Ford-GoBike-System-Data-Exploration

February 17, 2022

1 Ford GoBike System Data Exploration

1.1 by Sultanah Aldossari

1.2 Introduction

About dataset: a dataset include information about bike trips on February and March 2019. The dataset include

- `duration_sec`: Trip duration in seconds
- `start_time`: Trip start time and date
- `end_time`: Trip end time and date
- `start_station_id`: Trip start station id
- `start_station_name`: Station name
- `start_station_latitude`: Start station latitude
- `start_station_longitude` Start station longitude
- `end_station_id`: Trip end station ID
- `end_station_name`: Trip end station name
- `end_station_latitude`: End Station Latitude
- `end_station_longitude`: End Station Longitude
- `bike_id`: Bike ID
- `user_type`: User type whether a subscriber or a customer -- ("Subscriber" = Member or "Customer" = Casual)
- `member_birth_year`: User birth year
- `member_gender`: User gender whether a female or male
- `bike_share_for_all_trip`

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
```

1.4 Data Gathering

```
In [71]: # Read data from a Csv file
         df = pd.read_csv('tripdata.csv')
         df.head()
```

```
Out[71]:    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
```

1.4.1 What is the structure of your dataset?

The dataset consist of 183,412 observations and 16 features. 9 of the features are numeric the rest are catagorical variable. As it appears there are some missing values

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

We can derive several valuable information from the dataset. For instance, we can answer these questions using the above dataset:

- Which days have the highest number of trips?
- Which hours have the highest number of trips?
- Who have the highest number of trips customer or subscriber?
- In what age trips have the highest peaks? Is there a Significant relation between age the bike riding?
- Does gender affect bike riding?
- What is the longest trip time? does poeple tend to use bikes for long or short time?
- Which days customer and subscribers uses bikes?
- Top 5 stations

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

start and end time, station name, gender, user type, birth date, and duration in seconds

Derived features/variables to assist exploration and analysis: start_date, start_hourofday, start_dayofweek, member_age

2 Assess Data

```
In [3]: df.shape
```

```
Out[3]: (183412, 16)
```

```
In [4]: 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]: df.isna().sum()
```

```
Out[5]: 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
        dtype: int64
```

```
In [6]: df.duplicated().sum()
```

```
Out[6]: 0
```

```
In [7]: df.describe()
```

```
Out[7]:         duration_sec  start_station_id  start_station_latitude  \
        count  183412.000000     183215.000000           183412.000000
        mean      726.078435        138.590427               37.771223
        std      1794.389780        111.778864                0.099581
```

```
min        61.000000         3.000000               37.317298
25%       325.000000        47.000000               37.770083
50%       514.000000       104.000000               37.780760
75%       796.000000       239.000000               37.797280
max     85444.000000       398.000000               37.880222

       start_station_longitude  end_station_id  end_station_latitude  \
count            183412.000000   183215.000000         183412.000000
mean               -122.352664      136.249123             37.771427
std                   0.117097      111.515131              0.099490
min                -122.453704        3.000000             37.317298
25%                -122.412408       44.000000             37.770407
50%                -122.398285      100.000000             37.781010
75%                -122.286533      235.000000             37.797320
max                -121.874119      398.000000             37.880222

       end_station_longitude        bike_id  member_birth_year
count          183412.000000  183412.000000      175147.000000
mean             -122.352250    4472.906375        1984.806437
std                 0.116673    1664.383394          10.116689
min              -122.453704      11.000000        1878.000000
25%              -122.411726    3777.000000        1980.000000
50%              -122.398279    4958.000000        1987.000000
75%              -122.288045    5502.000000        1992.000000
max              -121.874119    6645.000000        2001.000000
```

In [8]:
```
df.nunique()
```

Out[8]:
```
duration_sec                 4752
start_time                 183401
end_time                   183397
start_station_id              329
start_station_name            329
start_station_latitude        334
start_station_longitude       335
end_station_id                329
end_station_name              329
end_station_latitude          335
end_station_longitude         335
bike_id                      4646
user_type                       2
member_birth_year              75
member_gender                   3
bike_share_for_all_trip         2
dtype: int64
```

In [9]:
```
df.member_gender.value_counts()
```

Out[9]:
```
Male      130651
Female     40844
```

```
Other        3652
Name: member_gender, dtype: int64
```

```
In [10]: df.user_type.value_counts()
```

```
Out[10]: Subscriber    163544
         Customer       19868
         Name: user_type, dtype: int64
```

```
In [11]: df.bike_share_for_all_trip.value_counts()
```

```
Out[11]: No     166053
         Yes     17359
         Name: bike_share_for_all_trip, dtype: int64
```

```
In [12]: df.start_station_name.value_counts()
```

```
Out[12]: Market St at 10th St                                           3904
         San Francisco Caltrain Station 2  (Townsend St at 4th St)      3544
         Berry St at 4th St                                             3052
         Montgomery St BART Station (Market St at 2nd St)               2895
         Powell St BART Station (Market St at 4th St)                   2760
         San Francisco Ferry Building (Harry Bridges Plaza)             2710
         San Francisco Caltrain (Townsend St at 4th St)                 2703
         Powell St BART Station (Market St at 5th St)                   2327
         Howard St at Beale St                                          2293
         Steuart St at Market St                                        2283
         The Embarcadero at Sansome St                                  2082
         Bancroft Way at Telegraph Ave                                  1796
         Bancroft Way at College Ave                                    1770
         2nd St at Townsend St                                          1765
         3rd St at Townsend St                                          1753
         Embarcadero BART Station (Beale St at Market St)               1746
         Beale St at Harrison St                                        1719
         Civic Center/UN Plaza BART Station (Market St at McAllister St)    1611
         Townsend St at 7th St                                          1573
         4th St at Mission Bay Blvd S                                   1552
         The Embarcadero at Steuart St                                  1458
         Post St at Kearny St                                           1376
         Downtown Berkeley BART                                         1375
         4th St at 16th St                                              1360
         Howard St at 8th St                                            1314
         Rhode Island St at 17th St                                     1303
         Esprit Park                                                    1290
         19th Street BART Station                                       1276
         8th St at Brannan St                                           1203
         Hearst Ave at Euclid Ave                                       1203
                                                                        ...
         10th Ave at E 15th St                                            57
```

```
45th St at MLK Jr Way                                        55
27th St at MLK Jr Way                                        55
San Antonio Park                                             53
Williams Ave at 3rd St                                       50
Delmas Ave and San Fernando St                               50
Locust St at Grant St                                        49
San Carlos St at Market St                                   48
Lane St at Revere Ave                                        36
Foothill Blvd at Harrington Ave                              35
Almaden Blvd at Balbach St                                   34
Mission St at 1st St                                         34
SAP Center                                                   32
George St at 1st St                                          31
Oak St at 1st St                                             30
Empire St at 7th St                                          29
Williams Ave at Apollo St                                    25
Foothill Blvd at 42nd Ave                                    23
San Pedro St at Hedding St                                   19
26th Ave at International Blvd                               19
23rd Ave at Foothill Blvd                                    18
Farnam St at Fruitvale Ave                                   18
Leavenworth St at Broadway                                   17
Backesto Park (Jackson St at 13th St)                        17
Taylor St at 9th St                                          13
Willow St at Vine St                                          9
Parker Ave at McAllister St                                   7
21st Ave at International Blvd                                4
Palm St at Willow St                                          4
16th St Depot                                                 2
Name: start_station_name, Length: 329, dtype: int64
```

In [13]:
```
df.start_station_name.nunique()
```

Out[13]:
```
329
```

In [14]:
```
df.member_birth_year.value_counts().sort_values()
```

Out[14]:
```
1878.0        1
1930.0        1
1928.0        1
1927.0        1
1910.0        1
1944.0        2
1934.0        2
1920.0        3
1938.0        3
1901.0        6
1941.0        9
1939.0       11
```

1902.0	11
1946.0	19
1933.0	20
1942.0	21
1943.0	30
2001.0	34
1948.0	51
1900.0	53
1931.0	89
1949.0	99
1945.0	105
1955.0	134
1947.0	135
1953.0	158
1950.0	178
1951.0	180
1952.0	189
1954.0	301
	...
1972.0	1909
1971.0	1924
1968.0	1928
1973.0	2080
1976.0	2442
1975.0	2503
1999.0	2528
1974.0	2633
1977.0	2725
1978.0	2830
1998.0	3208
1997.0	3481
1979.0	3756
1981.0	4345
1996.0	4640
1982.0	4990
1980.0	5024
1983.0	5954
1984.0	6562
1985.0	7028
1995.0	7423
1994.0	7660
1986.0	7973
1987.0	8018
1992.0	8250
1991.0	8498
1990.0	8658
1989.0	8972
1993.0	9325

