Ford GoBike System Data

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]:
```
# load in the dataset into a pandas dataframe, print statistics
df = pd.read_csv('201902-fordgobike-tripdata.csv')
```

In [3]:
```
df.head()
```

Out[3]:

	duration_sec	start_time	end_time	start_station_id	start_station_name	start_station_lati
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)	37.78
1	42521	2019-02-28 18:53:21.7890	2019-03-01 06:42:03.0560	23.0	The Embarcadero at Steuart St	37.79
2	61854	2019-02-28 12:13:13.2180	2019-03-01 05:24:08.1460	86.0	Market St at Dolores St	37.76
3	36490	2019-02-28 17:54:26.0100	2019-03-01 04:02:36.8420	375.0	Grove St at Masonic Ave	37.77
4	1585	2019-02-28 23:54:18.5490	2019-03-01 00:20:44.0740	7.0	Frank H Ogawa Plaza	37.80

In [4]:
```
df.tail()
```

Out[4]:

	duration_sec	start_time	end_time	start_station_id	start_station_name	start_static
183407	480	2019-02-01 00:04:49.7240	2019-02-01 00:12:50.0340	27.0	Beale St at Harrison St	
183408	313	2019-02-01 00:05:34.7440	2019-02-01 00:10:48.5020	21.0	Montgomery St BART Station (Market St at 2nd St)	
183409	141	2019-02-01 00:06:05.5490	2019-02-01 00:08:27.2200	278.0	The Alameda at Bush St	
183410	139	2019-02-01 00:05:34.3600	2019-02-01 00:07:54.2870	220.0	San Pablo Ave at MLK Jr Way	
183411	271	2019-02-01 00:00:20.6360	2019-02-01 00:04:52.0580	24.0	Spear St at Folsom St	

In [5]:
```
# high-level overview of data shape and composition
print(df.shape)
print(df.dtypes)
```

```
(183412, 16)
duration_sec                 int64
start_time                  object
end_time                    object
start_station_id           float64
start_station_name          object
start_station_latitude     float64
start_station_longitude    float64
end_station_id             float64
end_station_name            object
end_station_latitude       float64
end_station_longitude      float64
bike_id                      int64
user_type                   object
member_birth_year          float64
member_gender               object
bike_share_for_all_trip     object
dtype: object
```

In [6]:
```
#Understand the presence of Customers
Sub = df[df.user_type == 'Customer'].count()
Sub
```

Out[6]:
```
duration_sec                19868
start_time                  19868
end_time                    19868
start_station_id            19801
start_station_name          19801
start_station_latitude      19868
start_station_longitude     19868
end_station_id              19801
end_station_name            19801
end_station_latitude        19868
end_station_longitude       19868
bike_id                     19868
user_type                   19868
member_birth_year           16631
member_gender               16631
bike_share_for_all_trip     19868
dtype: int64
```

In [7]:
```
#Count the number of Custs
len(df[df['user_type' ] == 'Customer'])
```

Out[7]:
```
19868
```

In [8]:
```
# Counts the Number of Subs
len(df[df['user_type' ] == 'Subscriber'])
```

Out[8]:
```
163544
```

In [9]:
```
#Convert Times to_dateTime
df['start_time'] = pd.to_datetime(df['start_time'])
df['end_time'] = pd.to_datetime(df['end_time'])
```

In [10]:
```
#Add Days of the month, week, and hour that bikes were used.

df['day'] = df.start_time.dt.strftime('%A')
df['Day'] = df.start_time.dt.day
df['Hour'] = df.start_time.dt.hour
```

