Communicating Data Finding : Ford GoBike System Data

by Dain Russell

Investigation Overview

In this investigation, I wanted to look at the characteristics of the bikeshare dataset that could be used to explore trip duration.

Dataset Overview

The data consisted of trip start and end times and other variables.

- There were 4646 bikes.
- There are 183412 fordgobike trips in the cleaned dataset with 16 specifications or columns.

Connect to the dataset

In [41]:
```
# load in dataset into pandas dataframe
df = pd.read_csv('201902-fordgobike-tripdata.csv')
# display first 5 rows dataframe
df.head()
```

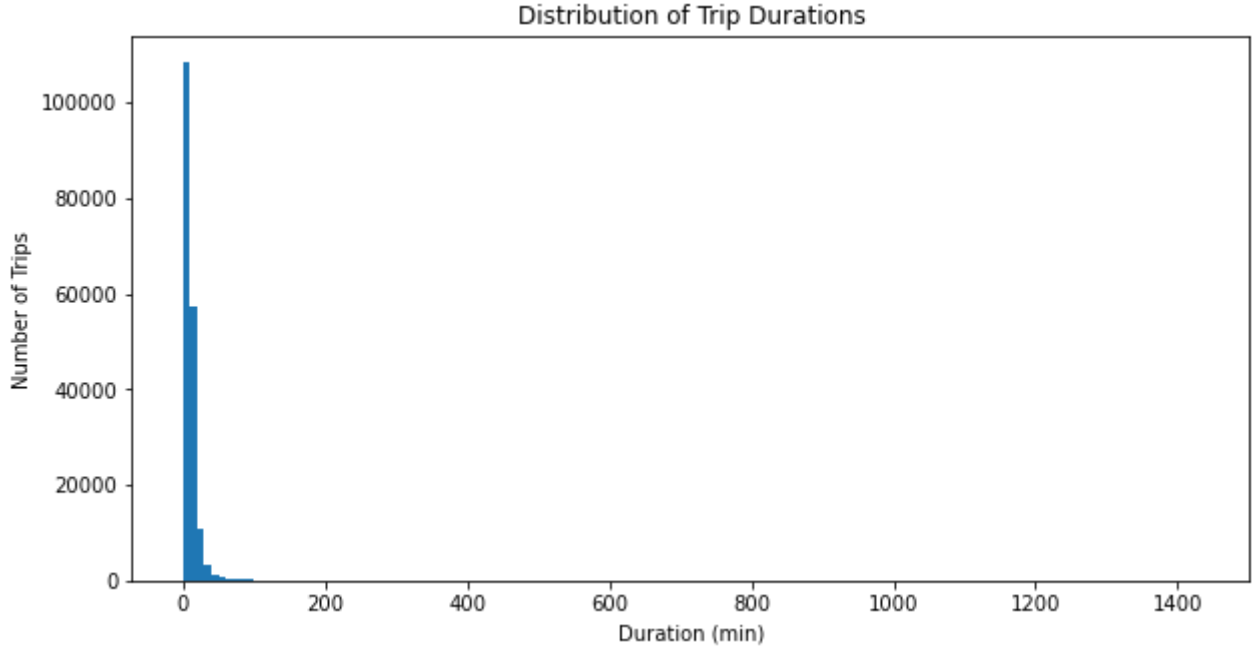
Out[41]:

	duration_sec	start_time	end_time	start_station_id	start_station_name	start_station_latitude	start_station_longitude	end_s
0	52185	2019-02-28 17:32:10.1450	2019-03-01 08:01:55.9750	21.0	Montgomery St BART Station (Market St at 2nd St)	37.789625	-122.400811	13.0
1	42521	2019-02-28 18:53:21.7890	2019-03-01 06:42:03.0560	23.0	The Embarcadero at Steuart St	37.791464	-122.391034	81.0
2	61854	2019-02-28 12:13:13.2180	2019-03-01 05:24:08.1460	86.0	Market St at Dolores St	37.769305	-122.426826	3.0
3	36490	2019-02-28 17:54:26.0100	2019-03-01 04:02:36.8420	375.0	Grove St at Masonic Ave	37.774836	-122.446546	70.0
4	1585	2019-02-28 23:54:18.5490	2019-03-01 00:20:44.0740	7.0	Frank H Ogawa Plaza	37.804562	-122.271738	222.0