```
1988.0    10236
Name: member_birth_year, Length: 75, dtype: int64
```

3 Cleaning Data

define - Drop unwanted columns and columns with missing values - Change station name to string type - Change birth year from float to int type - Change gender type to string type - Change user_type to string type - Change duration(start, end) time to datetime format - Feature Engineering: days of week, months and hours

code

```
In [15]: from datetime import datetime
```

```
In [16]: df.drop(['start_station_latitude', 'start_station_longitude', 'start_station_id', 'end_
```

```
In [17]: df.drop(['end_station_latitude', 'end_station_longitude'], axis=1, inplace=True)
```

```
In [18]: df.dropna(inplace=True)
```

Changing Data Types

```
In [19]: df['start_time'] = pd.to_datetime(df['start_time'])
         df['end_time'] = pd.to_datetime(df['end_time'])
```

```
In [20]: df['user_type'] = df['user_type'].astype('category')
```

```
In [21]: df['member_gender'] = df['member_gender'].astype('category')
```

```
In [24]: df['start_station_name'] = df['start_station_name'].astype(str)
```

```
In [25]: df['end_station_name'] = df['end_station_name'].astype(str)
```

```
In [26]: df['dayofweek'] = df['dayofweek'].astype(str)
```

```
In [27]: df['member_birth_year'] = df['member_birth_year'].astype(int)
```

```
In [28]: df['hourofday'] = df['hourofday'].astype(int)
```

Feature Engineering

```
In [22]: df['start_date'] = df.start_time.dt.strftime('%Y-%m-%d')
         df['hourofday'] = df.start_time.dt.strftime('%H')
         df['dayofweek'] = df.start_time.dt.strftime('%A')
         df['month'] = df.start_time.dt.strftime('%B')
         df['duration_minute'] = df['duration_sec']/60
         df['member_age'] = 2019 - df['member_birth_year']
```

test

```
In [29]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 174952 entries, 0 to 183411
Data columns (total 16 columns):
duration_sec              174952 non-null int64
start_time                174952 non-null datetime64[ns]
end_time                  174952 non-null datetime64[ns]
start_station_name        174952 non-null object
end_station_name          174952 non-null object
bike_id                   174952 non-null int64
user_type                 174952 non-null category
member_birth_year         174952 non-null int64
member_gender             174952 non-null category
bike_share_for_all_trip   174952 non-null object
start_date                174952 non-null object
hourofday                 174952 non-null int64
dayofweek                 174952 non-null object
month                     174952 non-null object
duration_minute           174952 non-null float64
member_age                174952 non-null float64
dtypes: category(2), datetime64[ns](2), float64(2), int64(4), object(6)
memory usage: 20.4+ MB
```

In [30]: df.head()

```
Out[30]:    duration_sec              start_time                end_time  \
         0         52185 2019-02-28 17:32:10.145 2019-03-01 08:01:55.975
         2         61854 2019-02-28 12:13:13.218 2019-03-01 05:24:08.146
         3         36490 2019-02-28 17:54:26.010 2019-03-01 04:02:36.842
         4          1585 2019-02-28 23:54:18.549 2019-03-01 00:20:44.074
         5          1793 2019-02-28 23:49:58.632 2019-03-01 00:19:51.760

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

                                     end_station_name  bike_id   user_type  \
         0                 Commercial St at Montgomery St     4902    Customer
         2  Powell St BART Station (Market St at 4th St)     5905    Customer
         3                            Central Ave at Fell St     6638  Subscriber
         4                             10th Ave at E 15th St     4898  Subscriber
         5                                 Broadway at Kearny     5200  Subscriber

            member_birth_year member_gender bike_share_for_all_trip  start_date  \
         0               1984          Male                      No  2019-02-28
```

```
2               1972          Male                        No  2019-02-28
3               1989         Other                        No  2019-02-28
4               1974          Male                       Yes  2019-02-28
5               1959          Male                        No  2019-02-28

