Part_I_exploration_template

February 9, 2023

1 Ford Gobike Data Exploration

1.1 by Shahad Al-Khalifa

1.2 Introduction

Ford GoBike is a regional public bicycle sharing system in the San Francisco Bay Area, California. The Ford GoBike system, which began operations as Bay Area Bike Share in August 2013, now includes approximately 2,600 bicycles in 262 stations around San Francisco, the East Bay, and San Jose. In a collaboration with Ford Motor Company, the system was formally introduced as Ford GoBike on June 28, 2017.

Ford GoBike, like other bike sharing systems, is made up of a fleet of carefully constructed, strong, and durable bikes that are docked at a network of docking stations located around the city. The bikes may be unlocked at any station in the system and returned to any other station, making them perfect for one-way excursions. The bikes are accessible for use 24 hours a day, seven days a week, 365 days a year, and riders who become members or purchase passes have access to all bikes in the network.

1.3 Preliminary Wrangling

```
In [1]: # import all packages and set plots to be embedded inline
        import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sb

        %matplotlib inline
```

```
In [2]: # Read file and view the first five rows
        df = pd.read_csv('201902-fordgobike-tripdata.csv')
        df.head()
```

```
Out[2]:    duration_sec                start_time                  end_time  \
        0         52185  2019-02-28 17:32:10.1450  2019-03-01 08:01:55.9750
        1         42521  2019-02-28 18:53:21.7890  2019-03-01 06:42:03.0560
        2         61854  2019-02-28 12:13:13.2180  2019-03-01 05:24:08.1460
        3         36490  2019-02-28 17:54:26.0100  2019-03-01 04:02:36.8420
        4          1585  2019-02-28 23:54:18.5490  2019-03-01 00:20:44.0740
```

```
   start_station_id                                start_station_name  \
0              21.0  Montgomery St BART Station (Market St at 2nd St)
1              23.0                    The Embarcadero at Steuart St
2              86.0                          Market St at Dolores St
3             375.0                          Grove St at Masonic Ave
4               7.0                              Frank H Ogawa Plaza

   start_station_latitude  start_station_longitude  end_station_id  \
0               37.789625              -122.400811            13.0
1               37.791464              -122.391034            81.0
2               37.769305              -122.426826             3.0
3               37.774836              -122.446546            70.0
4               37.804562              -122.271738           222.0

                              end_station_name  end_station_latitude  \
0                Commercial St at Montgomery St             37.794231
1                           Berry St at 4th St             37.775880
2  Powell St BART Station (Market St at 4th St)             37.786375
3                        Central Ave at Fell St             37.773311
4                         10th Ave at E 15th St             37.792714

   end_station_longitude  bike_id   user_type  member_birth_year  \
0            -122.402923     4902    Customer             1984.0
1            -122.393170     2535    Customer                NaN
2            -122.404904     5905    Customer             1972.0
3            -122.444293     6638  Subscriber             1989.0
4            -122.248780     4898  Subscriber             1974.0

  member_gender bike_share_for_all_trip
0          Male                     No
1           NaN                     No
2          Male                     No
3         Other                     No
4          Male                    Yes
```

In [3]:
```
# Look at a sample of data
df.sample(10)
```

Out[3]:
```
        duration_sec                start_time                  end_time  \
148583           740  2019-02-06 21:17:50.0850  2019-02-06 21:30:10.1370
29942            231  2019-02-25 01:42:22.5220  2019-02-25 01:46:13.6160
125358           899  2019-02-11 07:27:02.1700  2019-02-11 07:42:02.0240
126131           186  2019-02-10 20:54:39.7840  2019-02-10 20:57:46.1790
62832            405  2019-02-20 16:23:37.5250  2019-02-20 16:30:23.1430
99182            216  2019-02-14 20:57:58.7090  2019-02-14 21:01:35.5350
69609            627  2019-02-19 19:10:48.0320  2019-02-19 19:21:15.0780
60891            354  2019-02-20 18:11:29.2160  2019-02-20 18:17:23.4220
```

```
103136           326  2019-02-14 14:33:00.2110  2019-02-14 14:38:26.2210
16768            394  2019-02-27 08:09:13.5290  2019-02-27 08:15:47.9130

        start_station_id          start_station_name  start_station_latitude  \
148583             245.0      Downtown Berkeley BART               37.870139
29942              310.0   San Fernando St at 4th St               37.335885
125358              53.0      Grove St at Divisadero               37.775946
126131              75.0    Market St at Franklin St               37.773793
62832               24.0       Spear St at Folsom St               37.789677
99182              285.0  Webster St at O'Farrell St               37.783521
69609               50.0       2nd St at Townsend St               37.780526
60891               47.0       4th St at Harrison St               37.780955
103136              50.0       2nd St at Townsend St               37.780526
16768               73.0      Pierce St at Haight St               37.771793

        start_station_longitude  end_station_id  \
148583              -122.268422           239.0
29942               -121.885660           357.0
125358              -122.437777            20.0
126131              -122.421239            77.0
62832               -122.390428            25.0
99182               -122.431158            54.0
69609               -122.390288            61.0
60891               -122.399749            81.0
103136              -122.390288            64.0
16768               -122.433708            29.0

                                      end_station_name  end_station_latitude  \
148583                    Bancroft Way at Telegraph Ave             37.868813
29942                              2nd St at Julian St             37.341132
125358  Mechanics Monument Plaza (Market St at Bush St)             37.791300
126131                             11th St at Natoma St             37.773507
62832                               Howard St at 2nd St             37.787522
99182             Alamo Square (Steiner St at Fulton St)             37.777547
69609                               Howard St at 8th St             37.776513
60891                                Berry St at 4th St             37.775880
103136                             5th St at Brannan St             37.776754
16768                    O'Farrell St at Divisadero St             37.782405

        end_station_longitude  bike_id   user_type  member_birth_year  \
148583            -122.258764      342  Subscriber             1992.0
29942             -121.892844     6167  Subscriber             1992.0
125358            -122.399051     5070  Subscriber             1987.0
126131            -122.416040     4752  Subscriber                NaN
62832             -122.397405     6095  Subscriber             1993.0
99182             -122.433274     5916  Subscriber             1991.0
69609             -122.411306     4335  Subscriber             1992.0
60891             -122.393170     5937  Subscriber             1992.0
```

```
103136            -122.399018     6545  Subscriber             1950.0
16768             -122.439446     4834    Customer             1973.0

       member_gender bike_share_for_all_trip
148583        Female                      No
29942           Male                     Yes
125358        Female                      No
126131           NaN                      No
62832         Female                      No
99182           Male                      No
69609           Male                      No
60891         Female                      No
103136          Male                      No
16768           Male                      No
```

In [4]:
```
# Look at the information of the data such as columns and their data types
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 183412 entries, 0 to 183411
Data columns (total 16 columns):
duration_sec               183412 non-null int64
start_time                 183412 non-null object
end_time                   183412 non-null object
start_station_id           183215 non-null float64
start_station_name         183215 non-null object
start_station_latitude     183412 non-null float64
start_station_longitude    183412 non-null float64
end_station_id             183215 non-null float64
end_station_name           183215 non-null object
end_station_latitude       183412 non-null float64
end_station_longitude      183412 non-null float64
bike_id                    183412 non-null int64
user_type                  183412 non-null object
member_birth_year          175147 non-null float64
member_gender              175147 non-null object
bike_share_for_all_trip    183412 non-null object
dtypes: float64(7), int64(2), object(7)
memory usage: 22.4+ MB
```

In [5]:
```
# Change the data type for start and end time to be datetime64[ns]
df['start_time'] = pd.to_datetime(df.start_time)
df['end_time'] = pd.to_datetime(df.end_time)
```

In [6]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 183412 entries, 0 to 183411
```

```
Data columns (total 16 columns):
duration_sec               183412 non-null int64
start_time                 183412 non-null datetime64[ns]
end_time                   183412 non-null datetime64[ns]
start_station_id           183215 non-null float64
start_station_name         183215 non-null object
start_station_latitude     183412 non-null float64
start_station_longitude    183412 non-null float64
end_station_id             183215 non-null float64
end_station_name           183215 non-null object
end_station_latitude       183412 non-null float64
end_station_longitude      183412 non-null float64
bike_id                    183412 non-null int64
user_type                  183412 non-null object
member_birth_year          175147 non-null float64
member_gender              175147 non-null object
bike_share_for_all_trip    183412 non-null object
dtypes: datetime64[ns](2), float64(7), int64(2), object(5)
memory usage: 22.4+ MB
```

In [7]:
```
# Create a new column containing the date, time, and day of the week of both start and e

# Date
df['start_date'] = pd.to_datetime(df['start_time']).dt.date
df['end_date'] = pd.to_datetime(df['end_time']).dt.date

# Day of the week
df['start_day'] = pd.to_datetime(df['start_time']).dt.day_name()
df['end_day'] = pd.to_datetime(df['end_time']).dt.day_name()

# Time
df['start_time'] = pd.to_datetime(df['start_time']).dt.time
df['end_time'] = pd.to_datetime(df['end_time']).dt.time

df.head(10)
```

Out[7]:

	duration_sec	start_time	end_time	start_station_id	\
0	52185	17:32:10.145000	08:01:55.975000	21.0	
1	42521	18:53:21.789000	06:42:03.056000	23.0	
2	61854	12:13:13.218000	05:24:08.146000	86.0	
3	36490	17:54:26.010000	04:02:36.842000	375.0	
4	1585	23:54:18.549000	00:20:44.074000	7.0	
5	1793	23:49:58.632000	00:19:51.760000	93.0	
6	1147	23:55:35.104000	00:14:42.588000	300.0	
7	1615	23:41:06.766000	00:08:02.756000	10.0	
8	1570	23:41:48.790000	00:07:59.715000	10.0	
9	1049	23:49:47.699000	00:07:17.025000	19.0	

```
                                start_station_name  start_station_latitude  \
0  Montgomery St BART Station (Market St at 2nd St)               37.789625
1                      The Embarcadero at Steuart St               37.791464
2                           Market St at Dolores St               37.769305
3                           Grove St at Masonic Ave               37.774836
4                               Frank H Ogawa Plaza               37.804562
5                       4th St at Mission Bay Blvd S               37.770407
6                              Palm St at Willow St               37.317298
7                         Washington St at Kearny St               37.795393
8                         Washington St at Kearny St               37.795393
9                               Post St at Kearny St               37.788975

   start_station_longitude  end_station_id  \
0              -122.400811            13.0
1              -122.391034            81.0
2              -122.426826             3.0
3              -122.446546            70.0
4              -122.271738           222.0
5              -122.391198           323.0
6              -121.884995           312.0
7              -122.404770           127.0
8              -122.404770           127.0
9              -122.403452           121.0

                              end_station_name  end_station_latitude  \
0                 Commercial St at Montgomery St             37.794231
1                             Berry St at 4th St             37.775880
2  Powell St BART Station (Market St at 4th St)             37.786375
3                          Central Ave at Fell St             37.773311
4                           10th Ave at E 15th St             37.792714
5                              Broadway at Kearny             37.798014
6                        San Jose Diridon Station             37.329732
7                          Valencia St at 21st St             37.756708
8                          Valencia St at 21st St             37.756708
9                              Mission Playground             37.759210

   end_station_longitude  bike_id   user_type  member_birth_year  \
0            -122.402923     4902    Customer             1984.0
1            -122.393170     2535    Customer                NaN
2            -122.404904     5905    Customer             1972.0
3            -122.444293     6638  Subscriber             1989.0
4            -122.248780     4898  Subscriber             1974.0
5            -122.405950     5200  Subscriber             1959.0
6            -121.901782     3803  Subscriber             1983.0
7            -122.421025     6329  Subscriber             1989.0
8            -122.421025     6548  Subscriber             1988.0
9            -122.421339     6488  Subscriber             1992.0
```

```
  member_gender bike_share_for_all_trip  start_date    end_date start_day  \
0          Male                      No  2019-02-28  2019-03-01  Thursday
1           NaN                      No  2019-02-28  2019-03-01  Thursday
2          Male                      No  2019-02-28  2019-03-01  Thursday
3         Other                      No  2019-02-28  2019-03-01  Thursday
4          Male                     Yes  2019-02-28  2019-03-01  Thursday
5          Male                      No  2019-02-28  2019-03-01  Thursday
6        Female                      No  2019-02-28  2019-03-01  Thursday
7          Male                      No  2019-02-28  2019-03-01  Thursday
8         Other                      No  2019-02-28  2019-03-01  Thursday
9          Male                      No  2019-02-28  2019-03-01  Thursday

  end_day
0  Friday
1  Friday
2  Friday
3  Friday
4  Friday
5  Friday
6  Friday
7  Friday
8  Friday
9  Friday
```

In [8]:
```
# The sum of null values
df.isnull().sum()
```

Out[8]:
```
duration_sec                0
start_time                  0
end_time                    0
start_station_id          197
start_station_name        197
start_station_latitude      0
start_station_longitude     0
end_station_id            197
end_station_name          197
end_station_latitude        0
end_station_longitude       0
bike_id                     0
user_type                   0
member_birth_year        8265
member_gender            8265
bike_share_for_all_trip     0
start_date                  0
end_date                    0
start_day                   0
end_day                     0
dtype: int64
```

```
In [9]: # Remove rows that does not have gender value
        df = df[df['member_gender'].isnull() == False]
        df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 175147 entries, 0 to 183411
Data columns (total 20 columns):
duration_sec               175147 non-null int64
start_time                 175147 non-null object
end_time                   175147 non-null object
start_station_id           174952 non-null float64
start_station_name         174952 non-null object
start_station_latitude     175147 non-null float64
start_station_longitude    175147 non-null float64
end_station_id             174952 non-null float64
end_station_name           174952 non-null object
end_station_latitude       175147 non-null float64
end_station_longitude      175147 non-null float64
bike_id                    175147 non-null int64
user_type                  175147 non-null object
member_birth_year          175147 non-null float64
member_gender              175147 non-null object
bike_share_for_all_trip    175147 non-null object
start_date                 175147 non-null object
end_date                   175147 non-null object
start_day                  175147 non-null object
end_day                    175147 non-null object
dtypes: float64(7), int64(2), object(11)
memory usage: 28.1+ MB
```

```
In [10]: df.isnull().sum()
```

```
Out[10]: duration_sec                0
         start_time                  0
         end_time                    0
         start_station_id          195
         start_station_name        195
         start_station_latitude      0
         start_station_longitude     0
         end_station_id            195
         end_station_name          195
         end_station_latitude        0
         end_station_longitude       0
         bike_id                     0
         user_type                   0
         member_birth_year           0
         member_gender               0
```

```
bike_share_for_all_trip      0
start_date                   0
end_date                     0
start_day                    0
end_day                      0
dtype: int64
```

```
In [11]: # Remove rows that does not have start_station_id
         df = df[df['start_station_id'].isnull() == False]
         df.isnull().sum()
```

```
Out[11]: duration_sec              0
         start_time                0
         end_time                  0
         start_station_id          0
         start_station_name        0
         start_station_latitude    0
         start_station_longitude   0
         end_station_id            0
         end_station_name          0
         end_station_latitude      0
         end_station_longitude     0
         bike_id                   0
         user_type                 0
         member_birth_year         0
         member_gender             0
         bike_share_for_all_trip   0
         start_date                0
         end_date                  0
         start_day                 0
         end_day                   0
         dtype: int64
```