Distribution of Bike Trip Duration

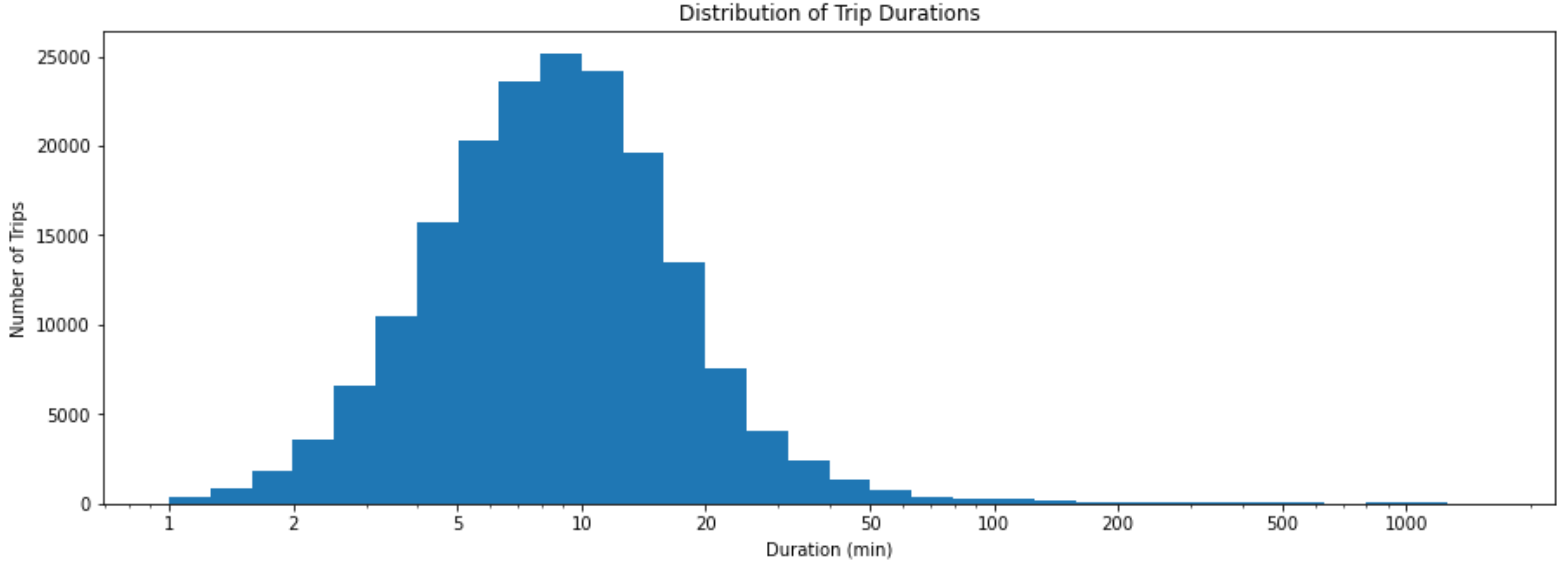
In [54]:

```
# histogram plot displaying bike trips in minutes
binsize = 10
bins = np.arange(0, df['duration_minutes'].max()+binsize, binsize)

plt.figure(figsize =[10,5])
plt.hist(data = df, x = 'duration_minutes', bins = bins)
plt.title('Distribution of Trip Durations')
plt.xlabel('Duration (min)')
plt.ylabel('Number of Trips')
plt.show()
```



In [55]:

```
# logarthmic scale transformation on a histogram
# there's a long tail in the distribution, so let's put it on a log scale instead
log_binsize = 0.1
log_bins = 10 ** np.arange(0.0, np.log10(df['duration_minutes'].max()) + log_binsize, lo
g_binsize)

plt.figure(figsize=[15, 5])
plt.hist(data = df, x = 'duration_minutes', bins = log_bins)
plt.title('Distribution of Trip Durations')
plt.xlabel('Duration (min)')
plt.ylabel('Number of Trips')
plt.xscale('log')
tick_locs = [1, 2, 5, 10, 20, 50, 100, 200, 500, 1000]
plt.xticks(tick_locs, tick_locs)
plt.show()
```



Observations

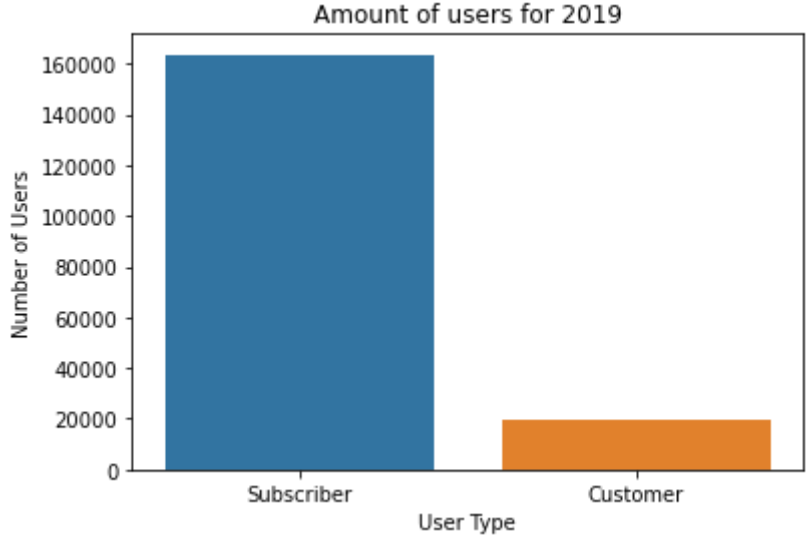
- Bike trip duration has a long tailed distribution.
- It apears bimodial when plotted on a log-scale. Peaks between 8 and 10.
- Most of the bike trips lasts between 8 and 15 minutes.
- The average bike trip is 12 minutes.
- The standard devaition is 29.9.
- 25% of the trips are over 5 minutes, 50% over 8 minutes and 75% over 13 minutes.
- The longest trip is 1424 minutes and the shortest being one minute.