   hourofday dayofweek     month  duration_minute  member_age
0         17  Thursday  February       869.750000        35.0
2         12  Thursday  February      1030.900000        47.0
3         17  Thursday  February       608.166667        30.0
4         23  Thursday  February        26.416667        45.0
5         23  Thursday  February        29.883333        60.0
```

In [31]:
```
df.isna().sum()
```

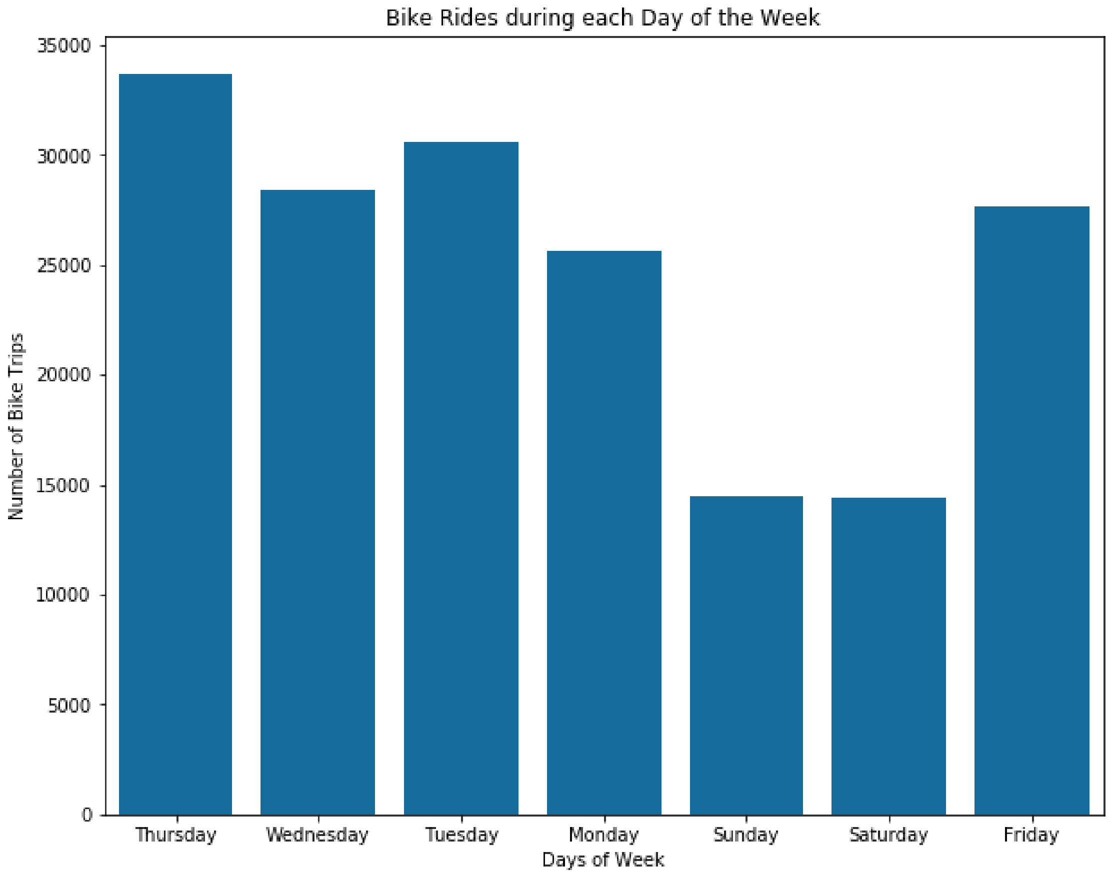
Out[31]:
```
duration_sec               0
start_time                 0
end_time                   0
start_station_name         0
end_station_name           0
bike_id                    0
user_type                  0
member_birth_year          0
member_gender              0
bike_share_for_all_trip    0
start_date                 0
hourofday                  0
dayofweek                  0
month                      0
duration_minute            0
member_age                 0
dtype: int64
```

3.1 Bivariate Exploration

3.2 Which days have the highest number of trips?

In [32]:
```
days_counts=df.dayofweek.value_counts()
```

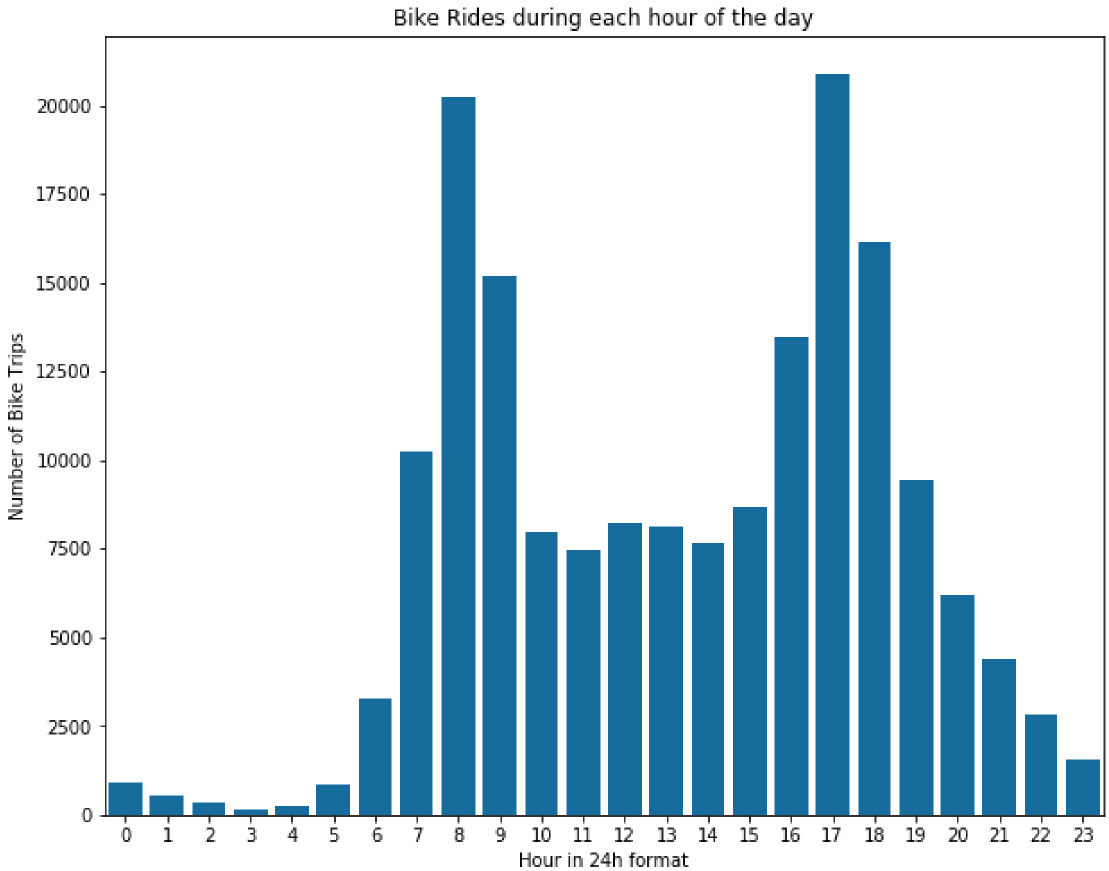
In [33]:
```
#Plot distribution of bikes trips on weekdays
import seaborn as sns
plt.figure(figsize=(10,8))
color = sb.color_palette('colorblind')[0]
sns.countplot(x=df['dayofweek'],  color=color)
plt.title("Bike Rides during each Day of the Week")
plt.xlabel("Days of Week")
plt.ylabel("Number of Bike Trips")
plt.show()
```



Insight We can see that during week days there are more bikes trips than on weekend, and usually because people on week days go to their work, shop, do some activities. And people tend to rest more on weekends.

3.3 Which hours have the highest number of trips?

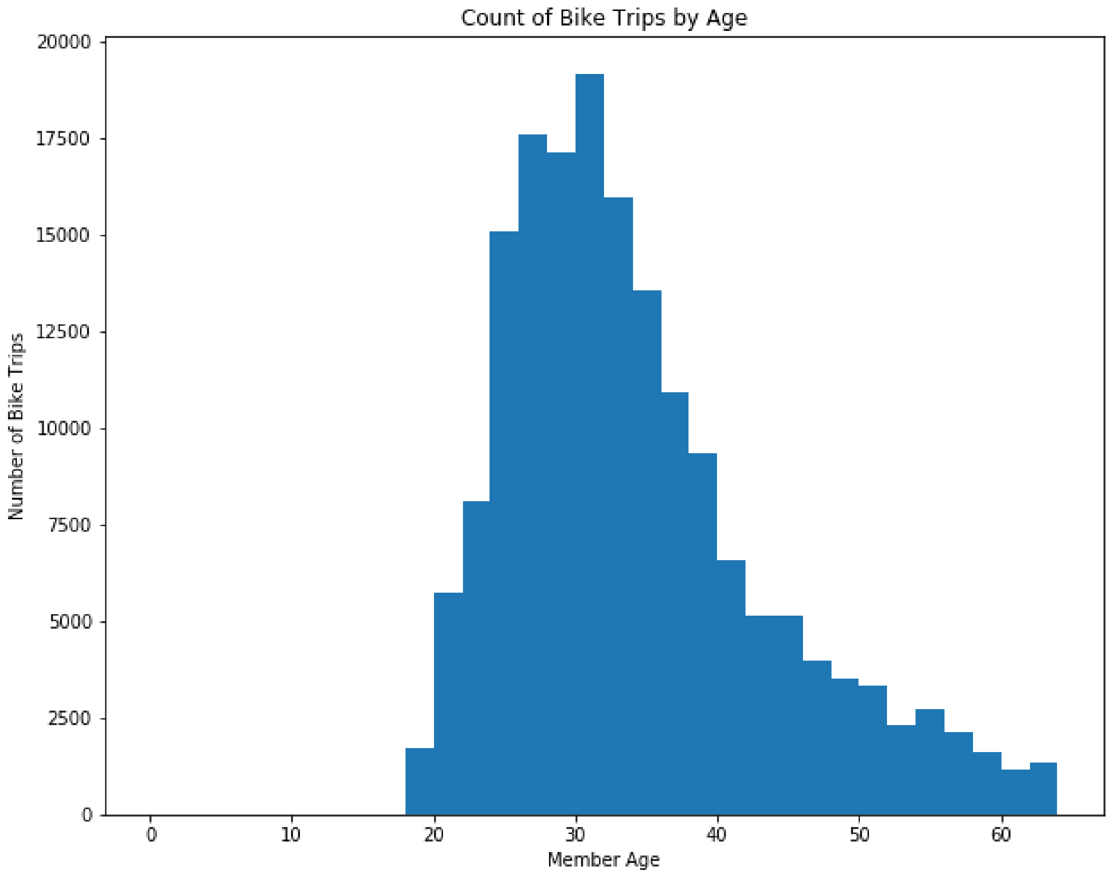
