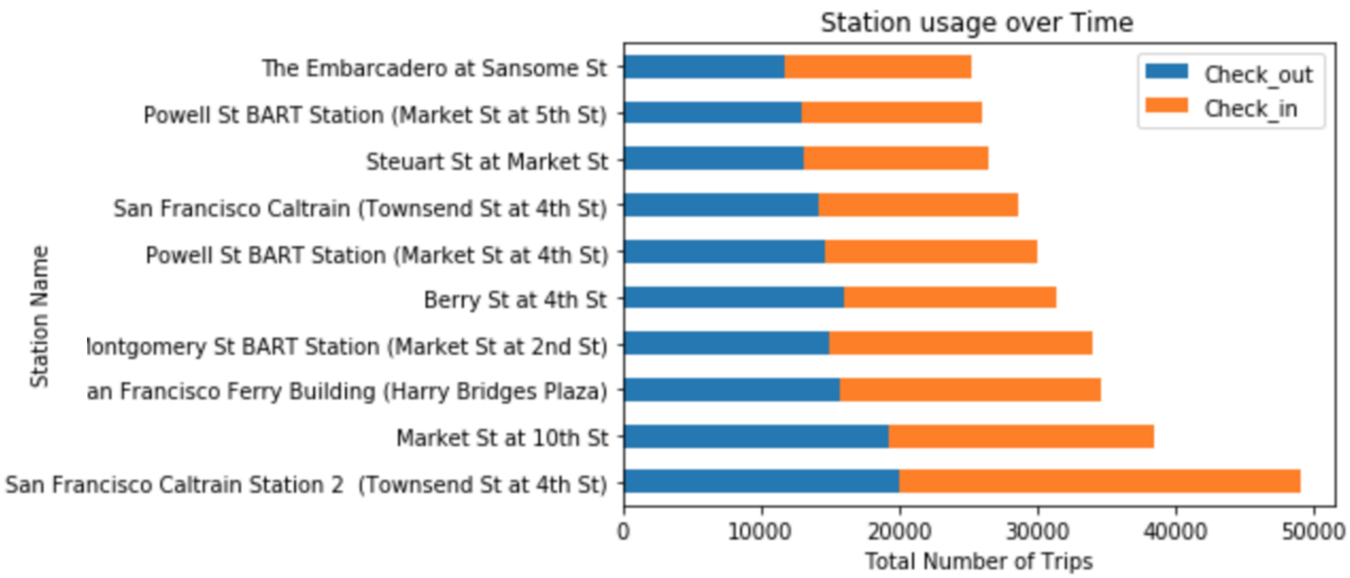
STEP 4

RESULT/ INTERPRETATION

Question #3 - Analysis of Bike Supply and Demand

STRATEGY #1:

Analysing popular stations based on usage



CHALLENGE

Due to data unavailability on daily/hourly basis, modified approach to analyse the overall demand and supply at popular stations.

STEPS

1. **Add a column** with total number of check ins and check outs for each station, arrange in descending order.
2. Get top 10 stations where

Total number of rides > 25000 (arbitrary).

RESULT/INTERPRETATION

Further investigation on the function and character of the area of these stops based on daily basis may show more clarity on the reasons of their popularity.

Question #3 - Analysis of Bike Supply and Demand

CODE SNIPPET

```
# Find the number of check-outs per station
station_start=sf_stations['start_station_name'].value_counts()
# Find the number of check-ins per station
station_end=sf_stations['end_station_name'].value_counts()
# Create a DataFrame with the check-outs and check-ins
# Create a column that sums check-outs and check-ins
# Create a column to get difference of check in - check out
station_counts = pd.concat([station_start, station_end], axis=1)
station_counts.rename(columns={'start_station_name':'Check_out', 'end_station_name':'Check_in'}, inplace=True)
station_counts['Total'] = station_counts['Check_out'] + station_counts['Check_in']
station_counts['CheckIn-CheckOut'] = station_counts['Check_in']- station_counts['Check_out']
# arrange in descending order to get top stations
station_counts = station_counts.sort_values('Total', ascending=False)
```

```
# top stations = total number of rides > 25000
top_stations= station_counts[station_counts['Total']>25000]
top_stations
```