```
In [12]: # Change the data type of start and end station id to integer
         df = df.astype({"start_station_id":"int","end_station_id":"int"})
         df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 174952 entries, 0 to 183411
Data columns (total 20 columns):
duration_sec               174952 non-null int64
start_time                 174952 non-null object
end_time                   174952 non-null object
start_station_id           174952 non-null int64
start_station_name         174952 non-null object
start_station_latitude     174952 non-null float64
start_station_longitude    174952 non-null float64
end_station_id             174952 non-null int64
end_station_name           174952 non-null object
```

```
end_station_latitude       174952 non-null float64
end_station_longitude      174952 non-null float64
bike_id                    174952 non-null int64
user_type                  174952 non-null object
member_birth_year          174952 non-null float64
member_gender              174952 non-null object
bike_share_for_all_trip    174952 non-null object
start_date                 174952 non-null object
end_date                   174952 non-null object
start_day                  174952 non-null object
end_day                    174952 non-null object
dtypes: float64(5), int64(4), object(11)
memory usage: 28.0+ MB
```

```
In [13]: # Check for duplicated values
         df.duplicated().sum()
```

```
Out[13]: 0
```

```
In [14]: # Change the data type for bike_share_for_all_trip to be bool
         df.bike_share_for_all_trip = (df.bike_share_for_all_trip == 'Yes')
```

```
In [15]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 174952 entries, 0 to 183411
Data columns (total 20 columns):
duration_sec               174952 non-null int64
start_time                 174952 non-null object
end_time                   174952 non-null object
start_station_id           174952 non-null int64
start_station_name         174952 non-null object
start_station_latitude     174952 non-null float64
start_station_longitude    174952 non-null float64
end_station_id             174952 non-null int64
end_station_name           174952 non-null object
end_station_latitude       174952 non-null float64
end_station_longitude      174952 non-null float64
bike_id                    174952 non-null int64
user_type                  174952 non-null object
member_birth_year          174952 non-null float64
member_gender              174952 non-null object
bike_share_for_all_trip    174952 non-null bool
start_date                 174952 non-null object
end_date                   174952 non-null object
start_day                  174952 non-null object
end_day                    174952 non-null object
dtypes: bool(1), float64(5), int64(4), object(10)
```

```
memory usage: 26.9+ MB
```

```
In [16]: df.shape

Out[16]: (174952, 20)

In [17]: df.user_type.value_counts()

Out[17]: Subscriber    158386
         Customer       16566
         Name: user_type, dtype: int64

In [18]: # Create column age to investigate it further by subtracting 2019, where the data was f
         df['age'] = 2019 - df['member_birth_year']

In [19]: df['age'].sample(10)

Out[19]: 88042     44.0
         134816    55.0
         124885    36.0
         125478    28.0
         134975    40.0
         83790     33.0
         151243    36.0
         19174     34.0
         52205     37.0
         82249     25.0
         Name: age, dtype: float64

In [20]: df['age'] = df['age'].astype('int64')
         df['age'].sample(5)

Out[20]: 78548     33
         154594    40
         14956     33
         182903    44
         33224     30
         Name: age, dtype: int64

In [21]: print(df['age'].min(), df['age'].max())

18 141
```

The maximum age value shows that some of the registered birth year are incorrect, thus we will drop the rows with the age value > 100.

```
In [22]: df[df['age'] > 100].count()
```

```
Out[22]: duration_sec               72
         start_time                 72
         end_time                   72
         start_station_id           72
         start_station_name         72
         start_station_latitude     72
         start_station_longitude    72
         end_station_id             72
         end_station_name           72
         end_station_latitude       72
         end_station_longitude      72
         bike_id                    72
         user_type                  72
         member_birth_year          72
         member_gender              72
         bike_share_for_all_trip    72
         start_date                 72
         end_date                   72
         start_day                  72
         end_day                    72
         age                        72
         dtype: int64

In [23]: df = df[df['age'] < 100]
         df.head()

Out[23]:    duration_sec       start_time         end_time  start_station_id  \
         0         52185  17:32:10.145000  08:01:55.975000                21
         2         61854  12:13:13.218000  05:24:08.146000                86
         3         36490  17:54:26.010000  04:02:36.842000               375
         4          1585  23:54:18.549000  00:20:44.074000                 7
         5          1793  23:49:58.632000  00:19:51.760000                93

                                        start_station_name  start_station_latitude  \
         0  Montgomery St BART Station (Market St at 2nd St)                37.789625
         2                          Market St at Dolores St                37.769305
         3                          Grove St at Masonic Ave                37.774836
         4                              Frank H Ogawa Plaza                37.804562
         5                    4th St at Mission Bay Blvd S                37.770407

            start_station_longitude  end_station_id  \
         0              -122.400811              13
         2              -122.426826               3
         3              -122.446546              70
         4              -122.271738             222
         5              -122.391198             323

                                      end_station_name  end_station_latitude ...  \
```

```
0                Commercial St at Montgomery St             37.794231 ...
2  Powell St BART Station (Market St at 4th St)             37.786375 ...
3                        Central Ave at Fell St             37.773311 ...
4                         10th Ave at E 15th St             37.792714 ...
5                            Broadway at Kearny             37.798014 ...

   bike_id   user_type member_birth_year  member_gender  \
0     4902    Customer            1984.0           Male
2     5905    Customer            1972.0           Male
3     6638  Subscriber            1989.0          Other
4     4898  Subscriber            1974.0           Male
5     5200  Subscriber            1959.0           Male

  bike_share_for_all_trip  start_date    end_date start_day end_day age
0                   False  2019-02-28  2019-03-01  Thursday  Friday  35
2                   False  2019-02-28  2019-03-01  Thursday  Friday  47
3                   False  2019-02-28  2019-03-01  Thursday  Friday  30
4                    True  2019-02-28  2019-03-01  Thursday  Friday  45
5                   False  2019-02-28  2019-03-01  Thursday  Friday  60

[5 rows x 21 columns]
```

In [24]:
```
df[df['age'] > 100].count()
```

Out[24]:
```
duration_sec              0
start_time                0
end_time                  0
start_station_id          0
start_station_name        0
start_station_latitude    0
start_station_longitude   0
end_station_id            0
end_station_name          0
end_station_latitude      0
end_station_longitude     0
bike_id                   0
user_type                 0
member_birth_year         0
member_gender             0
bike_share_for_all_trip   0
start_date                0
end_date                  0
start_day                 0
end_day                   0
age                       0
dtype: int64
```

In [25]:
```
df.shape
```

Out[25]:
```
(174880, 21)
```

1.3.1 What is the structure of your dataset?

The dataset has 183412 bike rides that happened in the San Francisco Bay Area. The dataset has 16 features, some of them are: - `duration_sec`: The duration of the trip in seconds. - `start_time` and `end_time` for the bike rides. - `start_station_name` and `end_station_name`, as well as latitude and longitude. - `user_type` of either a subscriber or a customer. - Some information of the members such as their gender and birth year.

After my modifications the dataset has 174880 bike rides and 21 features. The added features are the `age`, `start_date`, `end_date`, `start_day`, and `end_day`, since I want to investigate them further.

1.3.2 What is/are the main feature(s) of interest in your dataset?

Some of the features of interest in this dataset are the duration of trips and its relation to other features like the user type, age, gender, and day of the week. Also, the most popular start and end stations.

1.3.3 What features in the dataset do you think will help support your investigation into your feature(s) of interest?

I expect that trip duration will have the strongest effect on each the start stations and end stations due to the crowded places which is expected to receive more rides. Another feature is the user type since I expect subscribers to spend more time on bikes than customers.

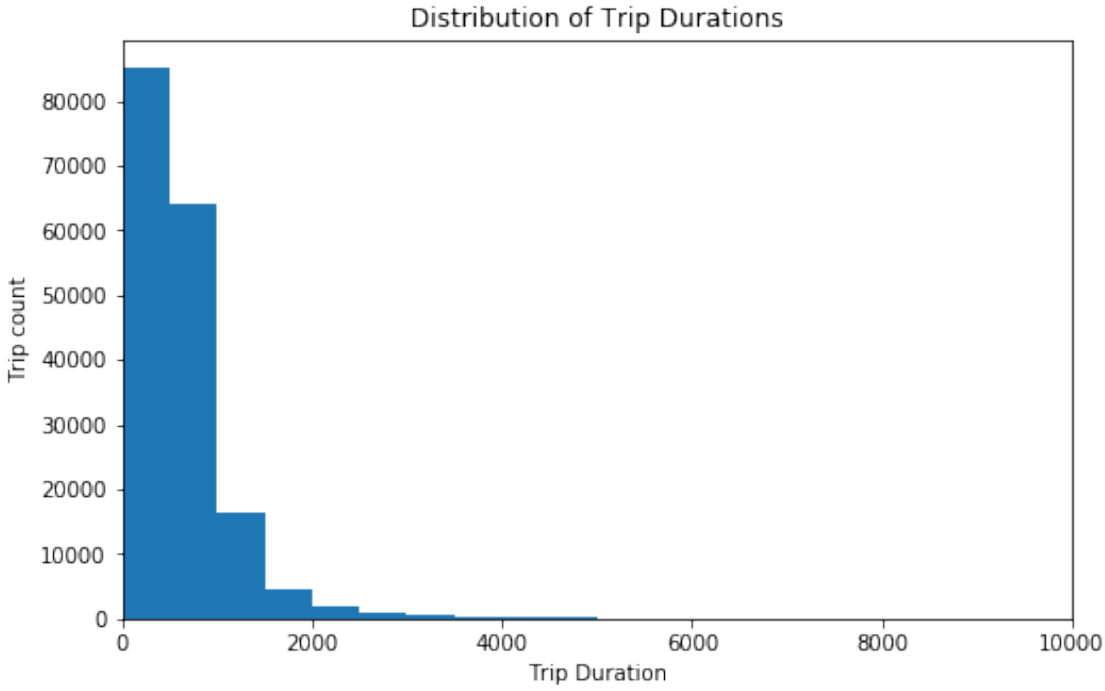
1.4 Univariate Exploration

This section provides a descriptive summary of the distribution of some variables. It also includes visual representations, such as histograms and bar plots, to help understand the shape and spread of the data. The goal of this section is to identify patterns, anomalies, and potential issues in the data that could impact the results of further analysis.

1.4.1 What is the distribution of `duration_sec`?

```
In [26]: max_duration = df['duration_sec'].max()
         max_duration

Out[26]: 84548

In [27]: min_duration = df['duration_sec'].min()
         min_duration

Out[27]: 61

In [28]: avg_duration = df['duration_sec'].mean()
         avg_duration

Out[28]: 704.02235818847214
```

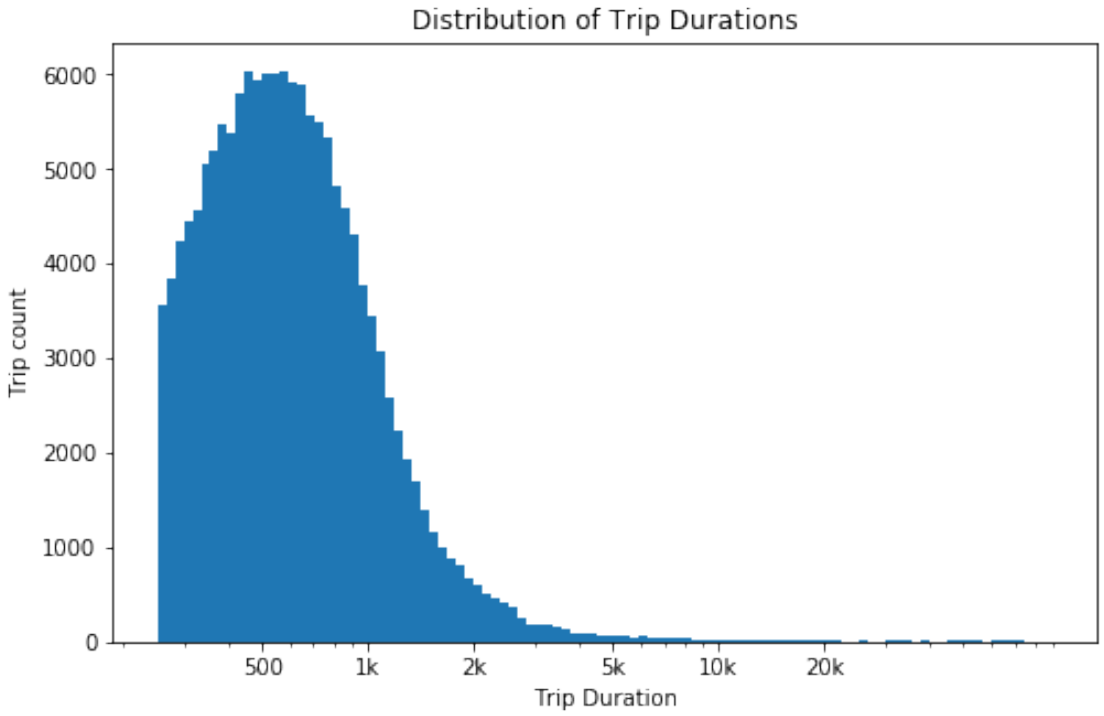
In [29]:
```
# Plot the histogram
plt.figure(figsize = [8, 5])
bin_size = 500
plt.hist(data = df, x = 'duration_sec', bins = np.arange(0, max_duration + bin_size, bi
plt.xlim([0, 10000])

# Add labels and formatting
plt.xlabel('Trip Duration')
plt.ylabel('Trip count')
plt.title('Distribution of Trip Durations')

plt.show()
```



There is a long tail in the distribution so we will put it in a logarthimic scale to make the distribution less skewed.

In [30]:
```
log_binsize = 0.025
bins = 10 ** np.arange(2.4, np.log10(max_duration) + log_binsize, log_binsize)

plt.figure(figsize=[8, 5])
plt.hist(data = df, x = 'duration_sec', bins = bins)
plt.xscale('log')
plt.xticks([500, 1e3, 2e3, 5e3, 1e4, 2e4], [500, '1k', '2k', '5k', '10k', '20k'])

# Add labels and formatting
```

```
plt.xlabel('Trip Duration')
plt.ylabel('Trip count')
plt.title('Distribution of Trip Durations')
plt.show()
```

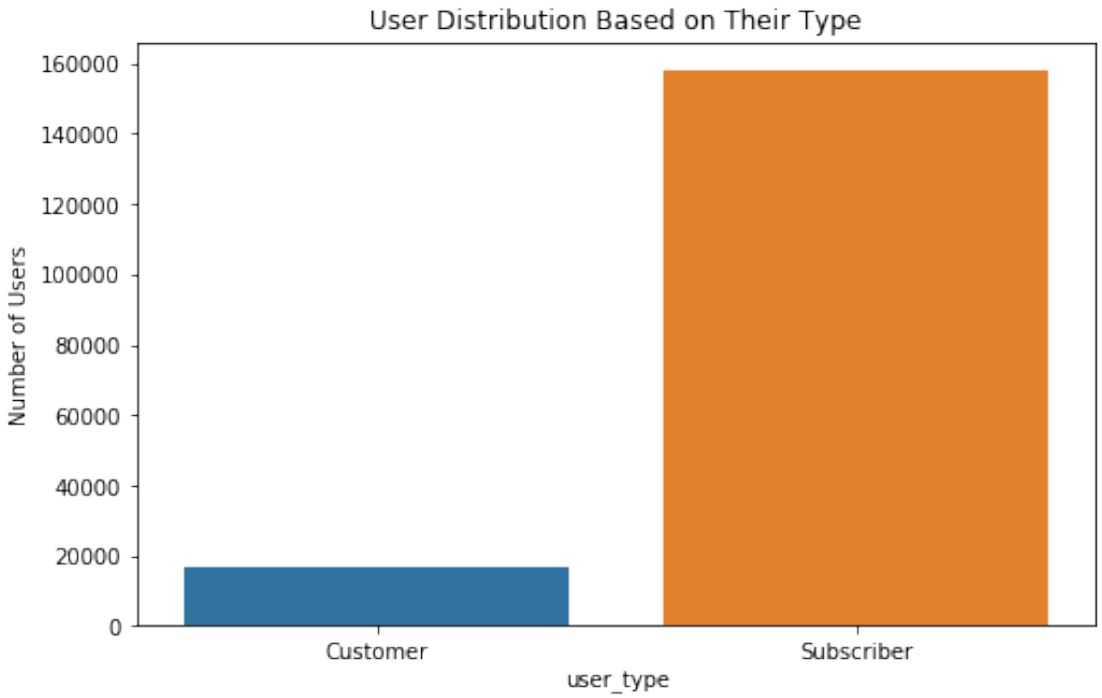


Most of the trip duration is focused in the lower spectrum. The majority of the numbers are less than 2k seconds, with the peak being about 600 seconds. When plotted on a log-scale, the distribution of trip durations looks right-skewed unimodal distribution. Overall, we conclude that trip durations are short on average.

1.4.2 What is the distribution of `user_type`?