Distribution of Users

User Type Key Customer = 24-hour pass or 3-day pass user Subscriber = Annual Member

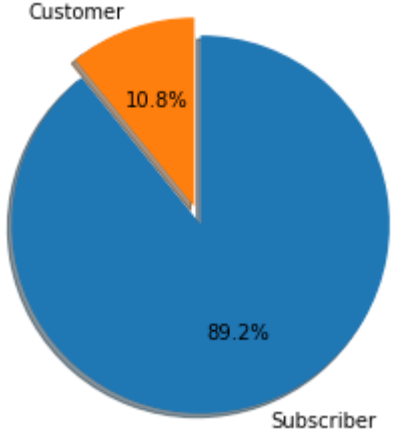
In [57]:

```
# plot a bar chart
user_category = df['user_type'].value_counts().index
sns.countplot(data = df, x = 'user_type', order = user_category)
plt.title('Amount of users for 2019')
plt.xlabel('User Type')
plt.ylabel('Number of Users')
plt.show()
```



In [58]:

```
# plot a pie chart
user_category = df['user_type'].value_counts()
plt.pie(user_category, explode =  (0, 0.1), labels = user_category.index, shadow = True,
startangle = 90,
        counterclock = False, autopct='%1.1f%%');
plt.axis('square')
plt.show()
```



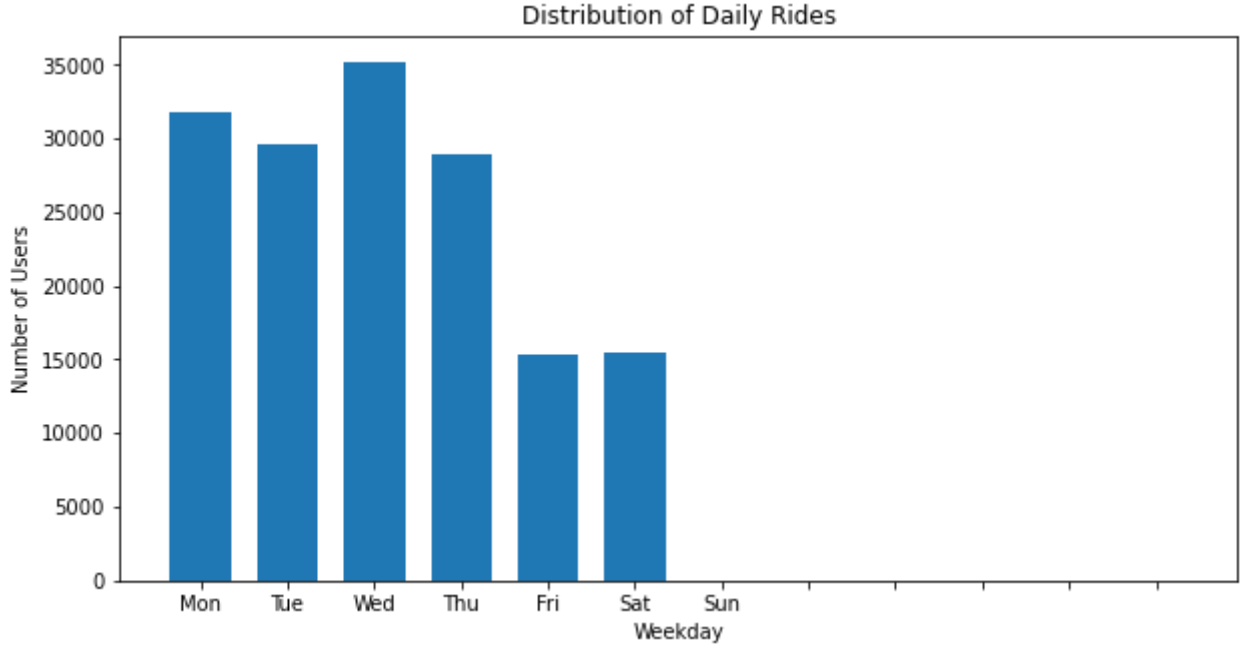
Observations

- Users included Customer that represents those with a 24-hour pass or 3-day pass, and subscribers those with anual membership.
- The bar chart shows over 160,000 subscribing users and 20,000 customers.
- Most users are actually subscribers with annual memberships.

Distribution of Daily Rides

In [60]:

```
# plot a histogram with gaps between bars
bin_edges = np.arange(0.5, 12.5 + 1, 1)
plt.figure(figsize=[10,5])
plt.hist(data = df, x = 'start_weekday', bins = bin_edges, rwidth = 0.7)
plt.xticks(np.arange(1, 12 + 1, 1),weekday_labels)
plt.xlabel('start_weekday')
plt.title('Distribution of Daily Rides')
plt.xlabel('Weekday')
plt.ylabel('Number of Users')
plt.show()
```



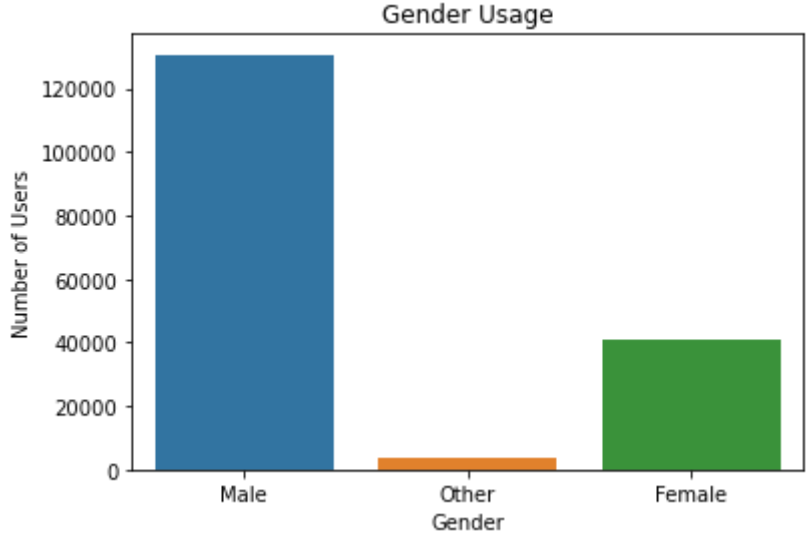
Observations

- Service most used on Wednesdays with over 35,000 for the year.
- The usage decreases significantly on the weekends and no activity on Sunday.