	Check_out	Check_in	Total	CheckIn-CheckOut
San Francisco Caltrain Station 2 (Townsend St at 4th St)	19961	29032	48993	9071
Market St at 10th St	19226	19235	38461	9
San Francisco Ferry Building (Harry Bridges Plaza)	15727	18805	34532	3078
Montgomery St BART Station (Market St at 2nd St)	15006	18967	33973	3961
Berry St at 4th St	15985	15346	31331	-639

Question #3 - Analysis of Bike Supply and Demand

STRATEGY #2:

Analysing popular routes based on frequency of check-in and check-out

For example, a negative number of check-in minus check-out indicates more bikes leaving station than entering. This suggests more demand for bikes at that station

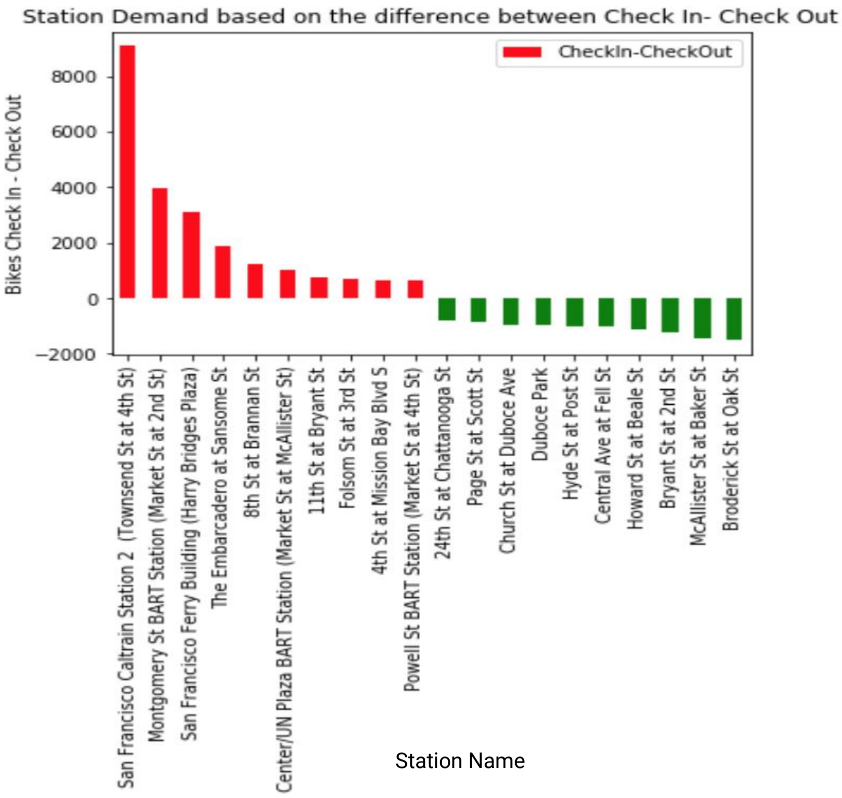
STEPS

1. **Add a column** with difference of check ins and check outs for each station, arrange in descending order.

2. Get top & bottom 10 stations.

RESULT/INTERPRETATION

This implies that the top 10 stations are the ones that had excess bikes at the station, while the bottom stations are the ones that should be monitored for reloading.



Question #3 - Analysis of Bike Supply and Demand

CODE SNIPPET

```
# BIKE DEMAND
# arrange in descending order to get difference
station_counts = station_counts.sort_values('CheckIn-CheckOut', ascending=False)
# 1. If more bikes check-in than check-out, implies bikes in excess
excess= station_counts.head(10)
# 2. If more bikes check-out than check-in, implies bikes in demand
demand= station_counts.tail(10)
# concatenating the two dataframes (excess and demand)
bike_demand = pd.concat([excess, demand])
bike_demand['Stations']= bike_demand.index
bike_demand.index = range(1,21)
bike_demand[['Stations','Check_out','Check_in','CheckIn-CheckOut']]
```

```
# table for data visualization
bike_demand_chart = bike_demand[['Stations','CheckIn-CheckOut']]
bike_demand_chart
```

	Stations	CheckIn-CheckOut
1	San Francisco Caltrain Station 2 (Townsend St...	9071
2	Montgomery St BART Station (Market St at 2nd St)	3961
3	San Francisco Ferry Building (Harry Bridges Pl...	3078
4	The Embarcadero at Sansome St	1858
5	8th St at Brannan St	1214
6	Civic Center/UN Plaza BART Station (Market St ...	992
7	11th St at Bryant St	736
8	Folsom St at 3rd St	676
9	4th St at Mission Bay Blvd S	640
10	Powell St BART Station (Market St at 4th St)	618
11	24th St at Chattanooga St	-822