```
In [31]: # Create a function to display the bar plots and prevent repetitive code

         def create_bar_plot(df, x, order = None, color = None):
             plt.figure(figsize = [8, 5])
             plot = sb.countplot(data = df, x = x, order = order, color = color)
             plt.xlabel(f'{x}')

In [32]: df['user_type'].value_counts()

Out[32]: Subscriber    158319
         Customer       16561
         Name: user_type, dtype: int64
```

```
In [33]: create_bar_plot(df, x = 'user_type')
         plt.ylabel('Number of Users')
         plt.title("User Distribution Based on Their Type")

         plt.show()
```

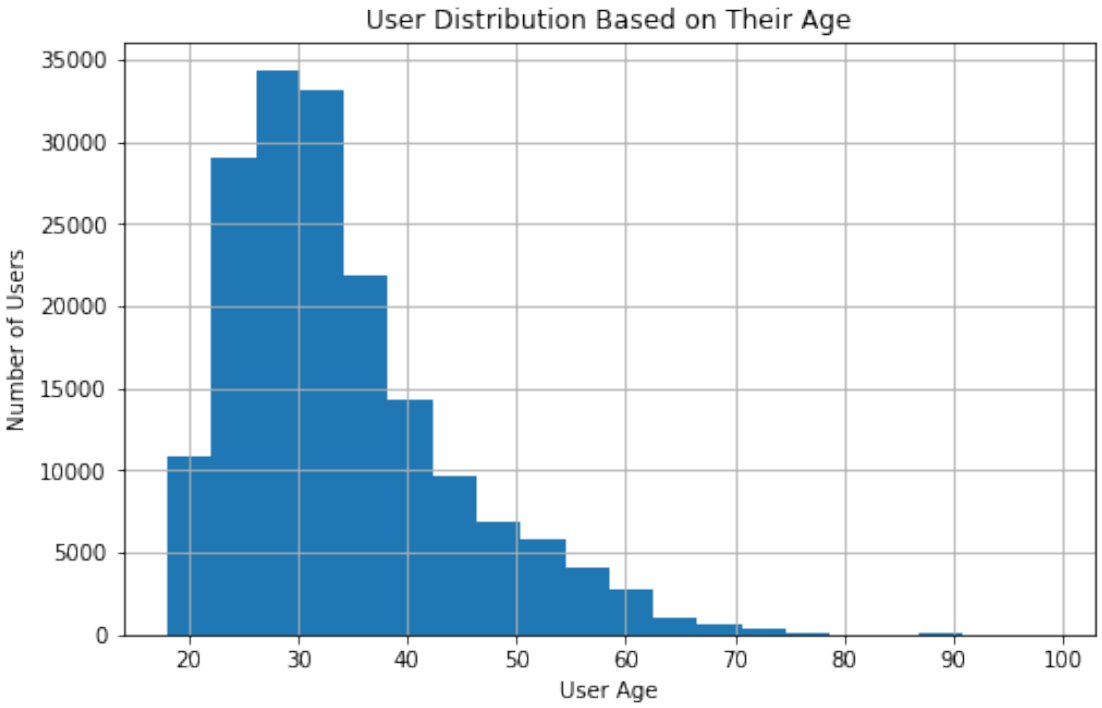


Most of the users are subscribers with a total of 158386, whereas the customers are way less with a total of 16566. This shows that the majority users of Ford Gobike system are subscribers.

1.4.3 What is the distribution of user's age?

```
In [34]: df['age'].min()

Out[34]: 18

In [35]: df['age'].max()

Out[35]: 99

In [36]: df['age'].mean()

Out[36]: 34.162042543458369

In [37]: plt.figure(figsize = [8, 5])
         plot = df['age'].hist(bins = 20)
         plt.xlabel('User Age')
```

```
plt.ylabel('Number of Users')
plt.title("User Distribution Based on Their Age")

plt.show()
```



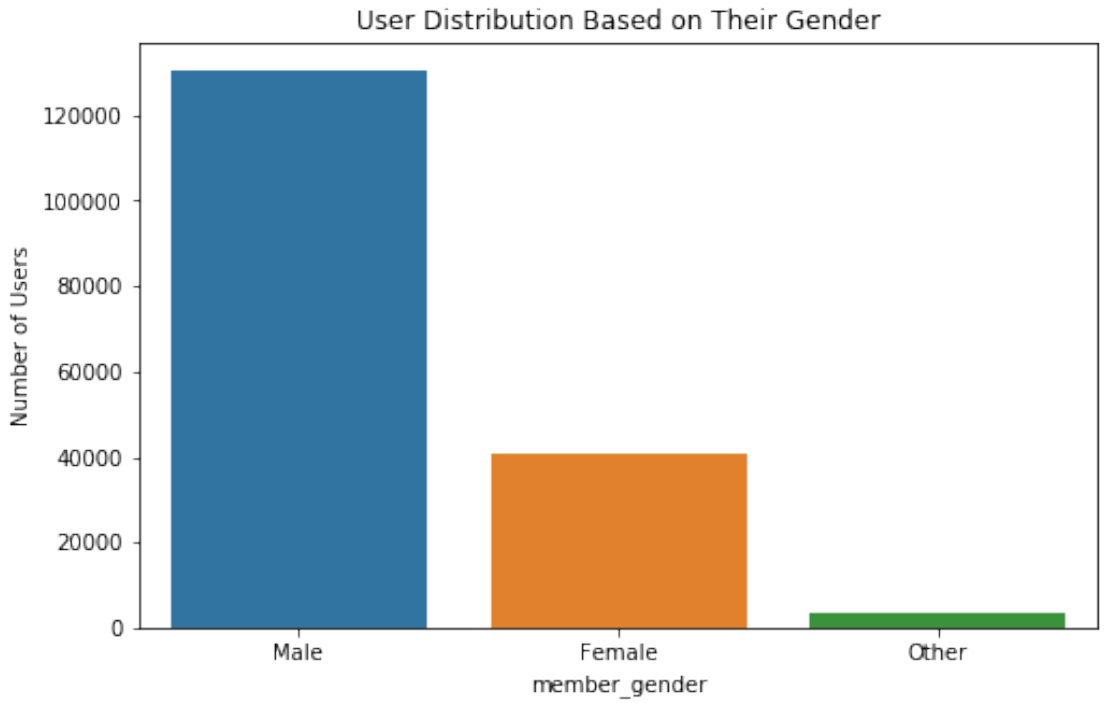
Most users are aged between 22 and 40 with the average being 34. This shows that most users are young.

1.4.4 What is the distribution of user's `gender`?

```
In [38]: df['member_gender'].value_counts()
```

```
Out[38]: Male      130443
         Female     40791
         Other       3646
         Name: member_gender, dtype: int64
```

```
In [39]: # Use function `create_bar_plot` defined earlier

         create_bar_plot(df, x = 'member_gender', order = ['Male', 'Female', 'Other'])
         plt.ylabel('Number of Users')
         plt.title("User Distribution Based on Their Gender")

         plt.show()
```



From the plot, we can see that the majority of users are male with a total of 130443 male users. Female users come next with a total of 40791 female users.

1.4.5 What are the top five start stations among users?

```
In [40]: df['start_station_name'].nunique()

Out[40]: 329

In [41]: df['start_station_name'].unique()

Out[41]: array(['Montgomery St BART Station (Market St at 2nd St)',
                'Market St at Dolores St', 'Grove St at Masonic Ave',
                'Frank H Ogawa Plaza', '4th St at Mission Bay Blvd S',
                'Palm St at Willow St', 'Washington St at Kearny St',
                'Post St at Kearny St', 'Jones St at Post St',
                'Civic Center/UN Plaza BART Station (Market St at McAllister St)',
                'Valencia St at 21st St', 'Bancroft Way at College Ave',
                'Howard St at Mary St', '22nd St at Dolores St',
                'Laguna St at Hayes St', '5th St at Folsom',
                'Telegraph Ave at 23rd St', 'Page St at Scott St',
                'Lake Merritt BART Station', 'West St at 40th St',
                'The Embarcadero at Sansome St', 'Folsom St at 9th St',
                'University Ave at Oxford St', 'MLK Jr Way at University Ave',
                'The Embarcadero at Bryant St', '17th St at Valencia St',
                'Valencia St at 16th St', 'Valencia St at 22nd St',
```

```
'Franklin Square', 'San Pablo Ave at MLK Jr Way',
'19th St at Mission St', 'Market St at 10th St',
'Folsom St at 13th St',
'San Francisco Ferry Building (Harry Bridges Plaza)',
'4th St at 16th St', 'Beale St at Harrison St',
'Broadway at Battery St', 'Cesar Chavez St at Dolores St',
'San Fernando St at 4th St', 'Grove St at Divisadero',
'Sanchez St at 17th St', 'Harmon St at Adeline St',
'Mission Playground', 'Davis St at Jackson St',
'Haste St at Telegraph Ave', 'Howard St at 8th St',
'Folsom St at 3rd St', 'Father Alfred E Boeddeker Park',
'Hubbell St at 16th St',
'San Francisco Public Library (Grove St at Hyde St)',
'Bancroft Way at Telegraph Ave', '19th Street BART Station',
'18th St at Noe St', 'Hyde St at Post St', '24th St at Market St',
'Vine St at Shattuck Ave',
'San Francisco Caltrain (Townsend St at 4th St)',
'Valencia St at Clinton Park',
'Union Square (Powell St at Post St)', 'Broderick St at Oak St',
'San Francisco Caltrain Station 2  (Townsend St at 4th St)',
'North Berkeley BART Station', 'Downtown Berkeley BART',
'Channing Way at Shattuck Ave', 'Fell St at Stanyan St',
'San Salvador St at 9th St', 'Marston Campbell Park',
'Oregon St at Adeline St', '11th St at Natoma St',
'Harrison St at 20th St', 'Haste St at College Ave',
'24th St at Bartlett St', 'Sanchez St at 15th St',
'Telegraph Ave at 19th St',
'Powell St BART Station (Market St at 5th St)',
'Jersey St at Castro St', 'Pierce St at Haight St',
'MacArthur BART Station', 'El Embarcadero at Grand Ave',
'23rd St at San Bruno Ave', 'Golden Gate Ave at Hyde St',
'S Van Ness Ave at Market St', 'Jackson Playground',
'San Fernando St at 7th St', 'West St at University Ave',
'Myrtle St at Polk St', 'Woolsey St at Sacramento St',
'Townsend St at 7th St', 'Harrison St at 17th St',
'West Oakland BART Station', 'Cyril Magnin St at Ellis St',
'Fulton St at Bancroft Way', '14th St at Mission St',
'San Pedro Square', 'Market St at Franklin St',
'Folsom St at 19th St', 'College Ave at Taft Ave',
'Rhode Island St at 17th St', 'Shattuck Ave at Hearst Ave',
'The Embarcadero at Vallejo St', 'The Embarcadero at Steuart St',
'Webster St at Grove St', 'Raymond Kimbell Playground',
'Victoria Manalo Draves Park', '20th St at Bryant St',
'S Park St at 3rd St', 'Lakeshore Ave at Trestle Glen Rd',
'Channing Way at San Pablo Ave', 'Mission Dolores Park',
'Lombard St at Columbus Ave', '17th St at Dolores St',
'Precita Park', 'Central Ave at Fell St', '4th St at Harrison St',
'Horton St at 40th St', 'Golden Gate Ave at Franklin St',
```

```
'Embarcadero BART Station (Beale St at Market St)',
'9th St at San Fernando St', '3rd St at Townsend St',
'McCoppin St at Valencia St', '13th St at Franklin St',
'Mission Bay Kids Park', 'Potrero Ave and Mariposa St',
'Emeryville Public Market', 'Union St at 10th St',
'Jackson St at 11th St', 'Broadway at Kearny',
'Paseo De San Antonio at 2nd St', 'Valencia St at Cesar Chavez St',
'Rockridge BART Station', '8th St at Brannan St',
'College Ave at Alcatraz Ave', '16th St Mission BART Station 2',
'San Jose Diridon Station', 'Masonic Ave at Turk St',
'17th & Folsom Street Park (17th St at Folsom St)',
'Grand Ave at Webster St', '7th St at Brannan St',
'Steuart St at Market St', 'Scott St at Golden Gate Ave',
'Parker St at Fulton St', 'Berkeley Civic Center',
'Clay St at Battery St', '11th St at Bryant St',
'Powell St BART Station (Market St at 4th St)',
'Doyle St at 59th St', '34th St at Telegraph Ave', 'Esprit Park',
'Emeryville Town Hall', 'Division St at Potrero Ave',
'Irwin St at 8th St', 'Pierce Ave at Market St',
'Howard St at Beale St', 'Washington St at 8th St',
'Dolores St at 15th St', 'Hearst Ave at Euclid Ave',
'Telegraph Ave at Ashby Ave', '8th St at Ringold St',
'14th St at Mandela Pkwy', 'Morrison Ave at Julian St',
'Commercial St at Montgomery St', 'Church St at Duboce Ave',
'Townsend St at 5th St', 'Valencia St at 24th St',
'16th St at Prosper St', '5th St at Virginia St',
"Webster St at O'Farrell St", 'Shattuck Ave at Telegraph Ave',
'Jackson St at 5th St', 'Berry St at 4th St',
'Telegraph Ave at Carleton St', 'Ellsworth St at Russell St',
'Adeline St at 40th St', 'Bay Pl at Vernon St',
'Russell St at College Ave', '22nd St Caltrain Station',
'Folsom St at 15th St', 'Snow Park', 'Ninth St at Heinz Ave',
'15th St at Potrero Ave', '23rd St at Tennessee St',
'McAllister St at Baker St', 'Bryant St at 2nd St',
'Mississippi St at 17th St', 'Ryland Park',
'Fountain Alley at S 2nd St', 'Ashby BART Station',
'Shattuck Ave at 51st St', 'Julian St at The Alameda',
'20th St at Dolores St', 'Broadway at Coronado Ave',
'Turk St at Fillmore St', 'Grand Ave at Santa Clara Ave',
'Eureka Valley Recreation Center', 'Parker Ave at McAllister St',
'Berry St at King St',
'Salesforce Transit Center (Natoma St at 2nd St)',
'San Antonio Park', 'Lakeside Dr at 14th St',
'16th St Mission BART', '2nd St at Townsend St',
'Stanford Ave at Hollis St', 'Broadway at 40th St',
'Mechanics Monument Plaza (Market St at Bush St)',
'Madison St at 17th St', 'Grand Ave at Perkins St',
'Garfield Square (25th St at Harrison St)', '53rd St at Hollis St',
```