In [11]:
```
#Min max date
a = df.start_time.max().strftime('%Y-%m-%d')
b = df.start_time.min().strftime('%Y-%m-%d')

c = df.duration_sec.max()
d = df.duration_sec.min()

e = df[df.user_type == 'Subscriber'].duration_sec.mean()
f = df[df.user_type == 'Customer'].duration_sec.mean()

# are subscribers consistent riders

g = df[df.user_type == 'Subscriber'].nunique().count()
h = df[df.user_type == 'Customer'].nunique().count()

#count the average time a subscriber or Customer uses the bikes
print(df[df.user_type == 'Customer'].nunique().value_counts().mean())
print(df[df.user_type == 'Subscriber'].nunique().value_counts().mean())
print(" ")

print(""""The recorded rides are from {} to {}
        \nMax and least time duration in seconds being {} and {}
        \nAverage duration for a subscriber {}
        \nAverage duration for a Customer {}
        \nNumber of unique Subscirbers {}
        \nNumber of unique Customer {}

        """.format(b,a,c,d, e, f, g, h))
```

```
1.3571428571428572
1.4615384615384615

"The recorded rides are from 2019-02-01 to 2019-02-28

Max and least time duration in seconds being 85444 and 61

Average duration for a subscriber 640.2636782761825

Average duration for a Customer 1432.465019126233

Number of unique Subscirbers 19

Number of unique Customer 19
```

In [12]:
```
# we need to know the distance between the start and end locations
from math import sin, cos, sqrt, atan2, radians

dis = []

for h,i,j,k in zip( df.start_station_latitude, df.start_station_longitud
e, df.end_station_latitude, df.end_station_longitude ):
    # approximate radius of earth in km
    R = 6378.0

    lat1 = radians(h)
    lon1 = radians(i)
    lat2 = radians(j)
    lon2 = radians(k)

    dlon = lon2 - lon1
    dlat = lat2 - lat1

    a = sin(dlat / 2)**2 + cos(lat1) * cos(lat2) * sin(dlon / 2)**2
    c = 2 * atan2(sqrt(a), sqrt(1 - a))

    distance = R * c

    dis.append(distance)
```

In [13]:
```
#add list to Dataframe
df['Distance'] = dis
```

In [14]:
```
df.head()
```

Out[14]:

	duration_sec	start_time	end_time	start_station_id	start_station_name	start_station_latitu
0	52185	2019-02-28 17:32:10.145	2019-03-01 08:01:55.975	21.0	Montgomery St BART Station (Market St at 2nd St)	37.7896
1	42521	2019-02-28 18:53:21.789	2019-03-01 06:42:03.056	23.0	The Embarcadero at Steuart St	37.7914
2	61854	2019-02-28 12:13:13.218	2019-03-01 05:24:08.146	86.0	Market St at Dolores St	37.7693
3	36490	2019-02-28 17:54:26.010	2019-03-01 04:02:36.842	375.0	Grove St at Masonic Ave	37.7748
4	1585	2019-02-28 23:54:18.549	2019-03-01 00:20:44.074	7.0	Frank H Ogawa Plaza	37.8045

What is the structure of your dataset?

There are 183,412 recorded rides in the dataset with 16 features. Most of which are numeric and some are objects.

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

-duration_sec
-start_time and end_time
-start station id and end station id -Start and end latitude and longitude -start station name and end station name
-user_type (were they a Subscriber?)
-member_birth_year
-member_gender

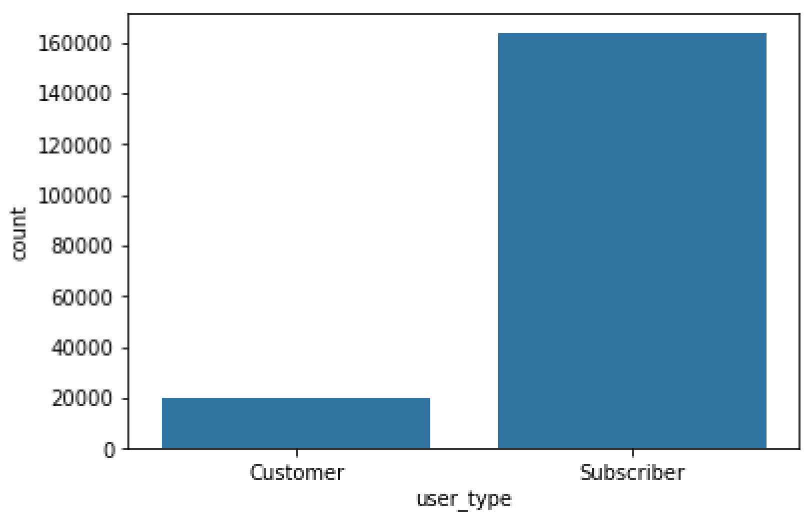
What features in the dataset do you think will help support your investigation into your feature(s) of interest?