Question #4 - Top Routes Contribution to Lyft's Revenue

STEP 1

APPROACH AND WHY

Might give pattern of usage

Implement specific marketing ideas

STEP 2

CHALLENGE AND STRATEGY

Calculating cost of a trip

Lack of unique identification of a subscriber

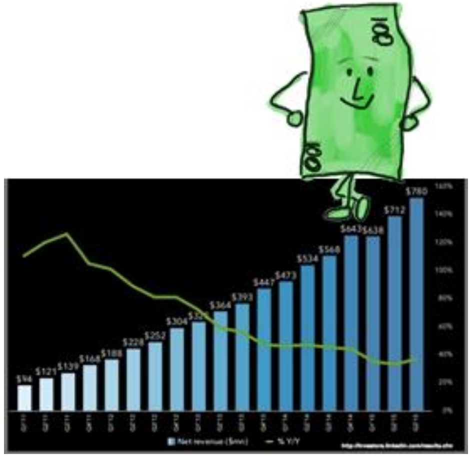
Can only analyse for a day user

STEP 3

DATA VISUALIZATION

STEP 4

RESULT/ INTERPRETATION



Question #4 - Top Routes Contribution to Lyft's Revenue

STRATEGY

Analysing popular stations based on usage

STEP 1- Cost of a trip

- **Subscriber** pays $15 or $149 monthly/annual member Receives unlimited 45-minute trips.
- **Customer** or a day user pays $2 and
- Receive a 30-minute trip.
- Over limit - $3 per additional 15 minutes for both.

STEP 2 - Calculation

- From the dataset, only day user data is extracted
- Made **cost function** and applied to calculate the cost of each day user ride based on duration.
- Made a dataframe with **top routes** based on frequency of that route(more than 250- arbitrary).
- **Calculated revenue contribution** of those top routes (9 routes have more than 250 frequency) towards total day user revenue generation of 5 months.

Question #4 - Top Routes Contribution to Lyft's Revenue

```
# extract only day user data
du_routes= sf_stations[sf_stations['user_type']=='Customer']
du_routes
# percentage contribution of day user to total ridership over 5 months
print("Total Number of users (Subscriber + day user): ",sf_stations.shape[0])
print("Number of day users: ",du_routes.shape[0])
print ("Percentage of day user riders out of total riders for 5 months is: ",
       (du_routes.shape[0]/sf_stations.shape[0])*100, "%" )

# function to calculate cost paid by day user
# day user pays $2 for first 30mins and $3 for every subsequent 15 minutes as per LYFT website
import math
def Calculate_Cost(x):

    #first 30 mins free every subsequent 15mins = $3
    if x>1800:
        total_time = x-1800
        if (total_time%900) == 0:
            total_cost = ((total_time/900)*3)+2
        else:
            total_cost = ((math.ceil(total_time/900))*3) +
    else:
        total_cost = 2
    return (total_cost)
# apply cost function to dataframe
du_routes['Cost_of_trip']= du_routes['trip_duration_sec'].
```

```
# top routes in SF based on frequency of rides (more than 200) for day user
dutrips_df = du_routes.groupby(['start_station_id','end_station_id']).size().reset_index(name = 'number of trips')
top_dutrips = dutrips_df.sort_values('number of trips', ascending=False)
top_9_dutrips = top_dutrips[top_dutrips['number of trips']>200]
top_9_dutrips
# extract data based on the 9 top routes in the day user model. Below one example:
du_route1 = du_routes[(du_routes['start_station_id']== 15) & (sf_stations['end_station_id']==6)]
# concatenate the 9 dataframes
du_bike_revenue = pd.concat([du_route1, du_route2,du_route3,du_route4,du_route5,du_route6,du_route7,du_route8,du_route9])

# calculate the revenue from total day users in 5 months
du_routes['Cost_of_trip'].sum()
# calculate the revenue from day users of top 9 routes in 5 months
top9_routes= du_bike_revenue['Cost_of_trip'].sum()
# percentage contribution wrt revenue generation of day user to total day user ridership over 5 months
print("Total revenue generation by day user for 5 months): ",total_du_revenue)
print("Total revenue generation by day user for top 9 routes in 5 months):",top9_routes)
print ("Percentage of revenue contribution of day user riders to total for 5 months is: ",(top9_routes/total_du_revenue)*100, "%
# Percentage of revenue contribution of day user riders to total for 5 months is:  6.62%
```