```
'2nd St at Julian St', 'Telegraph Ave at Alcatraz Ave',
'San Francisco City Hall (Polk St at Grove St)',
'5th St at Brannan St', '10th St at Fallon St',
'Yerba Buena Center for the Arts (Howard St at 3rd St)',
'30th St at San Jose Ave', '29th St at Tiffany Ave',
'Webster St at 2nd St', 'Koshland Park', 'Jersey St at Church St',
'Santa Clara St at 7th St', 'Telegraph Ave at 58th St',
'Fruitvale BART Station', 'Addison St at Fourth St',
'Leavenworth St at Broadway', 'Telegraph Ave at 27th St',
'Potrero del Sol Park (25th St at Utah St)',
'Spear St at Folsom St', 'College Ave at Harwood Ave',
"O'Farrell St at Divisadero St", '1st St at Folsom St',
'Golden Gate Ave at Polk St', '5th St at San Salvador St',
'29th St at Church St', 'Alamo Square (Steiner St at Fulton St)',
'Autumn Parkway at Coleman Ave', 'Fulton St at Ashby Ave',
'Bryant St at 15th St', 'Howard St at 2nd St',
'19th St at Florida St', 'Market St at 45th St',
'Derby St at College Ave', 'Market St at Brockhurst St',
'California St at University Ave', 'MLK Jr Way at 14th St',
'Market St at 40th St', 'Julian St at 6th St', 'Cahill Park',
'San Jose City Hall', 'Virginia St at Shattuck Ave',
'Jack London Square', 'Webster St at 19th St',
'24th St at Chattanooga St', 'The Alameda at Bush St',
'49th St at Telegraph Ave', 'Broadway at 30th St',
'Bryant St at 6th St', 'Empire St at 1st St',
'China Basin St at 3rd St', '47th St at San Pablo Ave',
'Milvia St at Derby St', 'San Salvador St at 1st St',
'45th St at Manila', 'San Carlos St at Market St',
'San Pablo Ave at 27th St', 'Market St at Park St',
'Franklin St at 9th St', 'Almaden Blvd at San Fernando St',
'Oak St at 1st St', 'William St at 10th St',
'Isabella St at San Pablo Ave', 'Guerrero Park',
'10th St at University Ave', 'DeFremery Park',
'Fifth St at Delaware St', 'Williams Ave at 3rd St',
'4th Ave at E 12th St (Temporary Location)',
'Shattuck Ave at 55th St', '59th St at Horton St', 'SAP Center',
'37th St at West St', 'Almaden Blvd at Balbach St',
'65th St at Hollis St', 'Santa Clara St at Almaden Blvd',
'Ninth St at Parker St', 'Bushrod Park', 'Empire St at 7th St',
'Mendell St at Fairfax Ave', '16th St Depot',
'Newhall St at 3rd St', 'George St at 1st St',
'Mission St at 1st St', 'Duboce Park', 'Locust St at Grant St',
'32nd St at Adeline St', 'Mosswood Park',
'Delmas Ave and San Fernando St', 'Lane St at Revere Ave',
'2nd Ave at E 18th St', 'San Carlos St at 11th St',
'Williams Ave at Apollo St', 'MacArthur Blvd at Telegraph Ave',
'Bestor Art Park', 'College Ave at Bryant Ave',
'Miles Ave at Cavour St', 'Saint James Park',
```

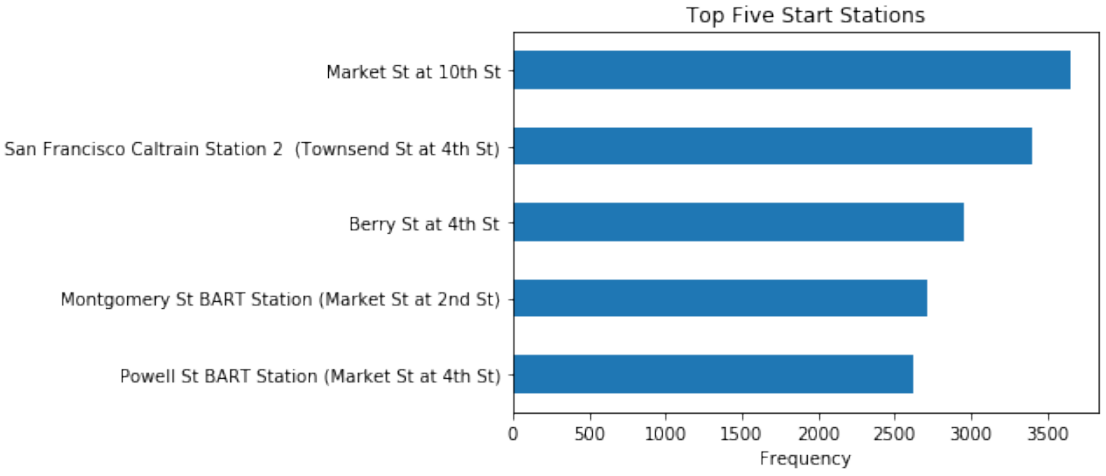
```
       '14th St at Filbert St', 'Foothill Blvd at Fruitvale Ave',
       'Market St at 8th St', 'Backesto Park (Jackson St at 13th St)',
       '10th Ave at E 15th St', 'Alcatraz Ave at Shattuck Ave',
       '55th St at Telegraph Ave', 'Genoa St at 55th St',
       'Dover St at 57th St', 'San Pablo Park',
       '6th Ave at E 12th St (Temporary Location)', 'Taylor St at 9th St',
       '27th St at MLK Jr Way', 'Foothill Blvd at Harrington Ave',
       '23rd Ave at Foothill Blvd', 'San Pedro St at Hedding St',
       '45th St at MLK Jr Way', '5th St at Taylor St',
       'Foothill Blvd at 42nd Ave', 'Willow St at Vine St',
       '26th Ave at International Blvd', 'Farnam St at Fruitvale Ave',
       '21st Ave at International Blvd', '2nd St at Folsom St'], dtype=object)
```

```
In [42]: df['start_station_name'].value_counts()
```

```
Out[42]: Market St at 10th St                                          3649
         San Francisco Caltrain Station 2  (Townsend St at 4th St)     3406
         Berry St at 4th St                                            2951
         Montgomery St BART Station (Market St at 2nd St)              2711
         Powell St BART Station (Market St at 4th St)                  2620
         San Francisco Caltrain (Townsend St at 4th St)                2572
         San Francisco Ferry Building (Harry Bridges Plaza)            2540
         Howard St at Beale St                                         2216
         Steuart St at Market St                                       2191
         Powell St BART Station (Market St at 5th St)                  2144
         The Embarcadero at Sansome St                                 1975
         Bancroft Way at Telegraph Ave                                 1761
         Bancroft Way at College Ave                                   1712
         2nd St at Townsend St                                         1702
         Beale St at Harrison St                                       1678
         3rd St at Townsend St                                         1675
         Embarcadero BART Station (Beale St at Market St)              1646
         4th St at Mission Bay Blvd S                                  1496
         Townsend St at 7th St                                         1479
         Civic Center/UN Plaza BART Station (Market St at McAllister St)  1476
         The Embarcadero at Steuart St                                 1420
         Downtown Berkeley BART                                        1322
         Post St at Kearny St                                          1311
         4th St at 16th St                                             1274
         Howard St at 8th St                                           1257
         19th Street BART Station                                      1231
         Esprit Park                                                   1226
         Rhode Island St at 17th St                                    1224
         Hearst Ave at Euclid Ave                                      1176
         8th St at Brannan St                                          1154
                                                                       ...
         45th St at MLK Jr Way                                           54
         27th St at MLK Jr Way                                           54
```

```
Mendell St at Fairfax Ave                                53
San Antonio Park                                         53
Locust St at Grant St                                    49
Williams Ave at 3rd St                                   48
Delmas Ave and San Fernando St                           47
San Carlos St at Market St                               37
Lane St at Revere Ave                                    36
Foothill Blvd at Harrington Ave                          35
Almaden Blvd at Balbach St                               32
Mission St at 1st St                                     32
George St at 1st St                                      31
Oak St at 1st St                                         30
Empire St at 7th St                                      28
SAP Center                                               27
Williams Ave at Apollo St                                25
Foothill Blvd at 42nd Ave                                23
26th Ave at International Blvd                           19
San Pedro St at Hedding St                               19
23rd Ave at Foothill Blvd                                17
Backesto Park (Jackson St at 13th St)                    17
Leavenworth St at Broadway                               16
Taylor St at 9th St                                      13
Farnam St at Fruitvale Ave                                9
Willow St at Vine St                                      9
Parker Ave at McAllister St                               7
21st Ave at International Blvd                            4
Palm St at Willow St                                      3
16th St Depot                                             2
Name: start_station_name, Length: 329, dtype: int64
```

```
In [43]: plt.title('Top Five Start Stations')
         plt.xlabel('Frequency')
         df['start_station_name'].value_counts(ascending=True).tail(5).plot.barh(color = '#1f77b
```



We can see that the top five start stations among users, respectively, are: - Market St at 10th St - San Francisco Caltrain Station 2 (Townsend St at 4th St) - Berry St at 4th St - Montgomery St BART Station (Market St at 2nd St) - Powell St BART Station (Market St at 4th St)

1.4.6 What are the top five end stations among users?

```
In [44]: df['end_station_name'].nunique()
```

```
Out[44]: 329
```

```
In [45]: df['end_station_name'].unique()
```

```
Out[45]: array(['Commercial St at Montgomery St',
                'Powell St BART Station (Market St at 4th St)',
                'Central Ave at Fell St', '10th Ave at E 15th St',
                'Broadway at Kearny', 'San Jose Diridon Station',
                'Valencia St at 21st St', 'Mission Playground',
                'San Francisco Public Library (Grove St at Hyde St)',
                'Bryant St at 2nd St', 'Channing Way at Shattuck Ave',
                '8th St at Ringold St', 'Broderick St at Oak St',
                'Potrero Ave and Mariposa St', 'Market St at Franklin St',
                'Telegraph Ave at 23rd St', '17th St at Dolores St',
                '6th Ave at E 12th St (Temporary Location)',
                'McAllister St at Baker St', 'Telegraph Ave at Carleton St',
                'Genoa St at 55th St', 'Grand Ave at Perkins St',
                'San Francisco Ferry Building (Harry Bridges Plaza)',
                'Folsom St at 9th St', 'Channing Way at San Pablo Ave',
                'Shattuck Ave at Hearst Ave', '2nd St at Townsend St',
                'Pierce St at Haight St',
                'Potrero del Sol Park (25th St at Utah St)',
                'Valencia St at 22nd St', 'Jackson Playground',
                'Dolores St at 15th St', '29th St at Church St',
                '19th St at Mission St', 'Bay Pl at Vernon St',
                'Post St at Kearny St',
                'Yerba Buena Center for the Arts (Howard St at 3rd St)',
                '4th St at Mission Bay Blvd S', 'Father Alfred E Boeddeker Park',
                'Market St at 10th St', '24th St at Chattanooga St',
                'Pierce Ave at Market St', 'Fell St at Stanyan St',
                '17th St at Valencia St', 'San Pablo Ave at 27th St',
                'Howard St at Mary St', 'Victoria Manalo Draves Park',
                'Davis St at Jackson St', 'Jersey St at Church St',
                'Haste St at Telegraph Ave', 'Eureka Valley Recreation Center',
                'Washington St at Kearny St', 'Grove St at Divisadero',
                'Berry St at 4th St', 'Parker St at Fulton St',
                'El Embarcadero at Grand Ave', 'Lake Merritt BART Station',
                'Hyde St at Post St', '24th St at Market St',
                '5th St at Brannan St', '24th St at Bartlett St',
```

```
'Townsend St at 5th St', 'Addison St at Fourth St',
'Broadway at Battery St', 'Market St at Dolores St',
'5th St at Virginia St', 'Marston Campbell Park',
'University Ave at Oxford St', 'Valencia St at 24th St',
'Valencia St at Cesar Chavez St', 'Ryland Park', 'Precita Park',
'Derby St at College Ave', 'Jersey St at Castro St',
'11th St at Natoma St', '45th St at MLK Jr Way',
'Valencia St at 16th St',
'San Francisco Caltrain Station 2  (Townsend St at 4th St)',
'Montgomery St BART Station (Market St at 2nd St)',
'18th St at Noe St', '37th St at West St', 'Newhall St at 3rd St',
'Haste St at College Ave', 'Cyril Magnin St at Ellis St',
'Beale St at Harrison St', 'Fulton St at Bancroft Way',
'San Fernando St at 4th St',
'Garfield Square (25th St at Harrison St)',
'29th St at Tiffany Ave', 'Bancroft Way at College Ave',
'Ashby BART Station', '11th St at Bryant St',
'14th St at Mandela Pkwy', 'Howard St at 8th St',
'Leavenworth St at Broadway', 'Locust St at Grant St',
'Lombard St at Columbus Ave', 'Sanchez St at 17th St',
'45th St at Manila', '23rd St at San Bruno Ave',
'Morrison Ave at Julian St', 'Sanchez St at 15th St',
'Koshland Park', 'Harrison St at 20th St', '2nd Ave at E 18th St',
'Steuart St at Market St', 'Church St at Duboce Ave',
'Page St at Scott St', 'Bancroft Way at Telegraph Ave',
'Mission Bay Kids Park', 'Folsom St at 3rd St',
'Valencia St at Clinton Park', 'Grand Ave at Santa Clara Ave',
'19th Street BART Station', 'Folsom St at 19th St',
'West Oakland BART Station', 'S Park St at 3rd St',
'5th St at Folsom',
'Embarcadero BART Station (Beale St at Market St)',
'Howard St at 2nd St', 'The Embarcadero at Sansome St',
'Backesto Park (Jackson St at 13th St)', 'Esprit Park',
'Myrtle St at Polk St', 'Franklin Square', 'Empire St at 7th St',
'Lakeside Dr at 14th St', 'Laguna St at Hayes St',
'65th St at Hollis St', '4th St at 16th St',
'49th St at Telegraph Ave', '16th St Mission BART Station 2',
'Lane St at Revere Ave', 'MLK Jr Way at University Ave',
'2nd St at Julian St', 'Webster St at Grove St',
'Telegraph Ave at Ashby Ave', 'Bryant St at 6th St',
'20th St at Dolores St',
'Powell St BART Station (Market St at 5th St)',
'San Francisco Caltrain (Townsend St at 4th St)',
'8th St at Brannan St', 'Downtown Berkeley BART',
'North Berkeley BART Station', 'Turk St at Fillmore St',
'Woolsey St at Sacramento St', '4th St at Harrison St',
'The Embarcadero at Bryant St', "O'Farrell St at Divisadero St",
'Grove St at Masonic Ave', 'Hubbell St at 16th St',
```