```
In [34]: plt.figure(figsize=(10,8))
         sns.countplot(data = df, x='hourofday', color = color)
         plt.title("Bike Rides during each hour of the day")
         plt.xlabel("Hour in 24h format")
         plt.ylabel("Number of Bike Trips")
         plt.show();
```



Insight In the plot above, it is obvious that in the morning and afternoon are the highest bike rides of the day.

3.4 In what age trips have the highest peaks? Is there a Significant relation between age the bike riding?

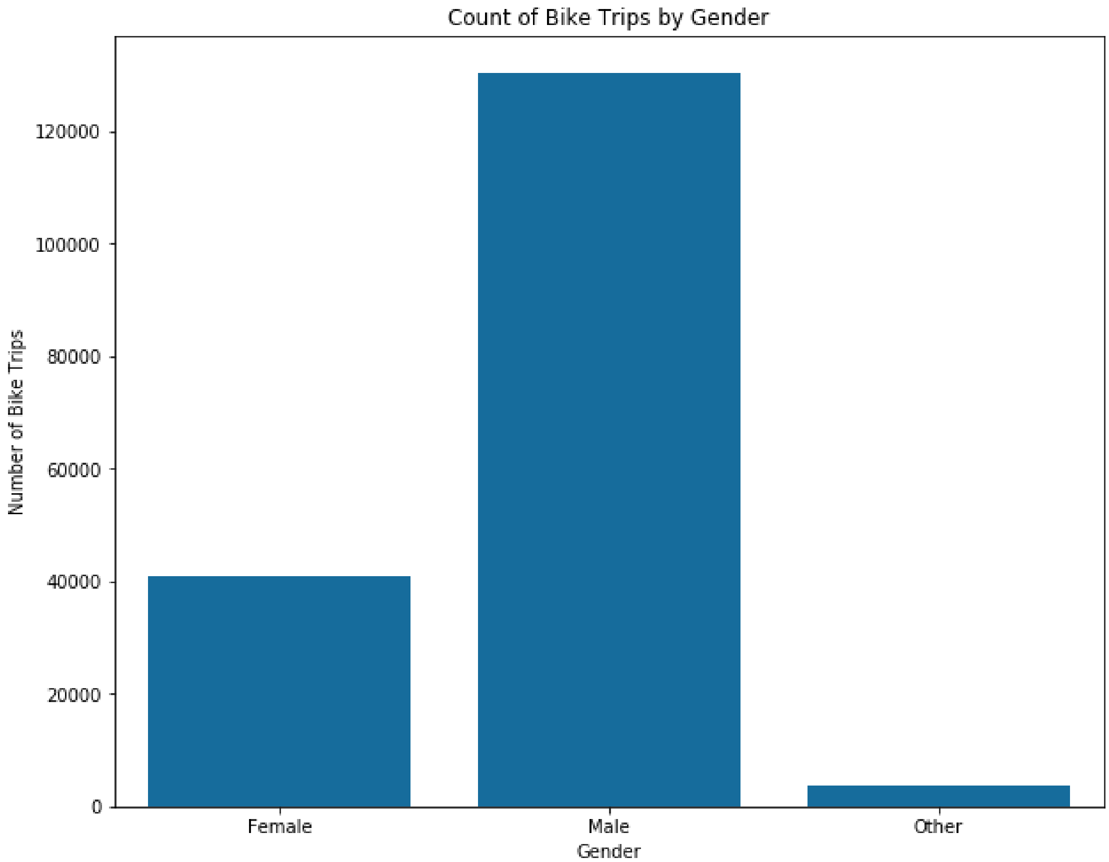
In [35]:
```
plt.figure(figsize=(10,8))
bins = np.arange(0, 65, 2)
plt.hist(df['member_age'], bins=bins)
plt.title('Count of Bike Trips by Age')
plt.ylabel('Number of Bike Trips')
plt.xlabel('Member Age');
```



Insight Bike Riders age peakes range in between 20 years to 35 years and then the usage of bikes starts decreasing.

3.5 Does gender affect bike riding?

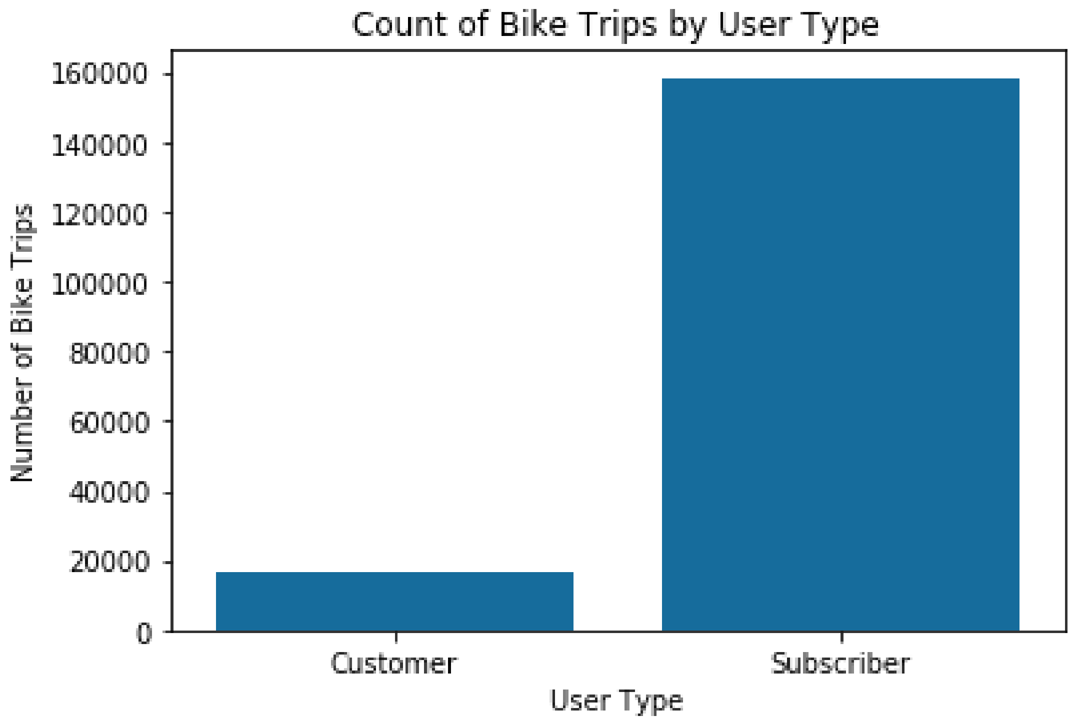
```
In [36]: plt.figure(figsize=(10,8))
         sns.countplot(data=df, x='member_gender',color=color)
         plt.title('Count of Bike Trips by Gender')
         plt.ylabel('Number of Bike Trips')
         plt.xlabel('Gender');
```



Insight Bike riders members are mainly Male members and few are females members, this certainly helps in marketing Campaigns as it simplify the targeted segment.

3.6 How many subscribers and customers rides bike?

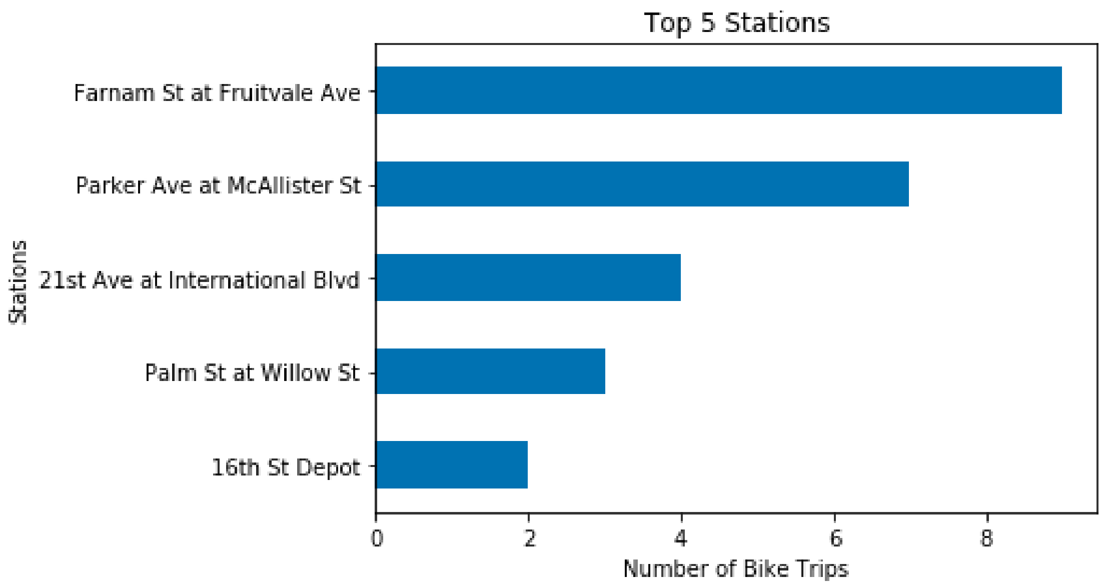
In [37]:
```
sns.countplot(data=df, x='user_type',color=color)
plt.title('Count of Bike Trips by User Type')
plt.ylabel('Number of Bike Trips')
plt.xlabel('User Type');
```



Insight Based on the above plot, subscribers use bikes more than regular customers.

3.7 Top 5 Start Station