If they choose to subsribe then my assumption is they did not want to miss out on discounts of deals since they intended to use the bikes often.

Univariate Exploration

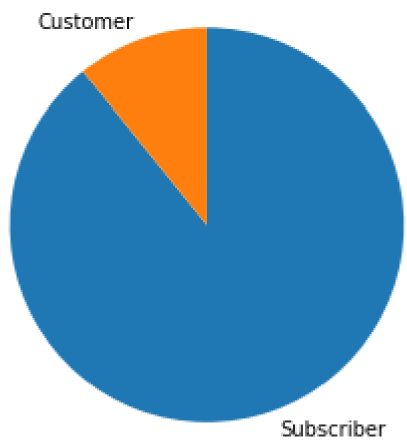
Ill start by looking at each variable to see if they have a corrilation with User type, Usage, and Gender.

In [15]:
```
# Who used the bikes more during February.
base_color = sb.color_palette()[0]
sb.countplot(data=df, x='user_type', color=base_color);
```



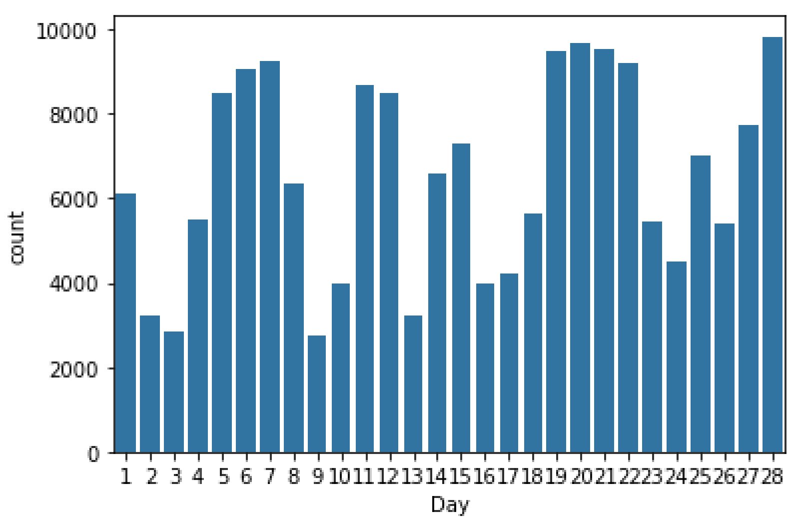
It Looks like there is a greatter count of Subscibers then Customers for the month of February. Something that is expected from Subscribers.

In [16]:
```
#Customers to subscribers
sorted_counts = df['user_type'].value_counts()
plt.pie(sorted_counts, labels = sorted_counts.index, startangle = 90, co
unterclock = False);
plt.axis('square');
```



Again there is a greater proportion of Subscribers to Customers

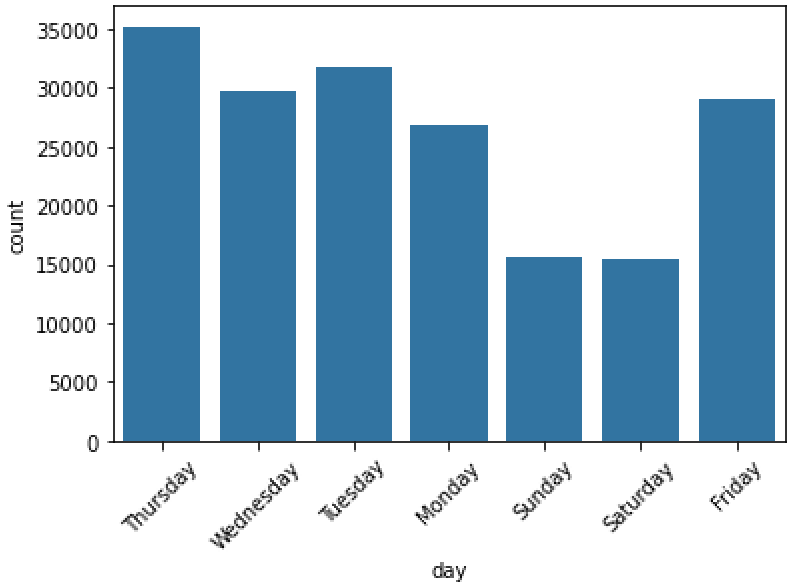
In [17]:
```
#Bike usage throughout the month of February
base_color = sb.color_palette()[0]
sb.countplot(data=df, x='Day', color=base_color);
```