```
'Civic Center/UN Plaza BART Station (Market St at McAllister St)',
'3rd St at Townsend St', 'Fountain Alley at S 2nd St',
'China Basin St at 3rd St', '59th St at Horton St',
'Mission Dolores Park', 'San Carlos St at 11th St',
'Jackson St at 11th St', '22nd St Caltrain Station',
'Townsend St at 7th St', '7th St at Brannan St',
'Webster St at 2nd St', 'Ellsworth St at Russell St',
"Webster St at O'Farrell St", 'Harmon St at Adeline St',
'1st St at Folsom St', 'Vine St at Shattuck Ave',
'Stanford Ave at Hollis St', 'Jones St at Post St',
'West St at University Ave', 'Paseo De San Antonio at 2nd St',
'Duboce Park', 'The Embarcadero at Steuart St',
'Russell St at College Ave', 'Golden Gate Ave at Hyde St',
'Berkeley Civic Center', '47th St at San Pablo Ave',
'George St at 1st St', '53rd St at Hollis St', 'West St at 40th St',
'15th St at Potrero Ave', 'Division St at Potrero Ave',
'San Pablo Ave at MLK Jr Way', 'Jackson St at 5th St',
'Union Square (Powell St at Post St)',
'4th Ave at E 12th St (Temporary Location)', 'Bushrod Park',
'Rhode Island St at 17th St', 'Folsom St at 13th St',
'Virginia St at Shattuck Ave', '16th St Mission BART',
'Lakeshore Ave at Trestle Glen Rd', 'Masonic Ave at Turk St',
'Harrison St at 17th St', 'McCoppin St at Valencia St',
'17th & Folsom Street Park (17th St at Folsom St)',
'10th St at Fallon St', '34th St at Telegraph Ave',
'The Alameda at Bush St', '9th St at San Fernando St',
'20th St at Bryant St', 'Howard St at Beale St',
'Cesar Chavez St at Dolores St', '55th St at Telegraph Ave',
'S Van Ness Ave at Market St', 'Scott St at Golden Gate Ave',
'14th St at Mission St', 'Mississippi St at 17th St',
'Alamo Square (Steiner St at Fulton St)', 'Shattuck Ave at 51st St',
'MacArthur BART Station', 'Madison St at 17th St',
'Horton St at 40th St', 'Hearst Ave at Euclid Ave',
'Folsom St at 15th St', 'Alcatraz Ave at Shattuck Ave',
'San Fernando St at 7th St', 'MLK Jr Way at 14th St',
'Milvia St at Derby St', 'College Ave at Alcatraz Ave',
'Washington St at 8th St', 'Guerrero Park',
'Oregon St at Adeline St', 'Parker Ave at McAllister St',
'23rd St at Tennessee St', 'Clay St at Battery St',
'Broadway at 40th St',
'Salesforce Transit Center (Natoma St at 2nd St)',
'Telegraph Ave at 19th St', 'Emeryville Public Market',
'Golden Gate Ave at Polk St', 'Telegraph Ave at 58th St',
'Foothill Blvd at Harrington Ave', 'The Embarcadero at Vallejo St',
'16th St at Prosper St', 'Berry St at King St',
'Broadway at Coronado Ave', 'Market St at 45th St',
'Mechanics Monument Plaza (Market St at Bush St)',
'Dover St at 57th St', '19th St at Florida St',
```

```
       'Miles Ave at Cavour St', 'Rockridge BART Station',
       'Fifth St at Delaware St', 'College Ave at Harwood Ave',
       'California St at University Ave', '5th St at San Salvador St',
       'Mosswood Park', 'William St at 10th St', 'Union St at 10th St',
       '5th St at Taylor St', 'Julian St at The Alameda',
       'Irwin St at 8th St', 'Market St at Brockhurst St',
       'Adeline St at 40th St', '30th St at San Jose Ave',
       'Spear St at Folsom St', '27th St at MLK Jr Way',
       'San Francisco City Hall (Polk St at Grove St)',
       '22nd St at Dolores St', 'Frank H Ogawa Plaza',
       'Golden Gate Ave at Franklin St', 'Broadway at 30th St',
       'Bryant St at 15th St', 'Grand Ave at Webster St',
       'Julian St at 6th St', 'Shattuck Ave at 55th St',
       'Santa Clara St at 7th St', '14th St at Filbert St',
       'Emeryville Town Hall', 'Cahill Park', 'Raymond Kimbell Playground',
       'Autumn Parkway at Coleman Ave', 'Isabella St at San Pablo Ave',
       'San Salvador St at 9th St', 'Telegraph Ave at 27th St',
       '13th St at Franklin St', 'Doyle St at 59th St',
       'Jack London Square', 'SAP Center', 'Telegraph Ave at Alcatraz Ave',
       'San Carlos St at Market St', '10th St at University Ave',
       'Ninth St at Heinz Ave', 'Market St at 40th St',
       '23rd Ave at Foothill Blvd', 'Bestor Art Park',
       '32nd St at Adeline St', 'DeFremery Park', 'San Pedro Square',
       'San Salvador St at 1st St', 'Fulton St at Ashby Ave',
       'Ninth St at Parker St', 'Taylor St at 9th St',
       'Empire St at 1st St', 'Franklin St at 9th St',
       'Webster St at 19th St', 'San Pablo Park',
       'Shattuck Ave at Telegraph Ave', 'College Ave at Taft Ave',
       'Market St at 8th St', 'Snow Park', 'San Antonio Park',
       'San Jose City Hall', 'Delmas Ave and San Fernando St',
       'Mendell St at Fairfax Ave', 'Santa Clara St at Almaden Blvd',
       'College Ave at Bryant Ave', 'Foothill Blvd at Fruitvale Ave',
       'Palm St at Willow St', 'Saint James Park', 'Market St at Park St',
       'Almaden Blvd at Balbach St', 'Almaden Blvd at San Fernando St',
       'Foothill Blvd at 42nd Ave', 'Fruitvale BART Station',
       'MacArthur Blvd at Telegraph Ave', 'Williams Ave at Apollo St',
       'Williams Ave at 3rd St', 'Mission St at 1st St',
       'San Pedro St at Hedding St', 'Oak St at 1st St',
       'Farnam St at Fruitvale Ave', '26th Ave at International Blvd',
       '16th St Depot', 'Willow St at Vine St',
       '21st Ave at International Blvd', '2nd St at Folsom St'], dtype=object)
```

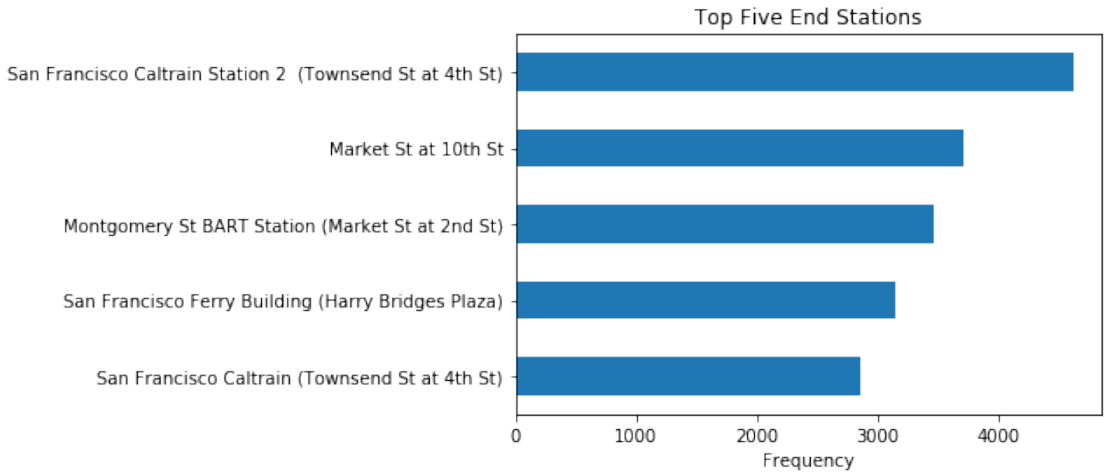
In [46]:
```
df['end_station_name'].value_counts()
```

Out[46]:
```
San Francisco Caltrain Station 2  (Townsend St at 4th St)    4622
Market St at 10th St                                         3709
Montgomery St BART Station (Market St at 2nd St)             3461
San Francisco Ferry Building (Harry Bridges Plaza)           3151
```

San Francisco Caltrain (Townsend St at 4th St)	2860
Powell St BART Station (Market St at 4th St)	2854
Berry St at 4th St	2782
The Embarcadero at Sansome St	2341
Steuart St at Market St	2264
Powell St BART Station (Market St at 5th St)	2153
Howard St at Beale St	1970
Bancroft Way at Telegraph Ave	1780
Beale St at Harrison St	1780
Civic Center/UN Plaza BART Station (Market St at McAllister St)	1733
2nd St at Townsend St	1667
3rd St at Townsend St	1628
Embarcadero BART Station (Beale St at Market St)	1622
4th St at Mission Bay Blvd S	1590
Townsend St at 7th St	1440
The Embarcadero at Steuart St	1414
19th Street BART Station	1392
Post St at Kearny St	1366
Downtown Berkeley BART	1321
Rhode Island St at 17th St	1271
8th St at Brannan St	1270
Folsom St at 3rd St	1234
Howard St at 8th St	1216
Esprit Park	1182
4th St at 16th St	1135
Spear St at Folsom St	1116
	...
Williams Ave at 3rd St	50
Market St at 40th St	48
27th St at MLK Jr Way	48
Lane St at Revere Ave	46
Locust St at Grant St	45
Empire St at 1st St	45
San Carlos St at Market St	42
Bestor Art Park	38
SAP Center	37
10th Ave at E 15th St	35
Almaden Blvd at Balbach St	34
George St at 1st St	33
San Pedro St at Hedding St	30
Oak St at 1st St	28
Empire St at 7th St	25
Mission St at 1st St	24
Williams Ave at Apollo St	20
Foothill Blvd at 42nd Ave	20
23rd Ave at Foothill Blvd	19
26th Ave at International Blvd	19
Backesto Park (Jackson St at 13th St)	18

```
Foothill Blvd at Harrington Ave                                  16
Leavenworth St at Broadway                                       12
Taylor St at 9th St                                              11
Farnam St at Fruitvale Ave                                       10
Parker Ave at McAllister St                                       9
Palm St at Willow St                                              7
16th St Depot                                                     6
Willow St at Vine St                                              5
21st Ave at International Blvd                                    5
Name: end_station_name, Length: 329, dtype: int64
```

```
In [47]: plt.title('Top Five End Stations')
         plt.xlabel('Frequency')
         df['end_station_name'].value_counts(ascending=True).tail(5).plot.barh(color = '#1f77b4'
```

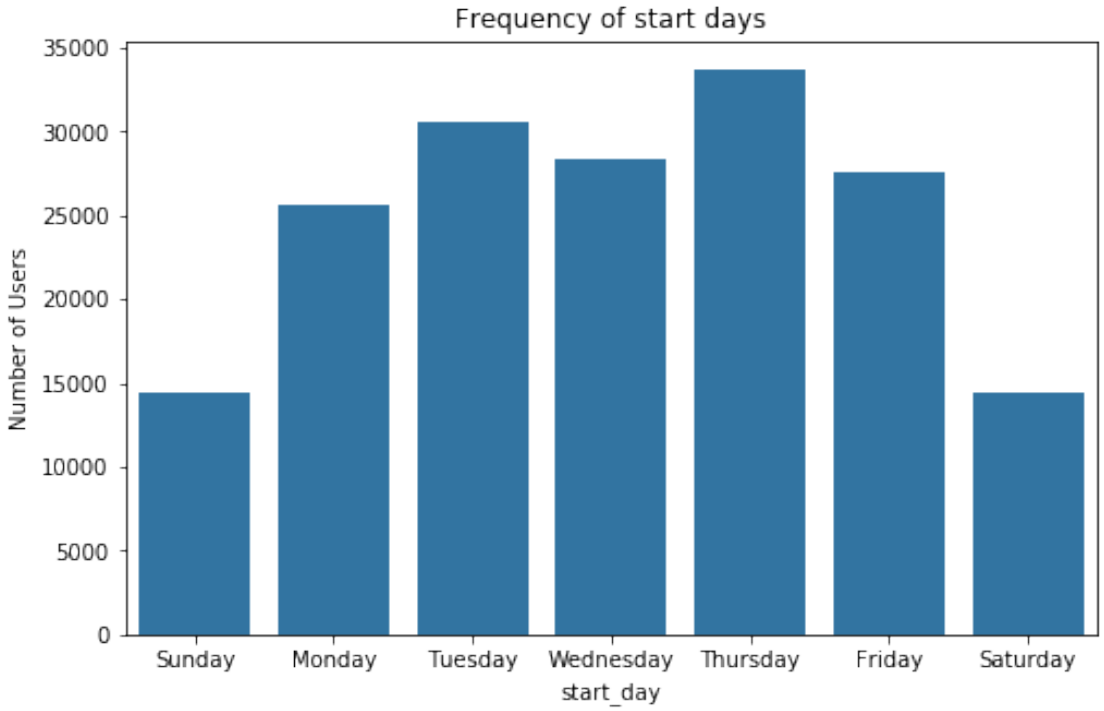


We can see that the top five end stations among users, respectively, are: - San Francisco Caltrain Station 2 (Townsend St at 4th St) - Market St at 10th St - Montgomery St BART Station (Market St at 2nd St) - San Francisco Ferry Building (Harry Bridges Plaza) - San Francisco Caltrain (Townsend St at 4th St)

We can observe that some stations are popular as start and end stations.

1.4.7 What is the most frequent start day of the week that users ride their bikes on?

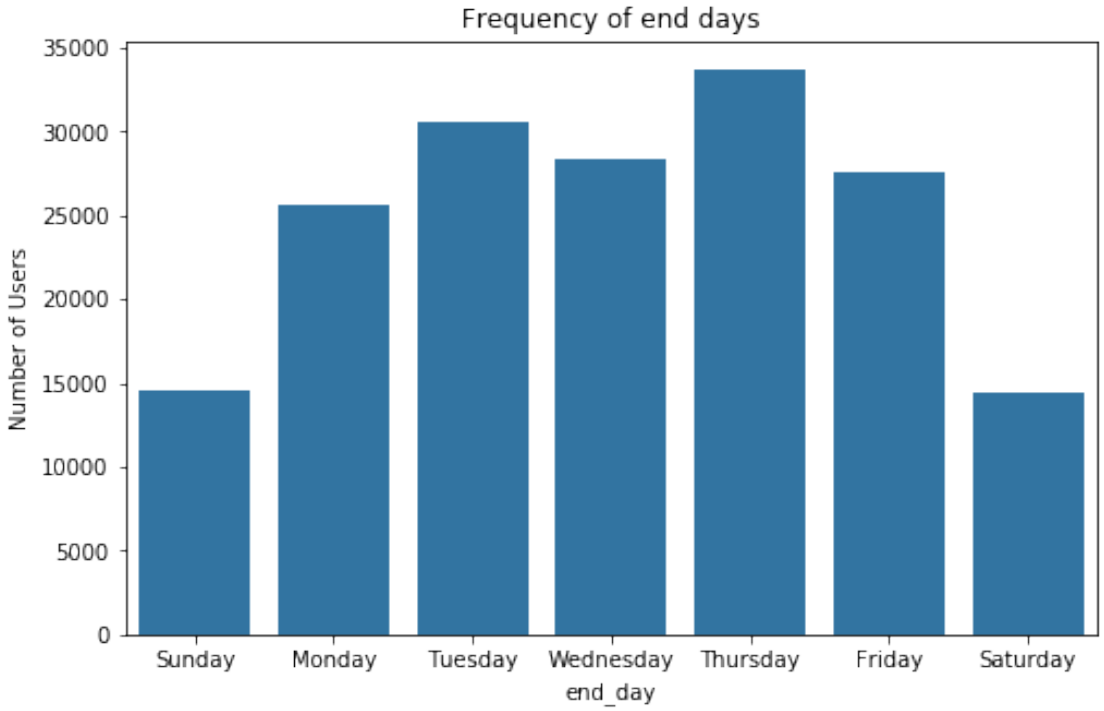
```
In [48]: # Use function defined earlier