```
In [38]: df.start_station_name.value_counts(sort=True, ascending=True)[:5].plot(kind='barh', col
         plt.title('Top 5 Stations')
         plt.xlabel('Number of Bike Trips')
         plt.ylabel('Stations');
```

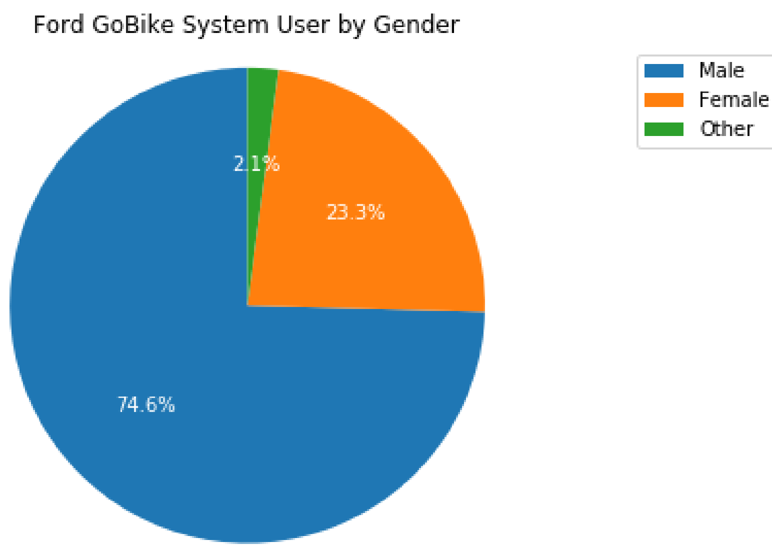


Insight Based on the above plot, Willow st Vine st have the highest traffic.

3.8 Gender Bike Trips

In [39]:
```
#Ref: https://matplotlib.org/stable/gallery/pie_and_polar_charts/pie_features.html