Distribution of Gender

In [62]:
```
# barplot of gender usage
sns.countplot(data = df, x = 'member_gender')
plt.title('Gender Usage')
plt.xlabel('Gender')
plt.ylabel('Number of Users')
```

Out[62]:
```
Text(0, 0.5, 'Number of Users')
```



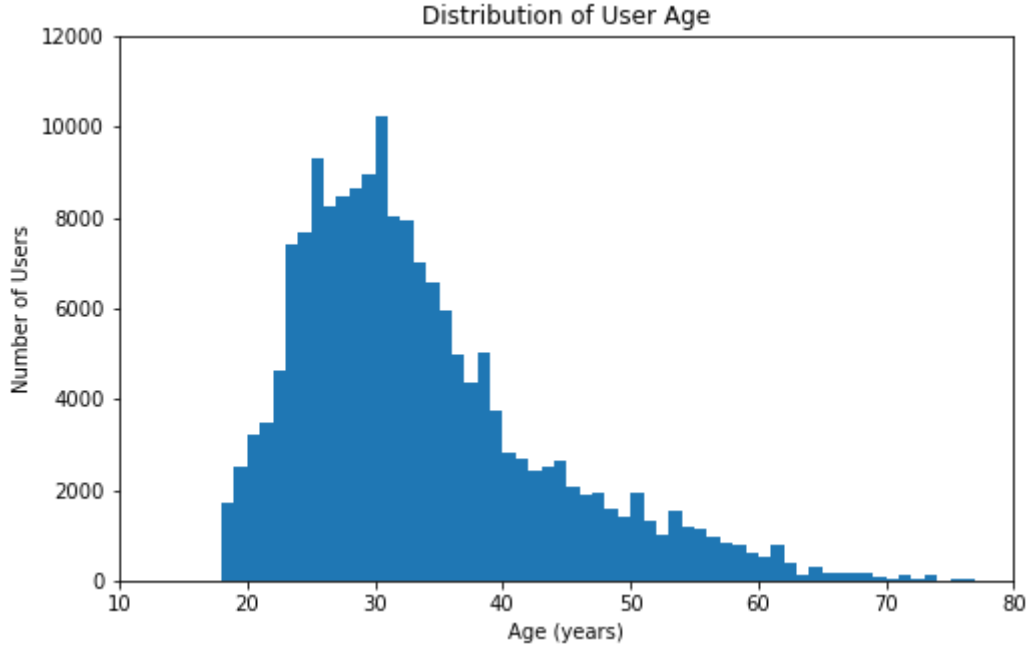
Observations

- Males use the bike service overwhelmingly more than females and other genders.
- Over 120,000 males used the service in 2019.

Age Distribution

In [63]:

```
# Plotting age distribution derived from member's birth year.
binsize = 1
bins = np.arange(0, df['member_birth_year'].astype(float).max()+binsize, binsize)

plt.figure(figsize=[8, 5])
plt.hist(data = df.dropna(), x = 'member_birth_year', bins = bins)
plt.axis([1939, 2009, 0, 12000])
plt.xticks([1939, 1949, 1959, 1969, 1979, 1989, 1999, 2009], [(2019-1939), (2019-1949),
(2019-1959), (2019-1969), (2019-1979), (2019-1989), (2019-1999), (2019-2009)])
plt.gca().invert_xaxis()
plt.title('Distribution of User Age')
plt.xlabel('Age (years)')
plt.ylabel('Number of Users')
plt.show()
```



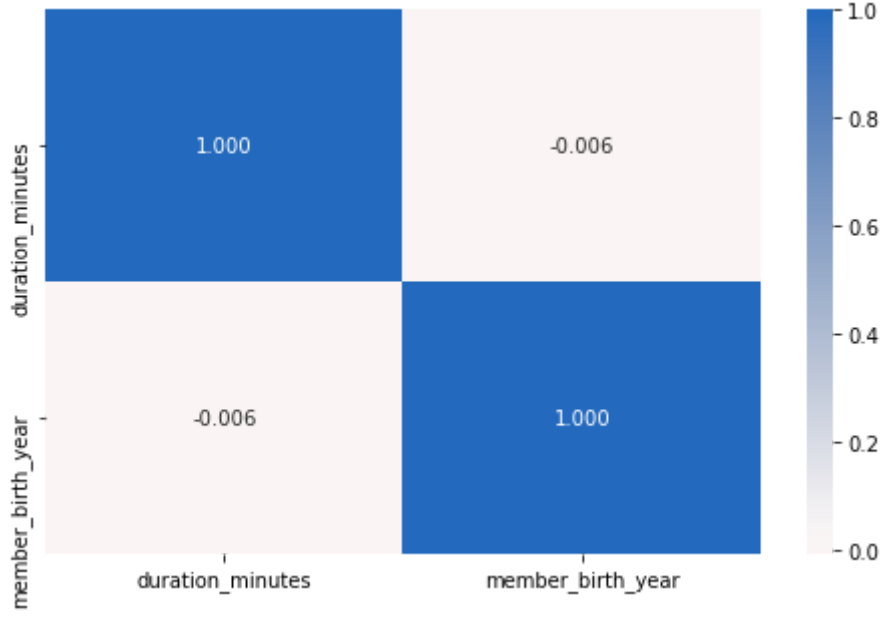
Observations

- Most users are between the age 25 and 35. There is a steady decline in usage from age 35 and up
- Males use the bike service three times more than females and other genders.
- Over 120,000 males used the service in 2019.