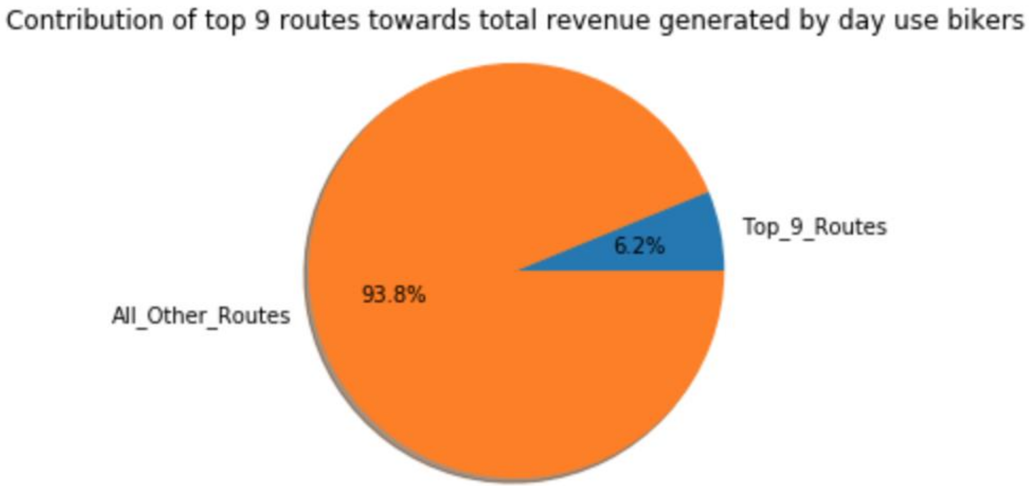
Question #4 - Top Routes Contribution to Lyft's Revenue

STRATEGY

Analysing popular stations based on usage

RESULT

- The total revenue generation by day users over the 5 months is **$283,160**.
- It is observed that the top 9 routes contributed **6.2%** of the total revenue generation by the day users over 5 months. Further study on duration could reveal more insight into the revenue contributions.



Question #4 - Top Routes Contribution to Lyft's Revenue

STRATEGY

Analysing popular stations based on usage

INTERPRETATION

- Out of the top 9 routes, 5 are **cyclic trips**, i.e. start and end station is same.
- When not cyclic, the trips permutate over 6 unique stations.
- Therefore could be leisurely tourist trips.
- As expected, when checked these are mostly stations around tourist attractions.
- As a suggestion, to improve revenue, this data can be used for target marketing like discount coupons at these stations or provide guided tours by collaboration with agencies.

	start_station_id	end_station_id	number of trips
1113	15.0	6.0	617
15695	377.0	377.0	503
16478	400.0	400.0	360
430	6.0	15.0	353
1237	15.0	371.0	291
16400	399.0	399.0	248
1120	15.0	15.0	237
423	6.0	6.0	228
1249	15.0	400.0	207

Question #5

PREDICT DURATION OF TRIP BASED ON VARIABLES

```
Intercept                      1341.1971
age_range[T.36-50]              -21.6903
age_range[T.Above 50]            35.3900
member_gender[T.Male]          -103.6008
user_type[T.Subscriber]        -594.1069
```

Age Range:
1. 18-35
2. 36-50
3. Above 50

Member Gender:
1. Female
2. Male

User Type:
1. Customer
2. Subscriber

trip_duration_sec ~ age_range + member_gender + user_type

- Trip duration **decreases by 22 seconds** if the **age range** changes from **18-35 to 36-50**.
- Trip duration **increases by 35 seconds** if the **age range** changes from **18-35 to Above 50**.
- Trip duration **decreases by 104 seconds** if **gender** changes from **female to male**.
- Trip duration **decreases by 594 seconds** if **user type** changes from **customers to subscribers**.

CONCLUSION & LEARNING

Lyft's Baywheels has performed well in the first five months of 2019!