gender = df.member_gender.value_counts()
fig1, ax1 = plt.subplots(figsize=(10,5))
ax1.pie(gender, labels = gender.index, autopct='%1.1f%%', shadow=False, startangle=90,
ax1.axis('equal')
plt.legend(labels =gender.index, loc="best")
plt.title("Ford GoBike System User by Gender");
```



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

There were more trips on work days (Mon-Fri) than on the weekends, peaking around 8-9am and 17-18pm during the day. Males constituted a larger proportion of riders than females, and subscribers were more common than casual riders. Furthermore, Most of the members did not use the bike share for all of their trips, and most were between 25 and 40 years old. Because the data was straight-forward, no transformations were required.

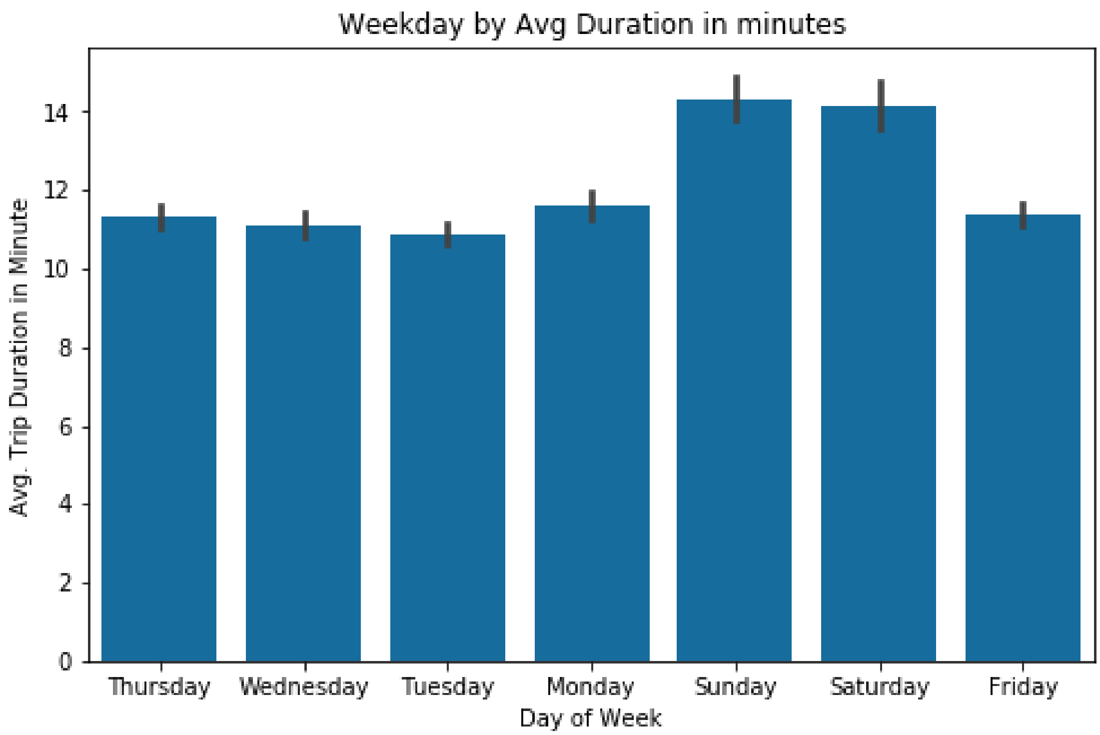
3.8.2 Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

There are no unusual expectations for a bike sharing system in a major city. As of now, the data indicates that adults in the average working age range are the primary users of the system, and they use the bikes daily for commuting.

3.9 Bivariate Exploration

3.9.1 Weekday by Avg Duration in minutes

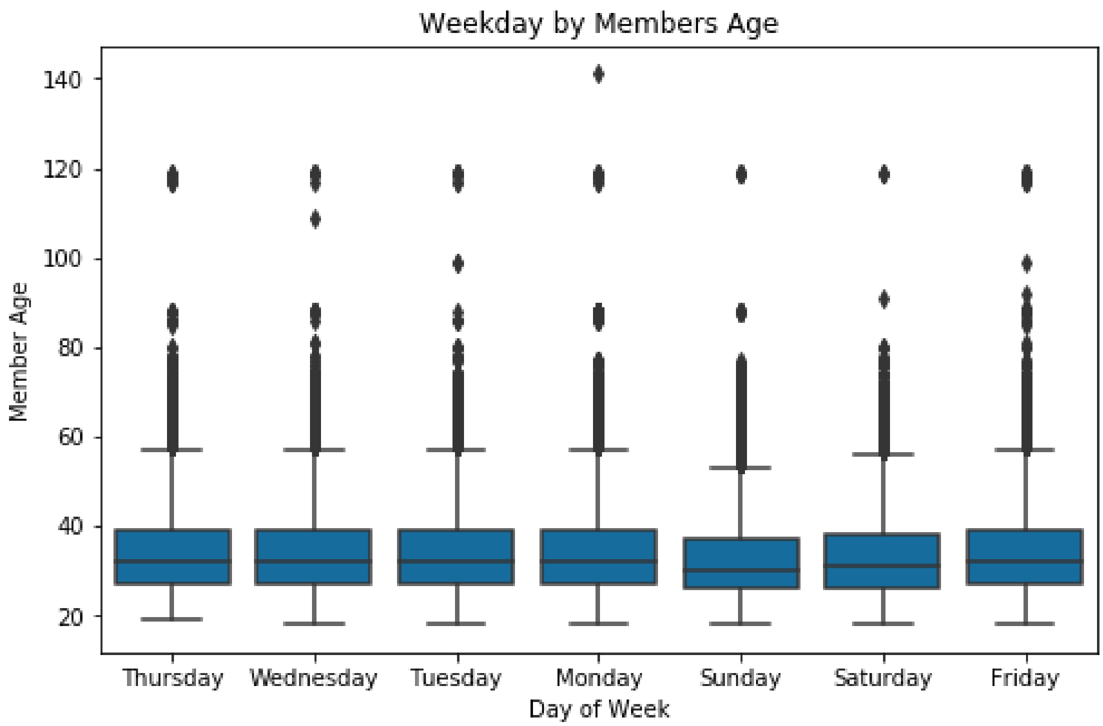
```
In [40]: plt.figure(figsize=(8,5))
         sns.barplot(data=df, x='dayofweek', y='duration_minute', color=color);
         plt.title('Weekday by Avg Duration in minutes');
         plt.xlabel('Day of Week');
         plt.ylabel('Avg. Trip Duration in Minute');
```



Insight: Comparatively to weekends, Monday through Friday riding trips are much shorter. According to that, the sharing system is used pretty reliably and efficiently on normal workdays, and more casually and flexible on weekends.

3.10 Weekday by Member's Age

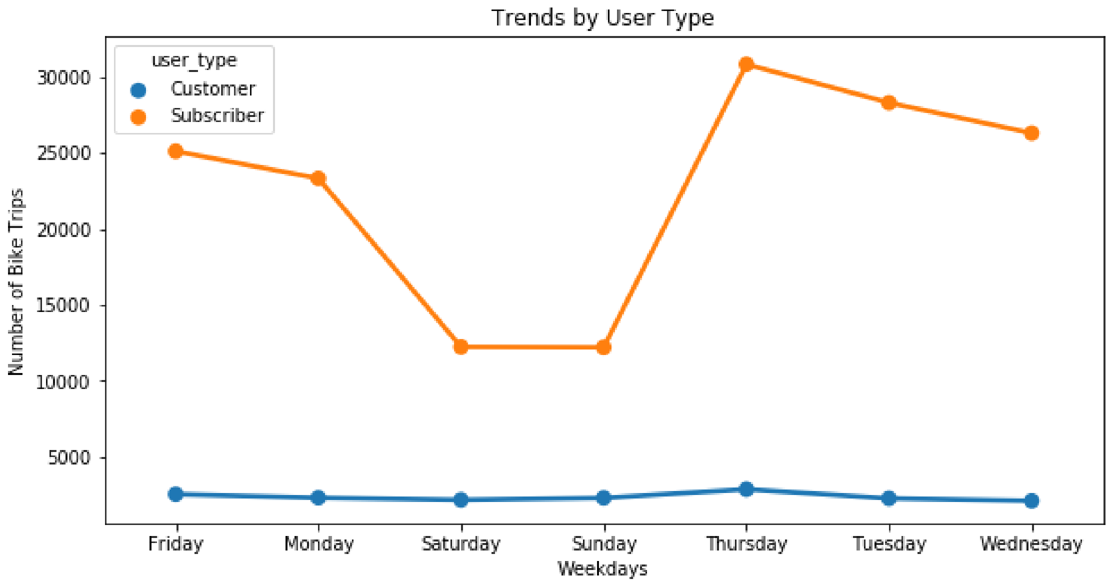
```
In [41]: plt.figure(figsize=(8,5))
         sns.boxplot(data=df, x='dayofweek', y='member_age', color=color)
         plt.title('Weekday by Members Age')
         plt.xlabel('Day of Week')
         plt.ylabel('Member Age');
```



Insight: There is a slight age difference between renters of bikes who ride from Monday through Friday and weekend renters, which corresponds to the commute to work patterns observed in the univariable exploration plots above.

3.11 User Type by Day Of Week

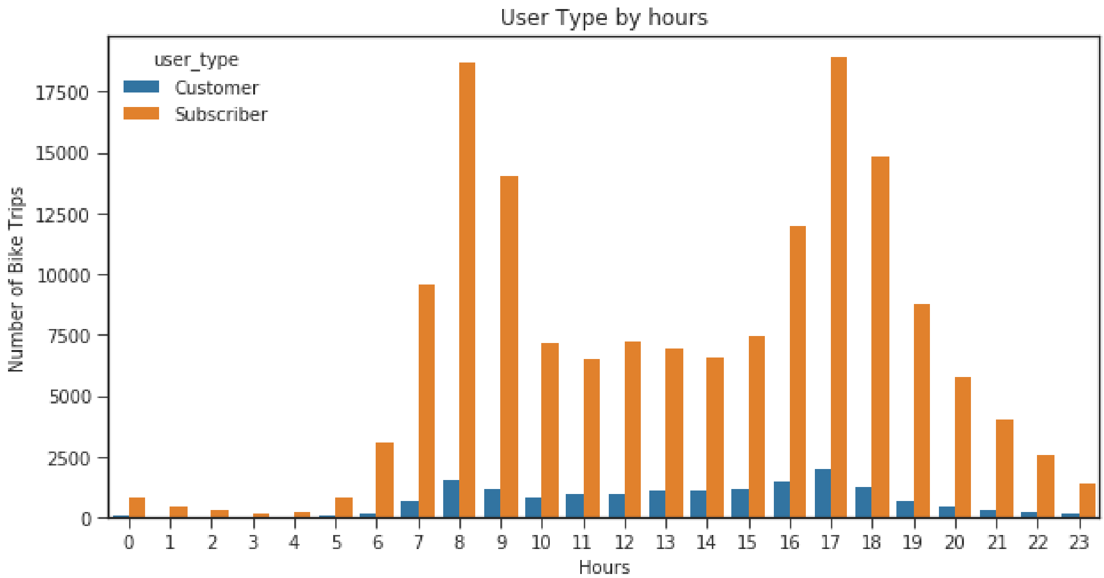
In [42]:
```
plt.figure(figsize=(10,5))
mask = df.groupby(['dayofweek', 'user_type']).size().reset_index()
ax = sns.pointplot(data=mask, x='dayofweek', y=0, hue = 'user_type')
plt.title('Trends by User Type')
plt.xlabel('Weekdays')
plt.ylabel('Number of Bike Trips');
```



Insights The above plot effectively illustrates the stark difference between Customers and Subscribers. In general, the bike share system is not very popular with customers; usage increases on weekends. The opposite is true for subscribers - on weekdays, usage has been high, but on weekends, usage has declined sharply.

3.12 User Type by hours

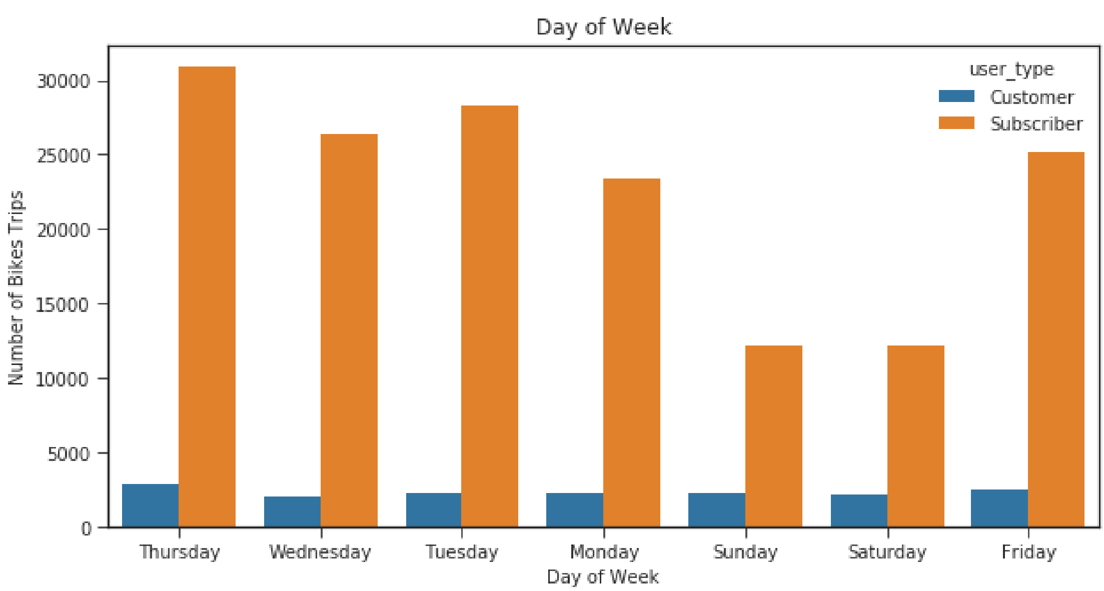
In [43]:
```
sns.set_style("ticks")
plt.figure(figsize=(10,5))
p = sns.countplot(data = df, x = 'hourofday', hue = 'user_type')
p.legend(loc = 2, framealpha = 0.2, title = 'user_type');
plt.title('User Type by hours')
plt.xlabel('Hours')
plt.ylabel('Number of Bike Trips');
```



Insights: The number of subscribers was higher than the number of casual customers. we can see from the plot clearly peaks out on typical rush hours when people go to work in the morning and getting off work in the afternoon

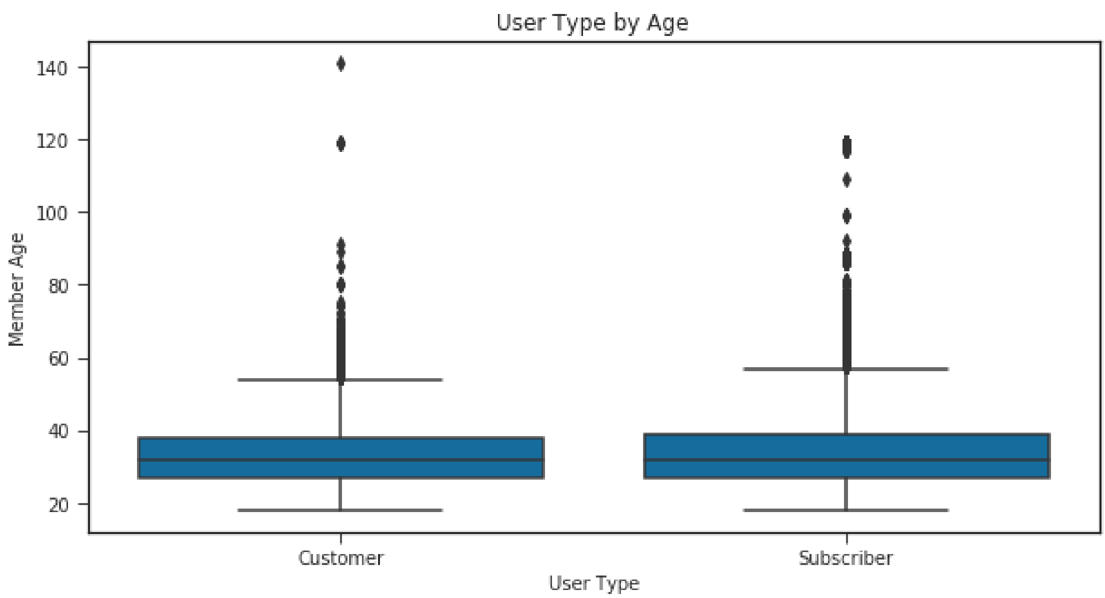
3.13 User Type by Day of week in bar graph