Trip Duration and Age

In [66]:

```
# correlation plot -  a negative correlation
plt.figure(figsize = [8, 5])
sns.heatmap(df[numeric_vars].corr(), annot = True, fmt = '.3f',cmap = 'vlag_r', center =
0)
plt.show()
```

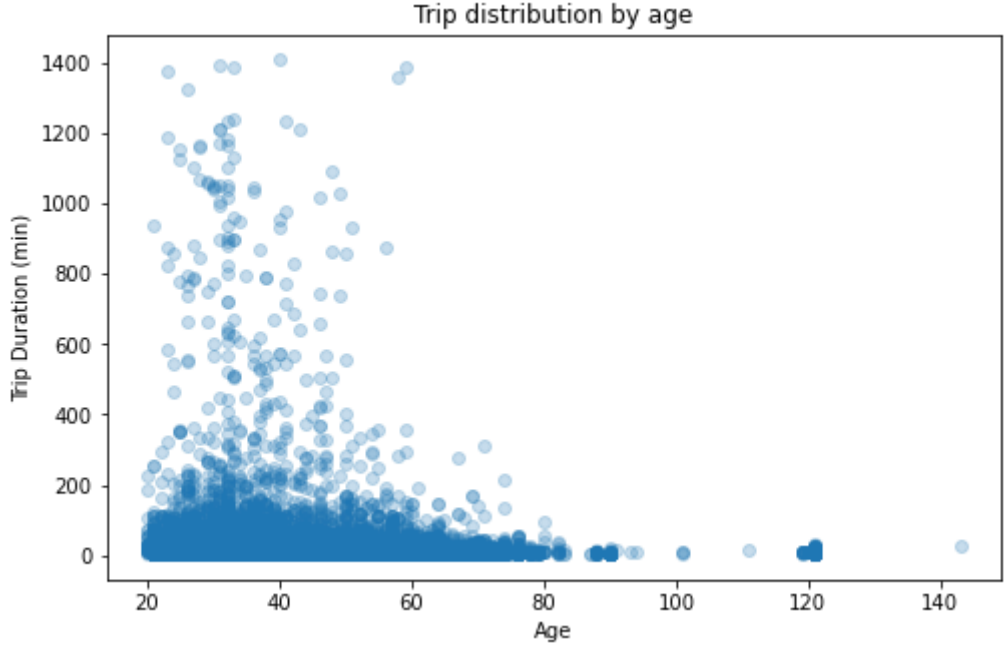


Observation

You can observe that there is a negative correlation where the age decreases as the trip duration increases.

In [68]:

```
# a scatter plot.
plt.figure(figsize=[8,5])
plt.scatter(data = df, x = 'age', y ='duration_minutes', alpha=.25)
plt.title('Trip distribution by age')
plt.xlabel('Age')
plt.ylabel('Trip Duration (min)')
plt.show()
```



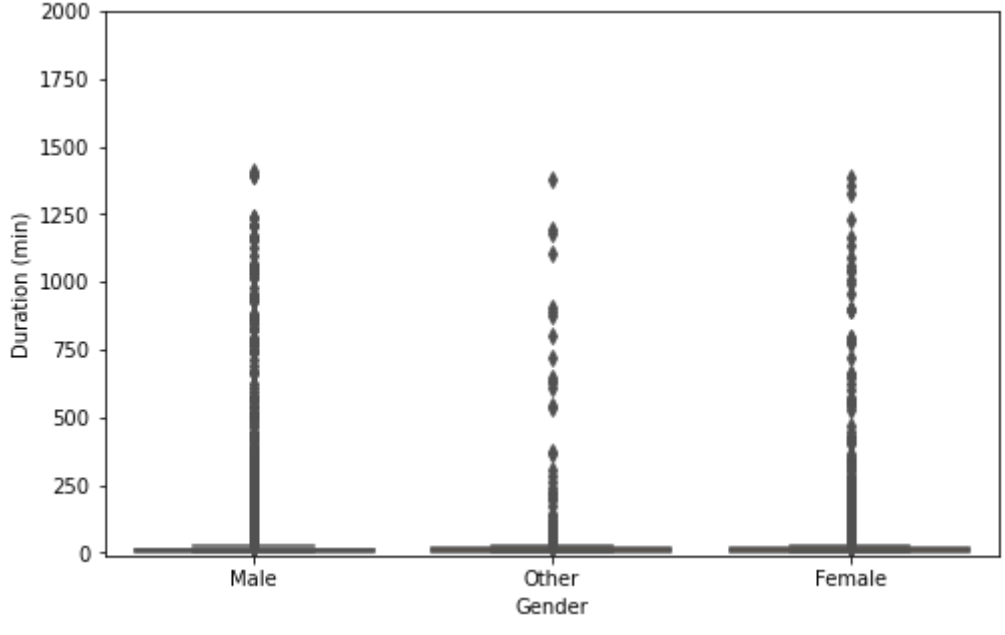
Observation

The concentration of rides are for persons between ages 25 and 45 showing the inverse relationship between age and the trip duration.