DATA AVAILABILITY

Use of relevant open source data for research- good strategy to gain a deeper understanding of the functioning LYFT.

OTHER FACTORS OF MARKET FAILURE

Bike vandalism, inappropriate bike maintenance, insufficient availability in certain locations or at certain times. Some of these can be tracked by data analysis.

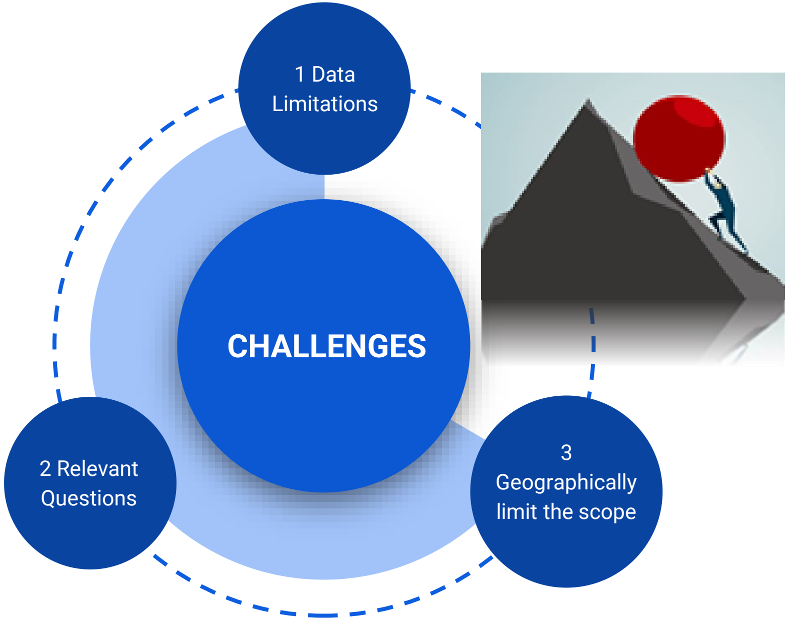
FURTHER RESEARCH

To discover additional factors like land use patterns, existing public transport network, bike related facilities, bike infrastructure etc that may affect the success

MARKET UPGRADABILITY/ COMPETITION

Research avenue to examine how BSS compares with emerging dock less bike systems.

CHALLENGES



1. Data Limitations

❖ Restricted by user unidentification
❖ Unique identifier= accurate data on frequency and user
❖ **<u>Overcame</u>** by deliberating other ways of classifications to allow comparative analysis.

1. Relevant Questions

❖ Structuring relevant questions to answer through data modeling, and that flowed well together.
❖ Overcame by deliberation on questions that fit the purpose along with data visualisation.
❖ Further, our analysis required more columns for example, “Start City” and “End City”, to compare data between cities and restrict our scope of study.

1. Geographical incorporation for comparative analysis and reduction

❖ With further exploration found data with latitude and longitude on lyft website.
❖ **<u>Overcame</u>** the hurdle on how to merge and extract relevant information with research and study on incorporating the latitudes and longitudes to identify cities and enable comparative analysis.

Challenges but beneficial in the long run!

LEARNING

CHALLENGES

1 Data Limitations

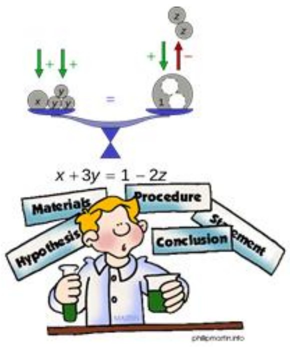
2 Relevant Questions

3 Geographically limit scope

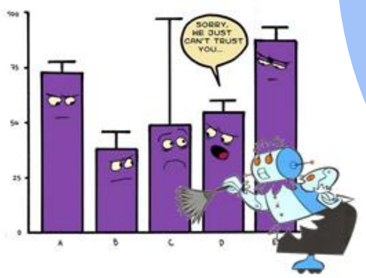
LEARNINGS

Never Assume
Rely on data analysis

Eg. Regression equation variable analysis based on correlation matrix



Effect of data cleaning and outlier identification



Eg. Trips with more than rideable hours might give misleading results.

Benefits of data visualisation

Eg. Demand and supply much better visualized graphically, heatmap of correlation matrix gives easily interpretable insight, animation might enhance the visual analysis

Questions?