         create_bar_plot(df, x = 'start_day',
                         order = ['Sunday', 'Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Frida
         plt.ylabel('Number of Users')
         plt.title("Frequency of start days")

         plt.show()
```



From the plot above, we can see that Thursday is the most frequent start day of the week, followed by Tuesday then Wednesday. We can also see that the bike rides drop on the weekends (Saturday and Sunday) which could mean that people highly use bikes for work instead of pleasure.

1.4.8 What is the most frequent end day of the week that users ride their bikes on?

In [49]:
```
# Use function defined earlier

create_bar_plot(df, x = 'end_day',
                order = ['Sunday', 'Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Frida
                color = '#1f77b4')
plt.ylabel('Number of Users')
plt.title("Frequency of end days")

plt.show()
```



We can see that there are not much difference in the frequency of the start and end days. This concludes that people rent their bikes daily instead of a multiple days basis.

1.4.9 Discuss the distribution(s) of your variable(s) of interest. Were there any unusual points? Did you need to perform any transformations?

From the first visualization regarding the duration of trips, the duration takes a large amount of values and is concentrated to a tail so we transformed it to a logarthmic scale and found that peak occurs at around 600 seconds starting from 0 and then distribution starts to drop and does not regain any more peak value.

1.4.10 Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

The birth year is converted by substracting the birth year of the users from 2019, the year of the published dataset. This gives us a distibution for age, this action is performed as age gives a better perception regarding trip duration dependency. Also, we modified the `start_time` and `end_time` to separate the date from the time and extracting the day of the week from the date column.

1.5 Bivariate Exploration

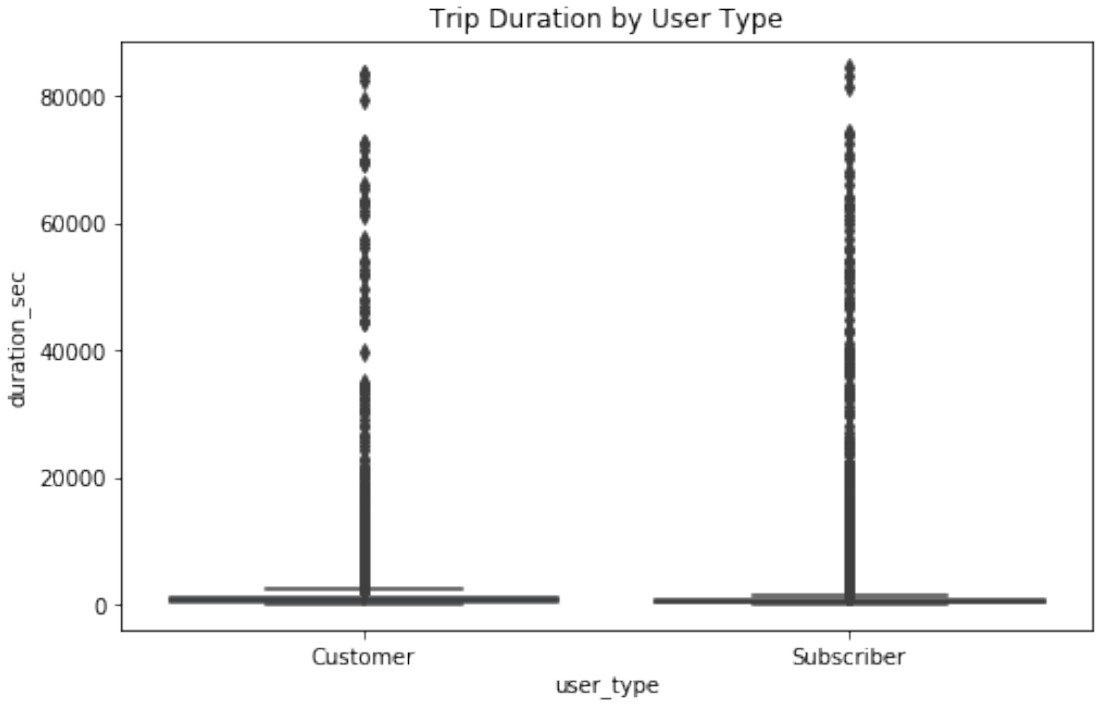
This section investigate relationships between pairs of variables. It also includes visual representations, such as clustered bar, bar plot, box plots, and scatter plot, to help understand the relation and spread of the data.

1.5.1 What is the average trip duration based on user type?

In [50]:
```
# Add user types into variables

customers = df['user_type'] == 'Customer'
subscribers = df['user_type'] == 'Subscriber'
```

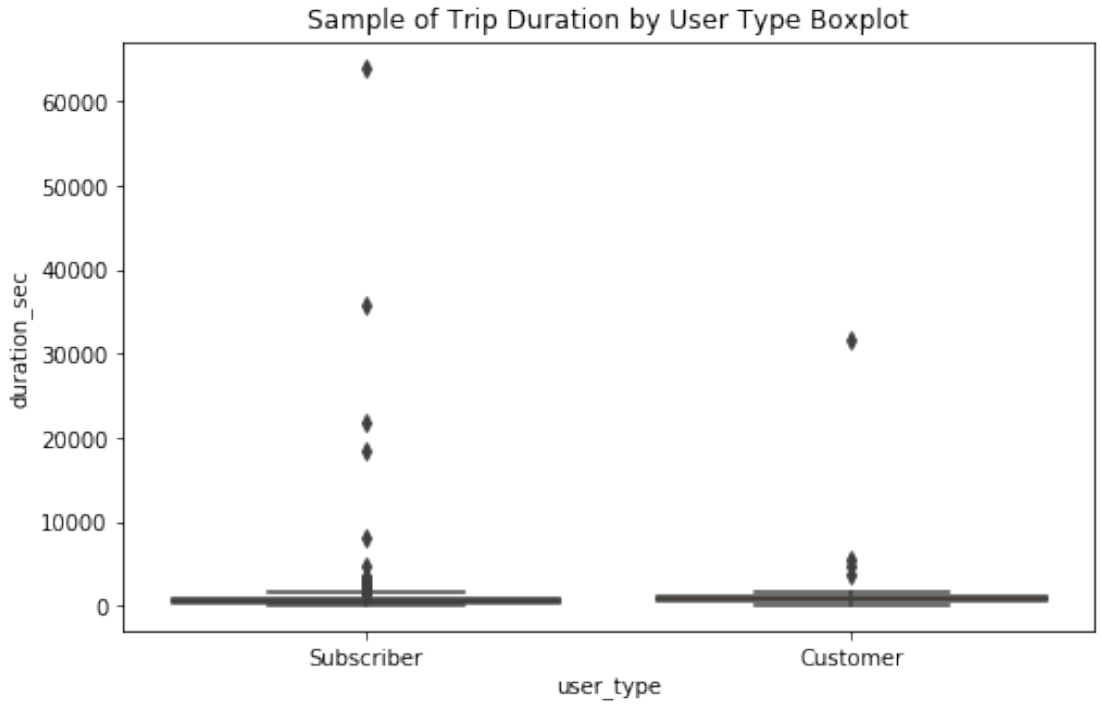
In [51]:
```
plt.figure(figsize = [8, 5])
plt.title('Trip Duration by User Type')
sb.boxplot(data = df, x = 'user_type', y = 'duration_sec')
```

Out[51]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f249a9b2668>
```

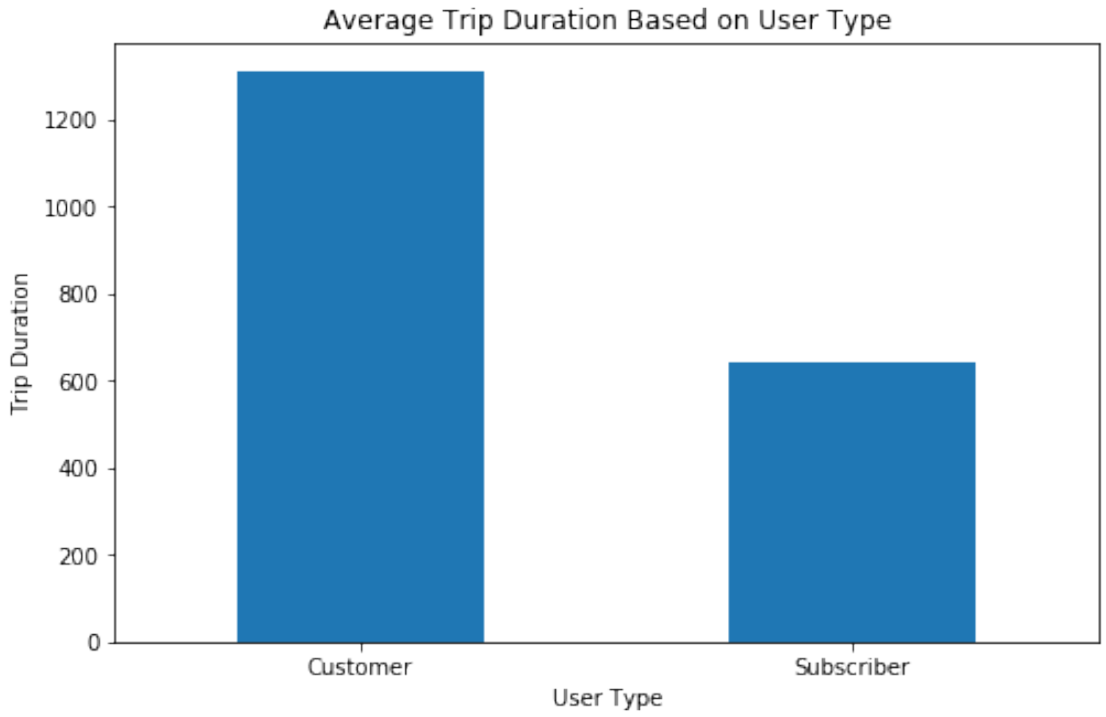


In [52]:
```
# Use a sample of data to make better visualization of the box plot

sample = df.sample(1000)

plt.figure(figsize = [8, 5])
plt.title('Sample of Trip Duration by User Type Boxplot')
sb.boxplot(data = sample, x = 'user_type', y = 'duration_sec')
```

Out[52]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f249a8e38d0>
```



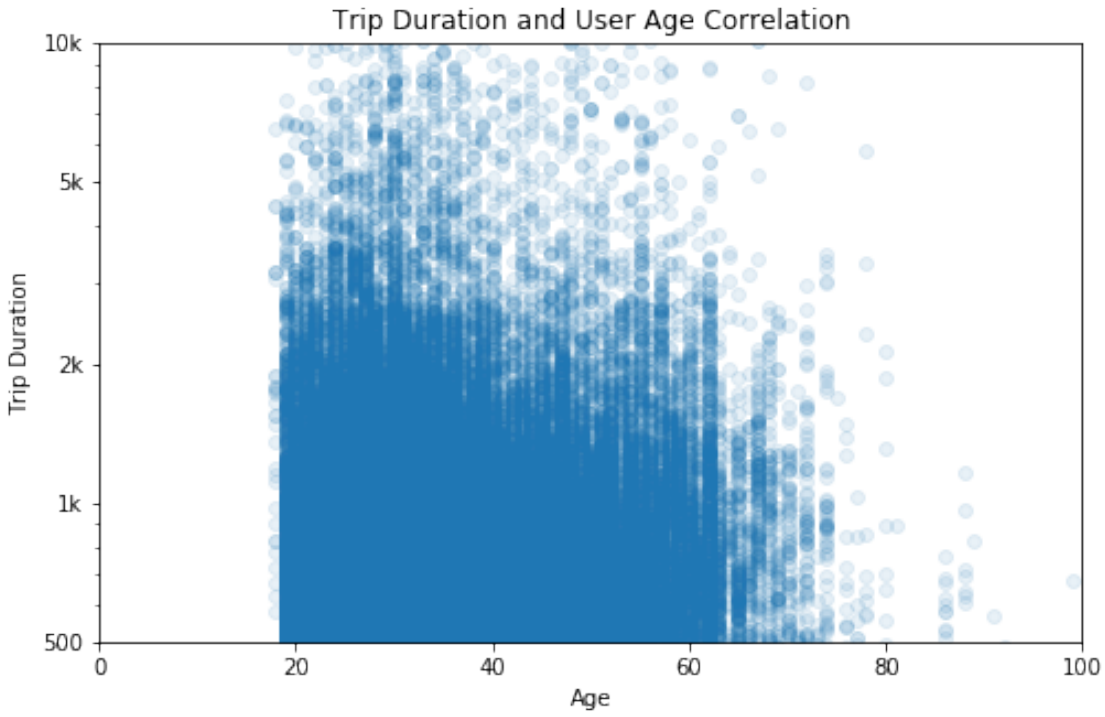

```
In [53]: # Average trip duration for Customer

         df[customers]['duration_sec'].mean()

Out[53]: 1310.7782138759737

In [54]: # Average trip duration for Subscriber

         df[subscribers]['duration_sec'].mean()