Trip Duration and Gender

In [69]:

```
# boxplot displaying the gender correlation with trip duration
plt.figure(figsize = [8, 5])
base_color = sns.color_palette()[1]
sns.boxplot(data = df, x = 'member_gender', y = 'duration_minutes', color = base_color)
plt.ylim([-10, 2000])
plt.xlabel('Gender')
plt.ylabel('Duration (min)')
plt.show()
```

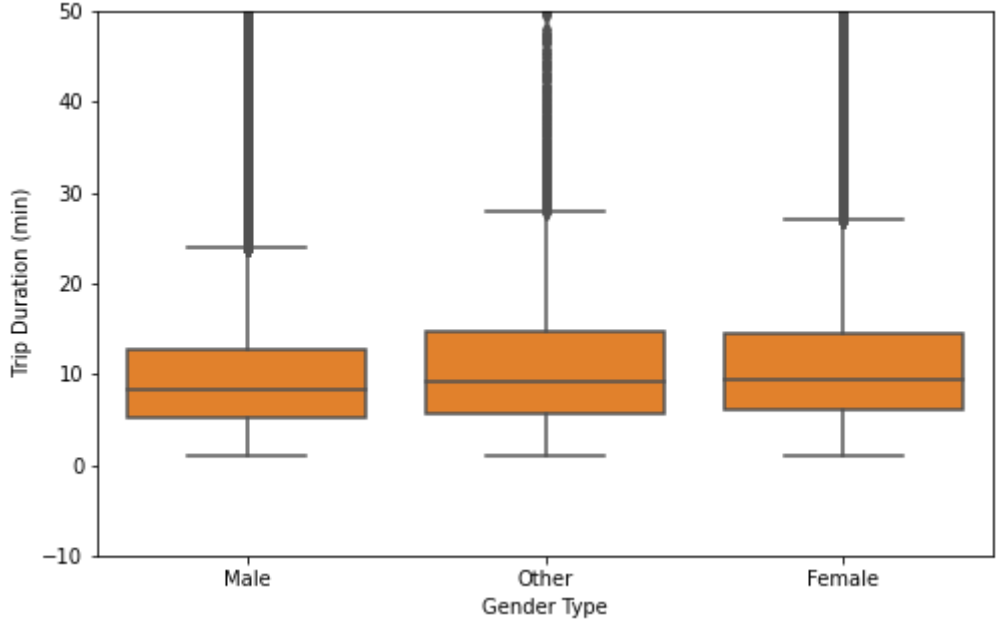


Observation

Trimming the trip duration y axis values is best so we can better view the box plot. Trimmed to 50 minutes.

In [70]:

```
plt.figure(figsize = [8, 5])
base_color = sns.color_palette()[1]
sns.boxplot(data = df, x = 'member_gender', y = 'duration_minutes', color = base_color)
plt.ylim([-10, 50])
plt.xlabel('Gender Type')
plt.ylabel(' Trip Duration (min)')
plt.show()
```



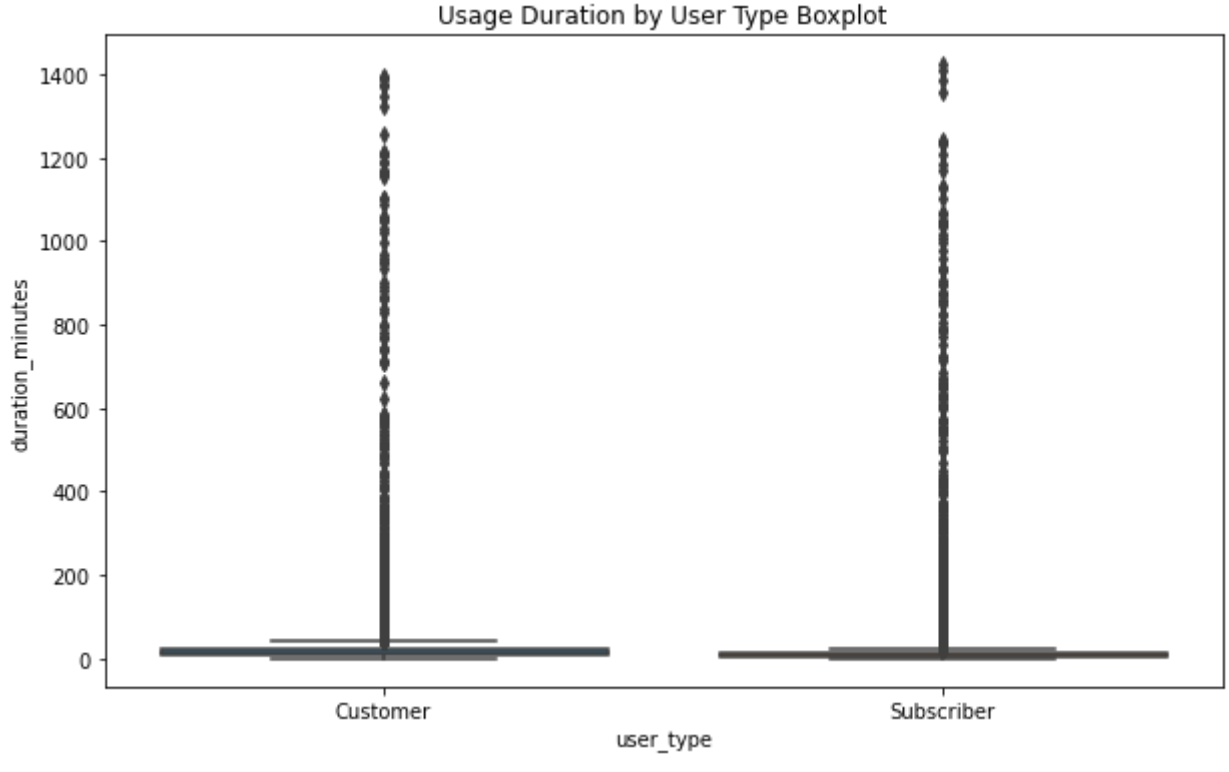
Observation

The boxplot does show that female and other gender have a higher trip duration than males.