```
In [44]: plt.figure(figsize=(10,5))
         sns.countplot(data=df, x='dayofweek', hue='user_type')
         plt.title('Day of Week')
         plt.xlabel('Day of Week')
         plt.ylabel('Number of Bikes Trips');
```



Insights: The number of subscribers was higher than the number of casual customers. On weekends, there is a severe decline in volume for subscribers, which suggests that they use their bicycles primarily to commute to work during the week, whereas on weekends, there is a slight increase in volume for customers, which suggests that the use is primarily leisure/touring and relaxing.

```
In [66]: plt.figure(figsize=(10,5))
         sns.boxplot(data=df, x='user_type', y='member_age', color=color);
         plt.title('User Type by Age')
         plt.xlabel('User Type');
         plt.ylabel('Member Age');
```



Insights: Monday through Friday, subscribers tend to be older than customers, ranging in age from 17 to 60.

3.13.1 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?

By analyzing the data for the type of user, we discovered different behavior usage between customers and subscribers. Customers are more likely to be casual riders, like tourists or students on vacation. Subscribers, on the other hand, tend to be daily commuters and full-time students who mostly use the system during weekdays, in better weather, and mainly for shorter distances. They tend to rent bikes during the morning and evening of a typical work or school day (8-9am and 5-6pm).

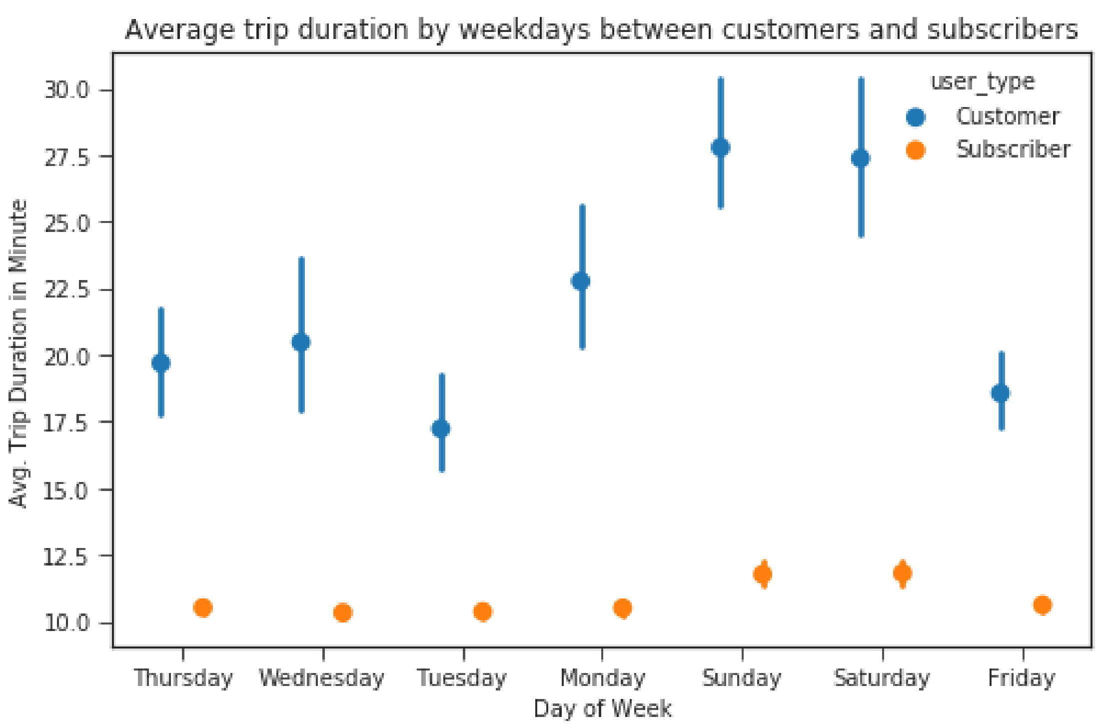
3.13.2 Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

It varies between subscribers and customers in the time it takes to use bikes. Subscribers during weekends use their bicycles largely to commute during the week, whereas on weekends, there is a slight increase in customers, which indicates that they use it primarily for leisure purposes.

3.14 Multivariate Exploration

3.14.1 Average trip duration by weekdays between customers and subscribers

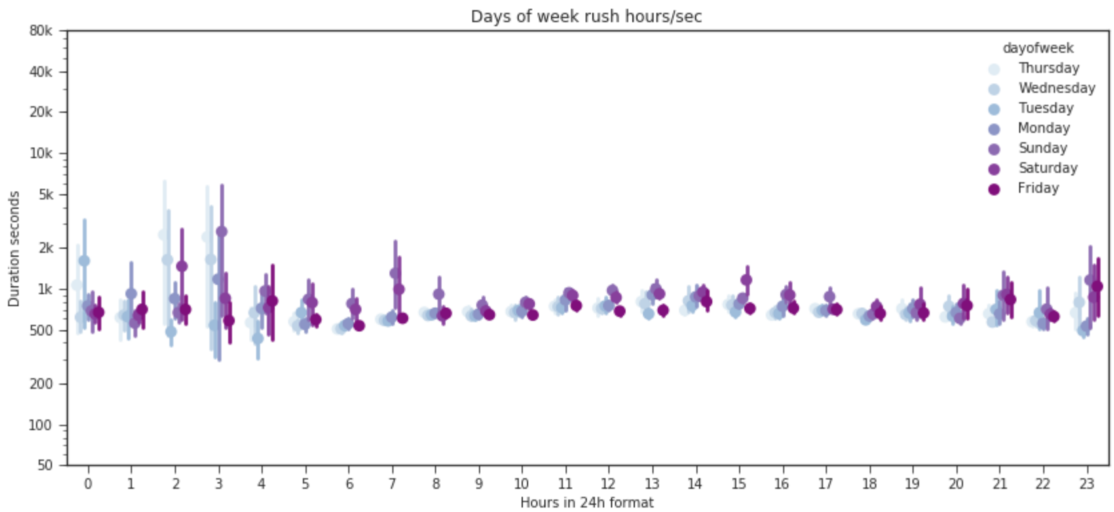
```
In [69]: plt.figure(figsize=[8,5])
         sns.pointplot(data=df, x='dayofweek', y='duration_minute', hue='user_type', dodge=0.3,
         plt.title('Average trip duration by weekdays between customers and subscribers');
         plt.xlabel('Day of Week');
         plt.ylabel('Avg. Trip Duration in Minute');
```



Insight: This plot shows that subscribers ride much shorter/quicker trips than customers on every day of the week. In particular, casual riders ride longer on Saturdays and Sundays than on other days of the week. The average duration of subscription usage seems to be more consistent between customers and subscribers.