Out[54]: 640.55250475306184

In [55]: # Plot trip duration comparison

         duration_user = df[['user_type', 'duration_sec']].groupby('user_type').mean()
         duration_user.plot(kind = 'bar', figsize = [8, 5], legend= None)
         plt.title("Average Trip Duration Based on User Type")
         plt.ylabel("Trip Duration")
         plt.xlabel('User Type')
         plt.xticks(rotation=0)

         plt.show()
```



Based on the plots above, we can see that the average trip duration of users is higher for customers in comparison with subscribers. Also, we can see that there exists some oulier values for both customers and subscribers.

1.5.2 What is the relation between trip duration and users' age?

In [56]:
```
# scatter plot of trip duration vs. users' age, with log transform on trip duration axi

plt.figure(figsize = [8, 5])
plt.scatter(df['age'], df['duration_sec'], alpha = 1/10)
plt.xlim([0, 100])
plt.ylim([500, 10000])
plt.xlabel('Age')
plt.yscale('log')
plt.yticks([500, 1e3, 2e3, 5e3, 1e4], [500, '1k', '2k', '5k', '10k'])
plt.ylabel('Trip Duration')
plt.title('Trip Duration and User Age Correlation')

plt.show()
```

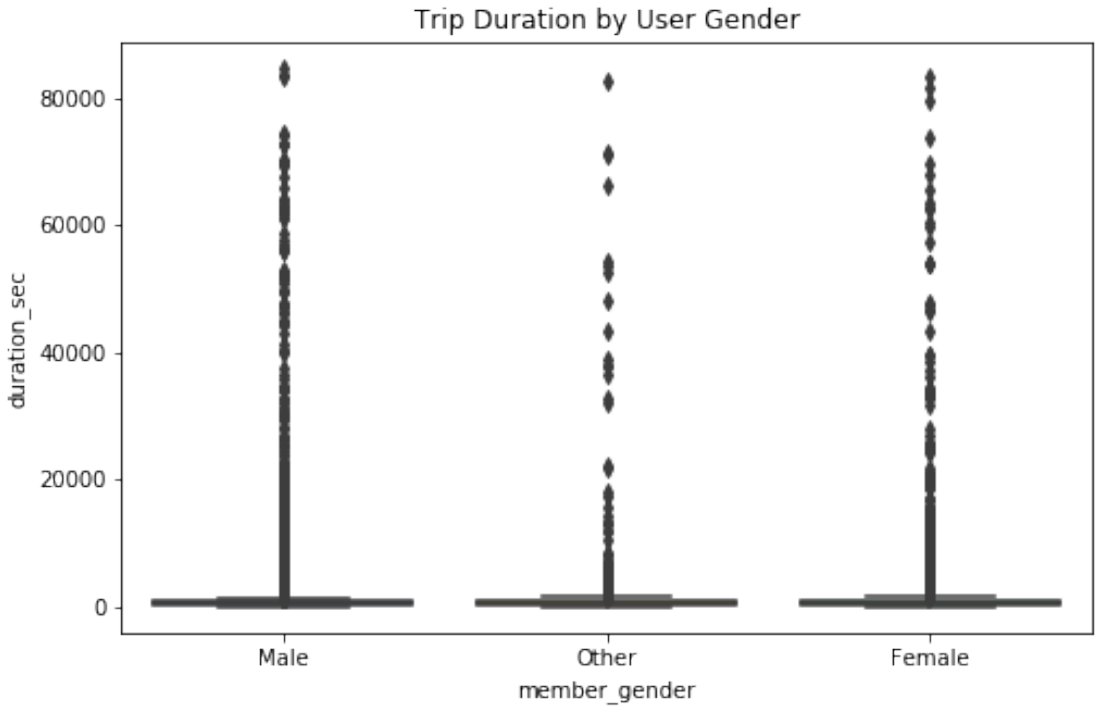


Most trip durations are below 2k and the age is below 65. We can conclude that most users who take longer trip durations are younger.

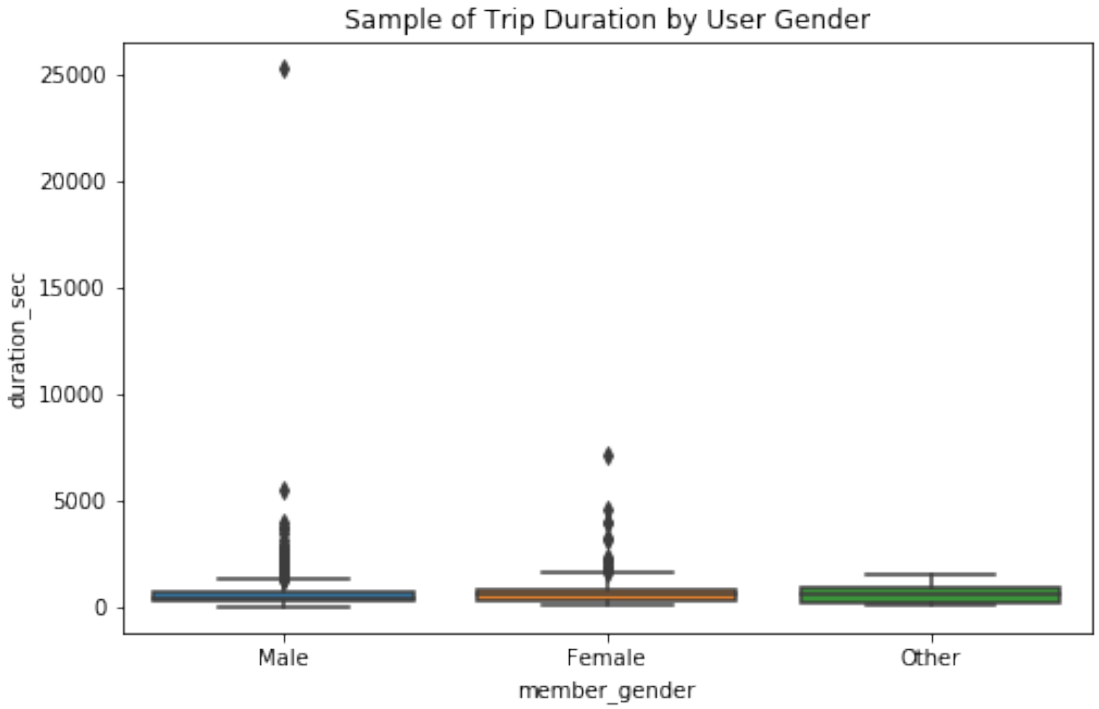
1.5.3 What is the average trip duration based on users' gender?

In [57]:
```
# Add user genders into variables

male = df['member_gender'] == 'Male'
female = df['member_gender'] == 'Female'
others = df['member_gender'] == 'Other'
```

In [58]:
```
plt.figure(figsize = [8, 5])
plt.title('Trip Duration by User Gender')
sb.boxplot(data = df, x = 'member_gender', y = 'duration_sec');
```

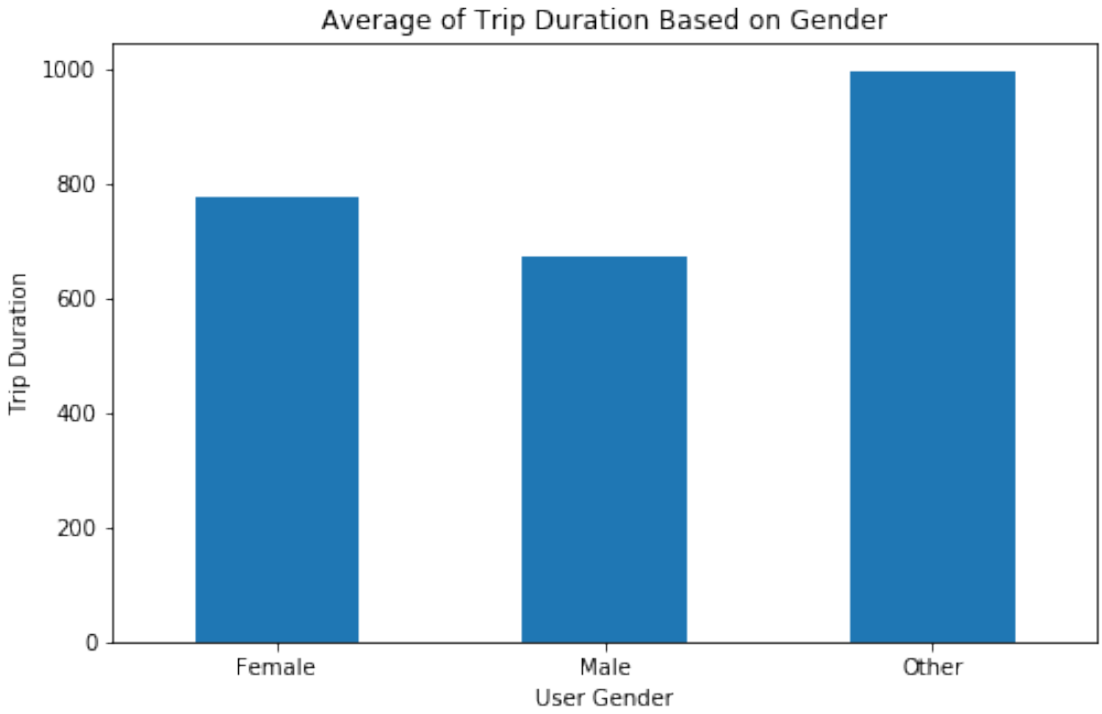


In [59]:
```
# Use a sample of data to make better visualization of the box plot

sample = df.sample(1000)

plt.figure(figsize = [8, 5])
plt.title('Sample of Trip Duration by User Gender')
sb.boxplot(data = sample, x = 'member_gender', y = 'duration_sec');
```




```
In [60]: # Average trip duration for Male

         df[male]['duration_sec'].mean()

Out[60]: 672.39138934247137

In [61]: # Average trip duration for Female

         df[female]['duration_sec'].mean()

Out[61]: 778.95001348336643

In [62]: # Average trip duration for Other

         df[others]['duration_sec'].mean()

Out[62]: 997.40263302249036

In [63]: # Plot trip duration comparison

         duration_gender = df[['member_gender', 'duration_sec']].groupby('member_gender').mean()
         duration_gender.plot(kind = 'bar', figsize = [8, 5], legend= None)
         plt.title("Average of Trip Duration Based on Gender")
         plt.ylabel("Trip Duration")
         plt.xlabel('User Gender')
```

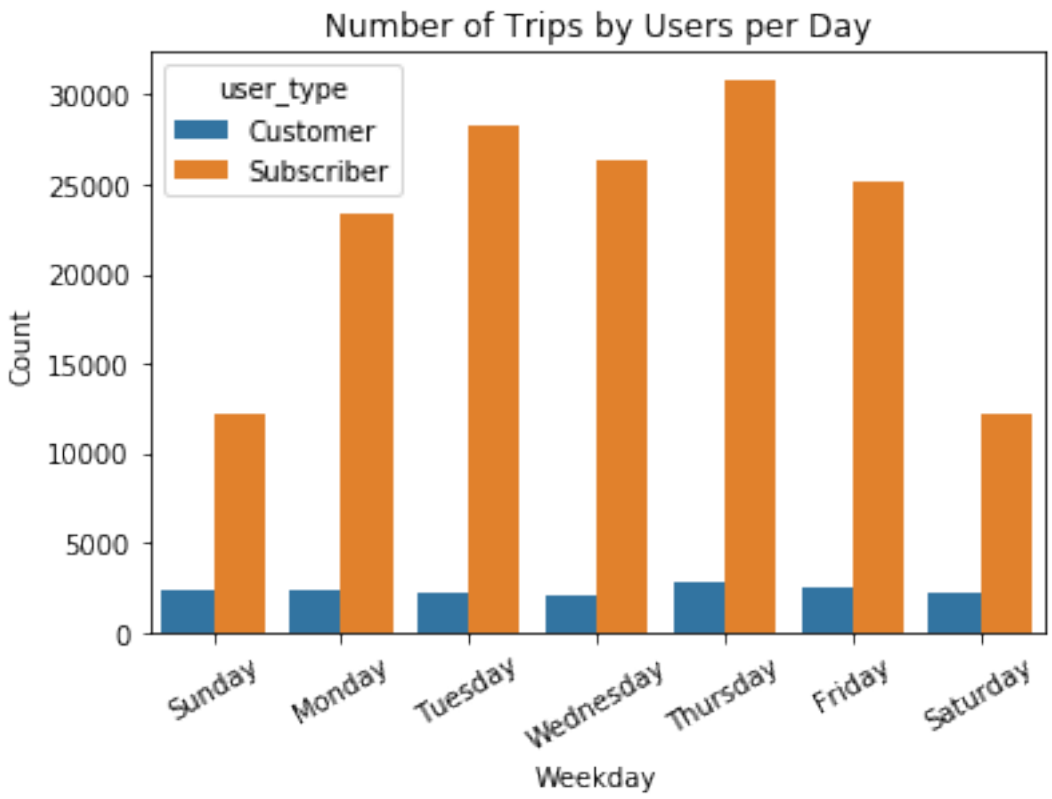
```
plt.xticks(rotation=0)

plt.show()
```



From the plots above, we can see that the trip duration of female users is more than male users with an average of 779 seconds, wheras male users have an average of 672 seconds of trip durations. However, `Other` tops male and female with the trip duration frequency having an average of 997 seconds.

1.5.4 What is the number of bike trips made by each user type per day?

In [64]:
```
# Plot clustered bar chart for comparison

plot = sb.countplot(data = df, x = 'start_day', hue = 'user_type',
            order = ['Sunday', 'Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday',
plot.set_title('Number of Trips by Users per Day')
plot.set_xlabel('Weekday')
plot.set_ylabel('Count')
plt.xticks(rotation=30)

plt.show()
```

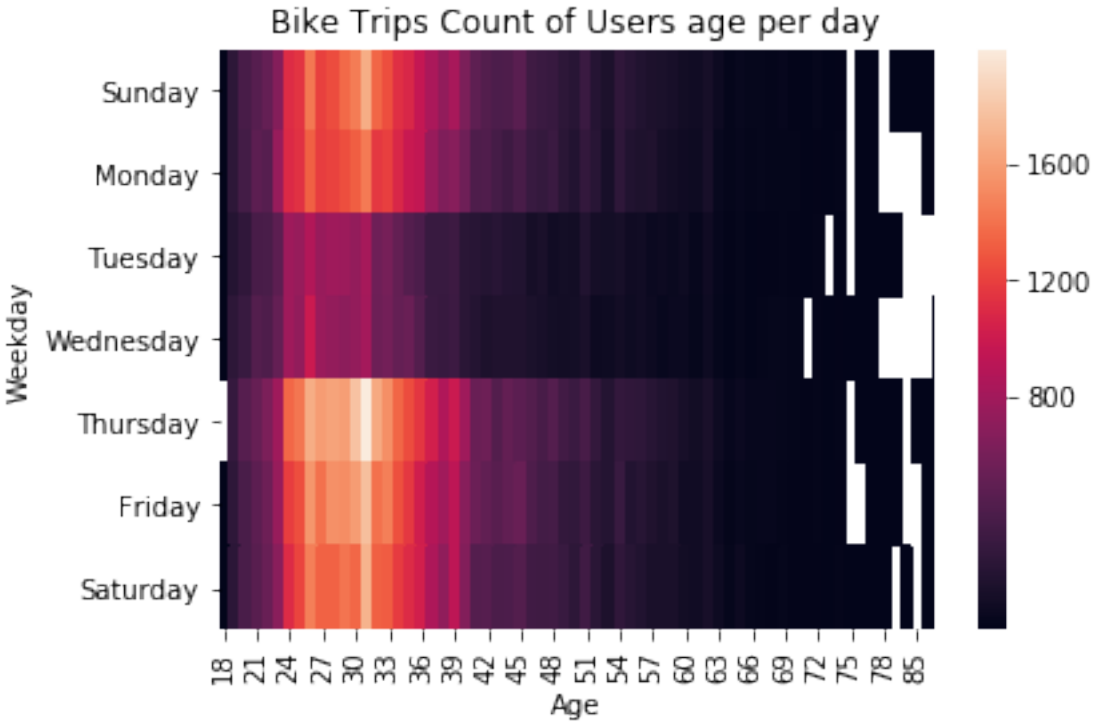


From the clustered bar above, we can see the number of trips taken by subscribers is dramatically more than the trips taken by customers. For subscribers, we can see that Thursday has the highest number of trips, followed by Tuesday then Wednesday. As for customers, we can see that the number of trips in all weekdays are very similar. However, we can see Thursday having a slight count rise than the rest of the days, followed by Friday.

1.5.5 What is the number of bike trips made by each user age per day?