Trip Duration and User Type

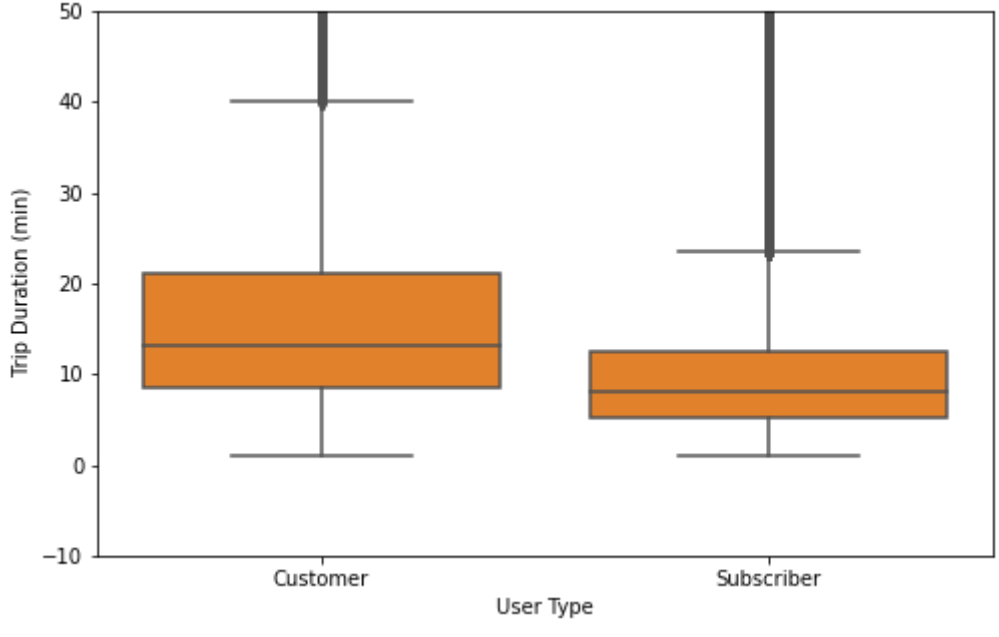
In [72]:

```
# box plot comparing customer and subscriber over time

plt.figure(figsize=(10,6))
plt.title('Usage Duration by User Type Boxplot')
sns.boxplot(data=df, x='user_type', y='duration_minutes');
```



In [73]:

```
plt.figure(figsize = [8, 5])
base_color = sns.color_palette()[1]
sns.boxplot(data = df, x = 'user_type', y = 'duration_minutes', color = base_color)
plt.ylim([-10, 50])
plt.xlabel('User Type')
plt.ylabel(' Trip Duration (min)')
plt.show()
```