```
In [70]: plt.figure(figsize=[14,6])
         sns.pointplot(data = df, x = 'hourofday', y = 'duration_sec', hue = 'dayofweek', dodge
```

```
plt.yscale('log')
plt.yticks([50,100,200,500, 1e3, 2e3, 5e3, 1e4, 2e4,4e4,8e4], [50,100,200,500, '1k', '2
plt.title('Days of week rush hours/sec')
plt.xlabel('Hours in 24h format')
plt.ylabel('Duration seconds');
```



Insight: most of weekdays have the most bikers than weekends. Thursday 3 AM has the most duration.

3.14.2 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?

In the multivariate exploration, several patterns were confirmed that had been discovered in the previous bivariate analysis as well as the univariate analysis. The relation between the multiple variables plotted is visually evident, and the data is presented in a combined form.According to subscribers, the majority of use occurs Monday through Friday during rush hours, indicating a primary use for work commutes. Based on the more relaxed and flexible use pattern of customers, it's clear that they might be using the system quite differently than subscribers, probably primarily over weekends and in the afternoon, for leisure purposes or city tours.

3.14.3 Were there any interesting or surprising interactions between features?

As a whole, the features complement each other and quite make sense when viewed in combination, it's not surprising at all. Throughout the exploration, there isn't a great deal of difference between male and female usage habits, which may be due to a greater number of male riders/records compared to female ones. If there were more data on females, it would be interesting to see how they use the system differently.

3.15 Conclusions

An impressive number of people will be able to benefit from this project: - The product is affordable and convenient transportation for anyone. - It provides for customers (students, tourists, etc.) with a flexible and sustainable way to tour the city. - Subscriptions (individuals who commute on a daily basis) benefit from the service in a convenient manner, according to the analysis - Using the Ford GoBike System is a great sustainable way to move around the city, both for leisure and for work. Users of the system can be either subscribers or customers. A majority of subscribers are daily commuters who have short trips to and from work. They rent bikes during weekdays at 8-9am and 5-6pm, and sometimes during lunch time. The system is mainly used by tourists and occasional riders on weekends to explore the Bay Area.

4 Sources

- https://docs.google.com/document/d/e/2PACX-1vQmkX4iOT6Rcrin42vslquX2_wQCjIa_hbwD0xmxrEI