```
In [65]: ct_counts = df.groupby(['start_day', 'age']).size()
         ct_counts = ct_counts.reset_index(name='count')
         ct_counts = ct_counts.pivot(index = 'start_day', columns = 'age', values = 'count')

         # Create heatmap
         heatmap = sb.heatmap(ct_counts,
                              yticklabels = ['Sunday', 'Monday', 'Tuesday', 'Wednesday', 'Thursd

         heatmap.set_xlim([0, 65])

         # set the colorbar ticks
         cbar = heatmap.collections[0].colorbar
         cbar.set_ticks([800, 1200, 1600])
```

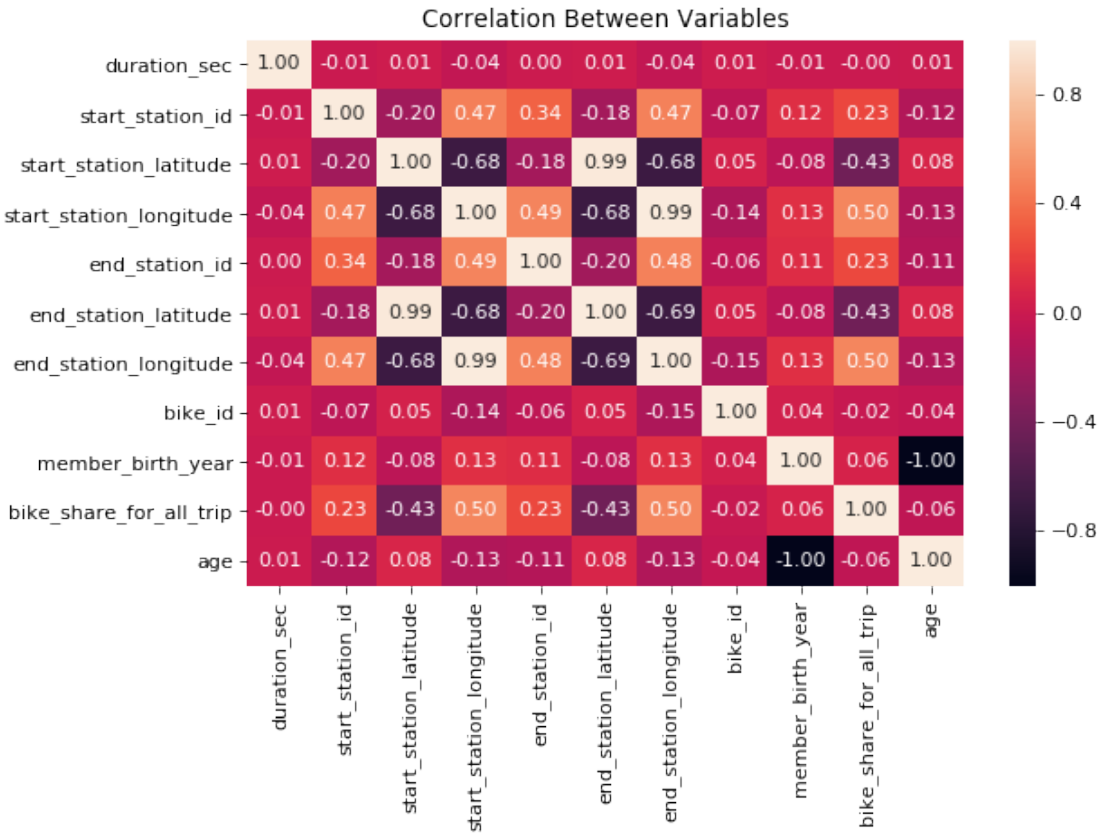
```
plt.xlabel('Age')
plt.ylabel('Weekday')
plt.title('Bike Trips Count of Users age per day')

plt.show()
```



From the heatmap, we can see that the majority of users aged between 24 and 36 frequently ride their bikes on Thursday, Friday, and Saturday. Also, the least frequent days for their bike rides are Tuesday and Wednesday. We can also see that users aged older than 42 don't drive their bikes much since the number of bike ride count is dropping.

1.5.6 What is the correlation between each feature in the data?

```
In [66]: fig, ax = plt.subplots(figsize = [8, 5], dpi = 80)
         correlation = df.corr()
         sb.heatmap(correlation, annot = True, fmt = '.2f')
         ax.set_title("Correlation Between Variables")
         plt.xticks(rotation=90)

         plt.show()
```



From the heatmap above, we can observe the following: - There are almost no correlation between the trip duration and the start and end station ids, in addition to their longtitude and latitude. - There is a weak correlation between the start station id and the start and end station longtitudes. Thus, we can say that the station location might be a good factor to further study in the future. - There is a strong correlation between the start station latitude and the end station latitude. The same goes for the longtitude. - There is a negative correlation between the birth year and the age.

1.5.7 Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

We observed that the trip duration depends on two important factors which are: **user type** and **user age**. The main reason is that the trip duration is significantly higher for customers in comparison with subscribers. Also, the age of users affects the duration of trips since users between 20 and 40 are the majority of bikers. We also observed that the subscribers have a higher bike ride count regardless of customers taking longer trip durations.

1.5.8 Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

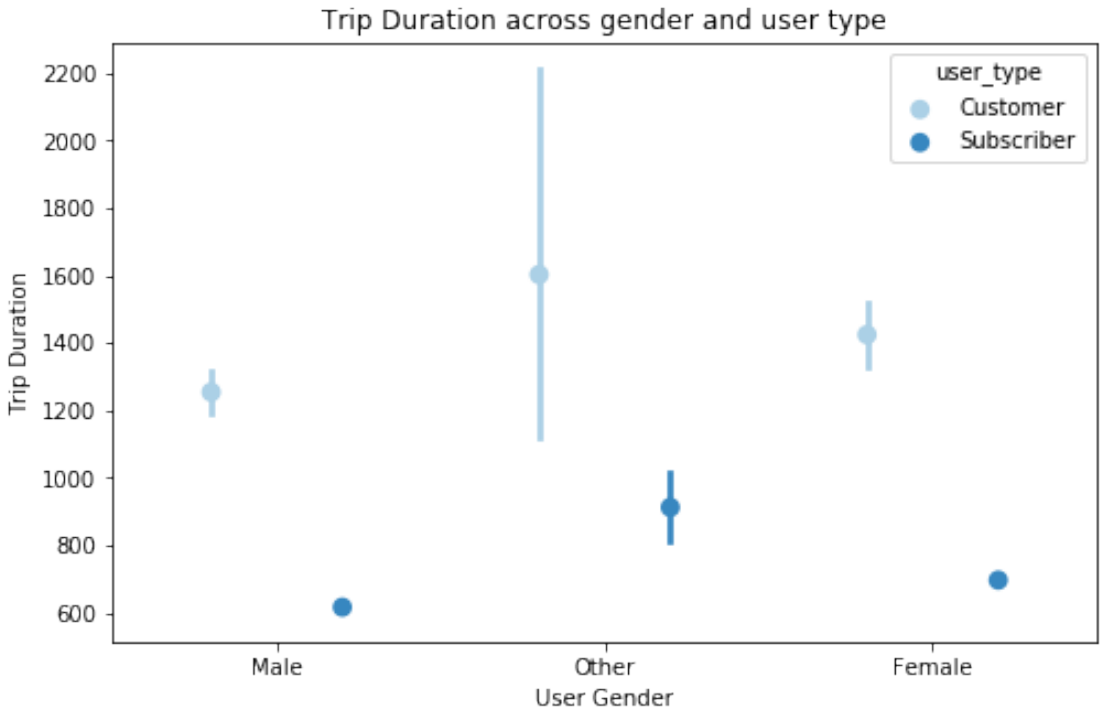
The gender of users and its dependance on the trip duration is very interesting. We observed that females have a higher trip duration than males. We also observed that Thursday is the most popular day for bike rides among subscriberes and customers. In addition, we observed that the location of the start station and end station have a strong correlation, meaning that they are important factors in the dataset. Also, having no correlation between trip duration and the location of the stations is surprising.

1.6 Multivariate Exploration

This section investigates the relation between three or more variables in the data. This section would help us in investigating the data further, in addition to concluding insights from previous sections.

1.6.1 What is the trip duration across gender and user type?

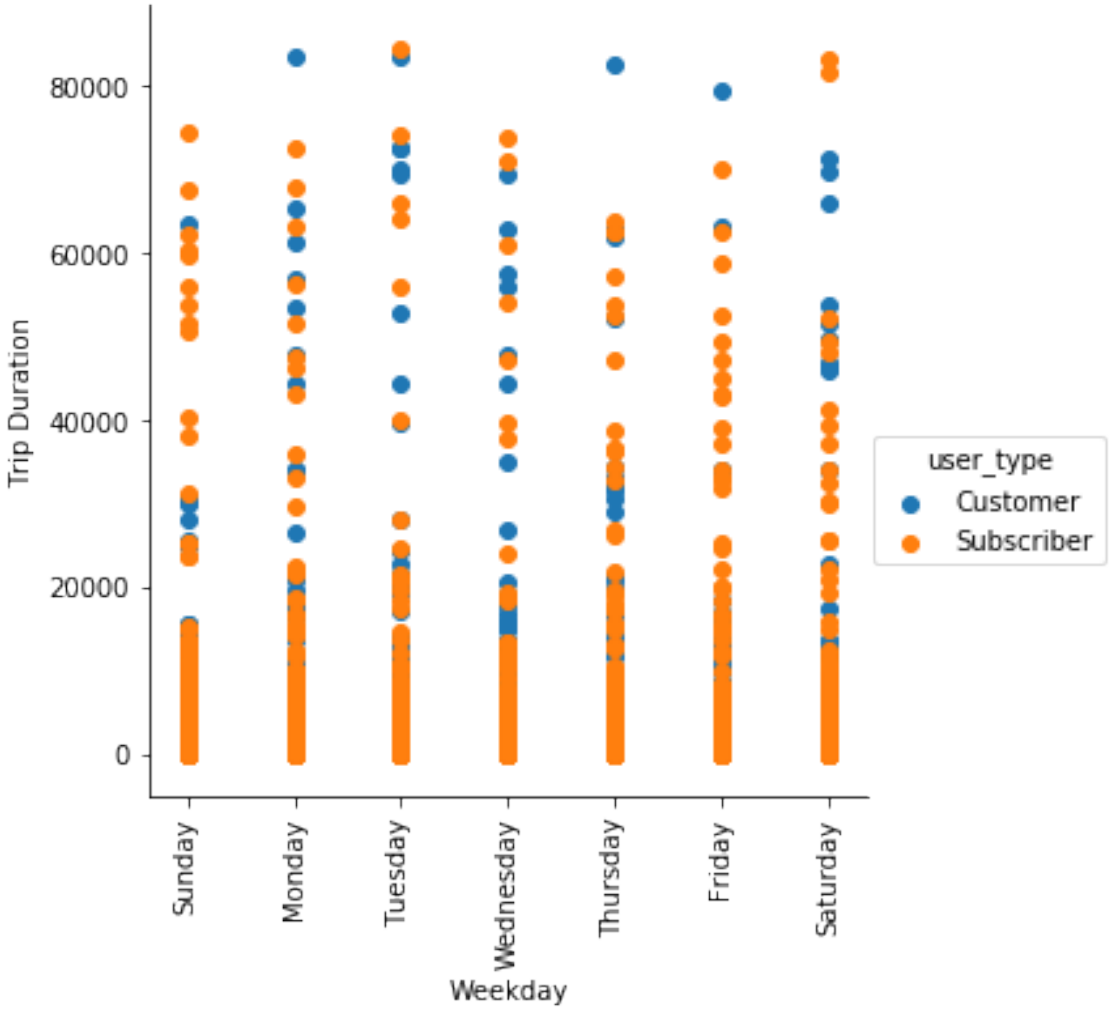
In [67]:
```
fig = plt.figure(figsize = [8,5])
ax = sb.pointplot(data = df, x = 'member_gender', y = 'duration_sec', hue = 'user_type'
                  palette = 'Blues', linestyles = '', dodge = 0.4)

plt.title('Trip Duration across gender and user type')
plt.ylabel('Trip Duration')
plt.xlabel('User Gender')

plt.show()
```



From the point plot above, we can say that **Other** gender and **Customer** user type spent the longest trip duration. For the male and female, we can see that **female customers** had longer trip durations than **male customers**. Overall, customers have longer trip durations than subscribers on a significant level.

1.6.2 What is the bike trip duration of user types per day?

```
In [68]: g = sb.FacetGrid(data = df, hue = 'user_type', size = 5)
         g.map(plt.scatter, 'start_day', 'duration_sec')
         g.add_legend()

         # Set the order of the x-axis
         day_order = ['Sunday', 'Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday', 'Saturda
         g.set_xticklabels(day_order, rotation=90)
         g.set_xlabels('Weekday')
         g.set_ylabels('Trip Duration')
         g.set_titles('Trip Duration for User Types per Day')

         plt.show()
```



From the plot above, we can see that the majority of both customers and subscribers trip durations are clustered below 20,000 seconds i.e. 333.33 minutes (5.56 hours). Surprisingly, we can see that we have outliers for both customers and subscribers where the trip duration can reach up to 80,000 seconds (1333.3333 hours), for all weekdays. In addition, we can see that customers have longer trip durations on Monday, Wednesday, and Saturday. Whereas, subscribers have longer trip durations on Sunday, Thursday and Friday.

1.6.3 Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?

We observed that the `other` gender are the majority of customers and the ones who take the longest trip durations. We also observed that the majority of subscribers and customers take around 5.56 hours of trip durations or less on all weekdays.

1.6.4 Were there any interesting or surprising interactions between features?

Customers who have `other` gender take the longest trip durations. Surprisingly, there exists some outliers for customers and subscribers where the trip duration can be as long as 80,000 seconds or more (1333.3333 hours).

1.7 Conclusions

After exploring many factors, we conclude the following: - The average of trip durations is around 600 seconds i.e. 10 minutes. - Even though the majority of users are subscribers, customers have longer trip durations that subscribers. - Most users are aged between 20 and 40 with the average being 33 years old. - While male users are the majority of users in the system, female users and others have longer trip durations than male users. - Most users aged between 20 and 40 have trip duartions below 2000 seconds i.e. 33 minutes. - The highest day of the week for bike rides is Thursday, and the lowest is the weekend (Saturday and Sunday). - There is a strong relation between the location of the start and end stations. - There is no clear relation between the trip duration and the location of stations. - The majority of subscribers and customers take around 5.56 hours of trip durations or less on all weekdays. - The most popular day for bike rides for subscribers is Thursday and customers have longer trip durations on Saturday.

In [69]:
```
# Export clean dataset to csv format

df.to_csv('df_clean.csv', index=None)
```