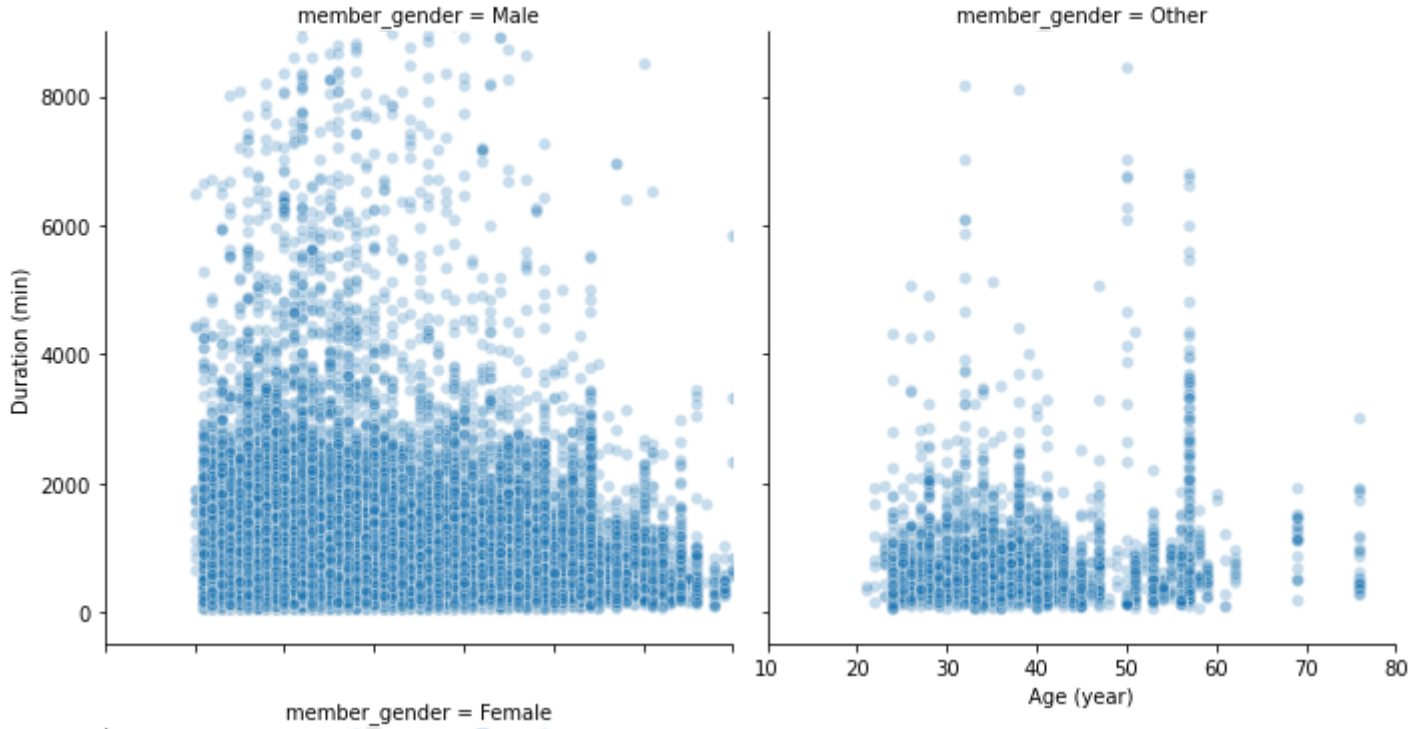
Observation

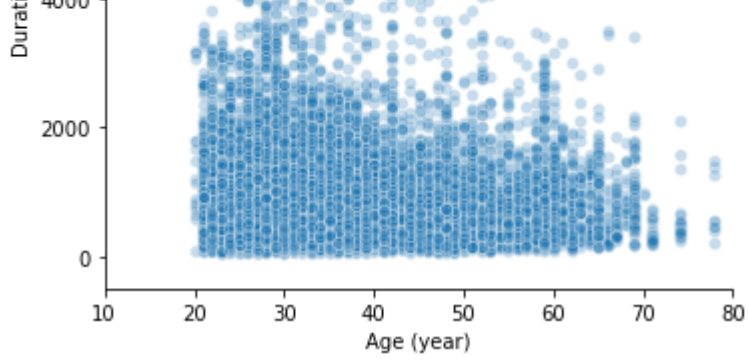
The customer is spending more time on a bike trip than subscribers.

Trip Duration vs Age and Gender Type

In [76]:

```
# https://seaborn.pydata.org/generated/seaborn.FacetGrid.html
genders = sns.FacetGrid(data = df, col = 'member_gender', col_wrap = 2, height = 5, xlim
= [10, 80], ylim = [-500, 9000])
genders.map(sns.scatterplot, 'age', 'duration_sec', alpha=0.25)
genders.set_xlabels('Age (year)')
genders.set_ylabels('Duration (min)')
plt.show()
```




4000
2000
0
10
20
30
40
50
60
70
80
Age (year)

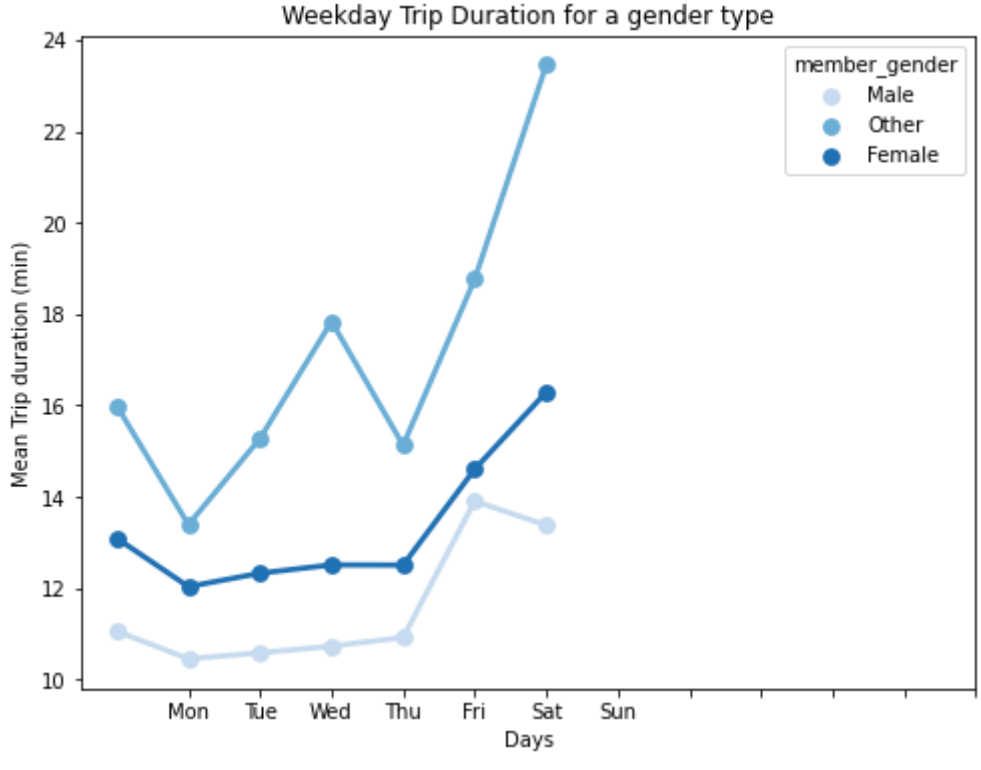
Observation

- Comparing the gender types as it relates to trip duration, the age 20 to 40 represents the group that does most of the rides. * Females and males do appear to have similar ride average.

Trip Duration vs Weekday and Gender Type

In [78]:

```
fig = plt.figure(figsize = [8,6])
sns.pointplot(data = df, x = 'start_weekday', y = 'duration_minutes', hue = 'member_gend
er', palette = 'Blues', ci=None)
plt.title('Weekday Trip Duration for a gender type')
plt.ylabel('Mean Trip duration (min)')
plt.xlabel('Days')
plt.xticks(np.arange(1, 12 + 1, 1),weekday_labels)
plt.show()
```



Observation

- The trip duration start trending up on the weekends from Thursdays to Saturdays.
- Males still have the shortest bike trip.

The End