It looks like there is a drop every weekend during the Month of February.

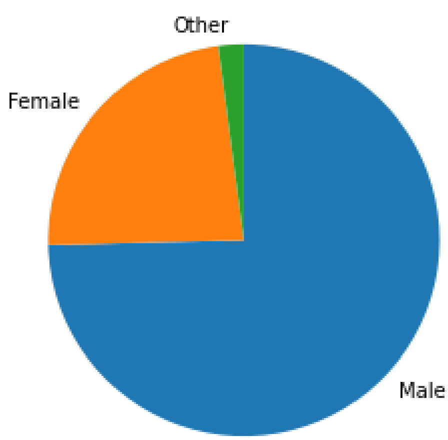
In [ ]:

In [18]:
```
#Bike usage throughout the week

base_color = sb.color_palette()[0]
sb.countplot(data=df, x='day', color=base_color);
plt.xticks(rotation = 45);
```



It looks like over all there is more usage activity during the weekdays and then a drop during the weekend.

In [19]:
```
# Males to Females in the data
sorted_counts = df['member_gender'].value_counts()
plt.pie(sorted_counts, labels = sorted_counts.index, startangle = 90, co
unterclock = False);
plt.axis('square');
```



There is significantly more men riding bikes then there are women or other genders.

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

I am interested in looking at how the usage of the customers and Subscirbers differ throughout the week and seeing if being a subscriber really does make a difference. A difference in how long they ride for and the kind of distance that is covered. I want to answer whether companys should try to have more subscirbers.

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?

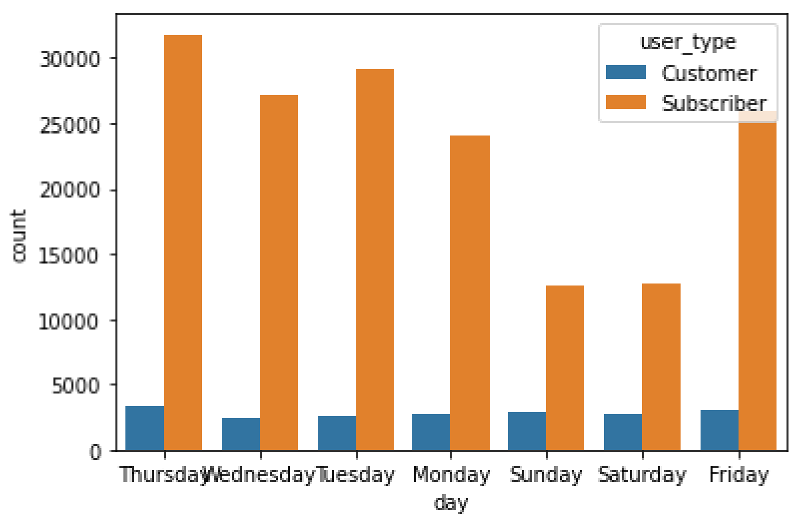
One of the variabels thaht i had to adjust was the time to be able to catergorize the days of the month, and the month date along with it. I also had to know the distance between each start and end location to know how the distance of each Bike ride.

Bivariate Exploration

To start off with, I want to look at the the relations ships that being a Customer or Subsciber has with how long and hwo far you bike.

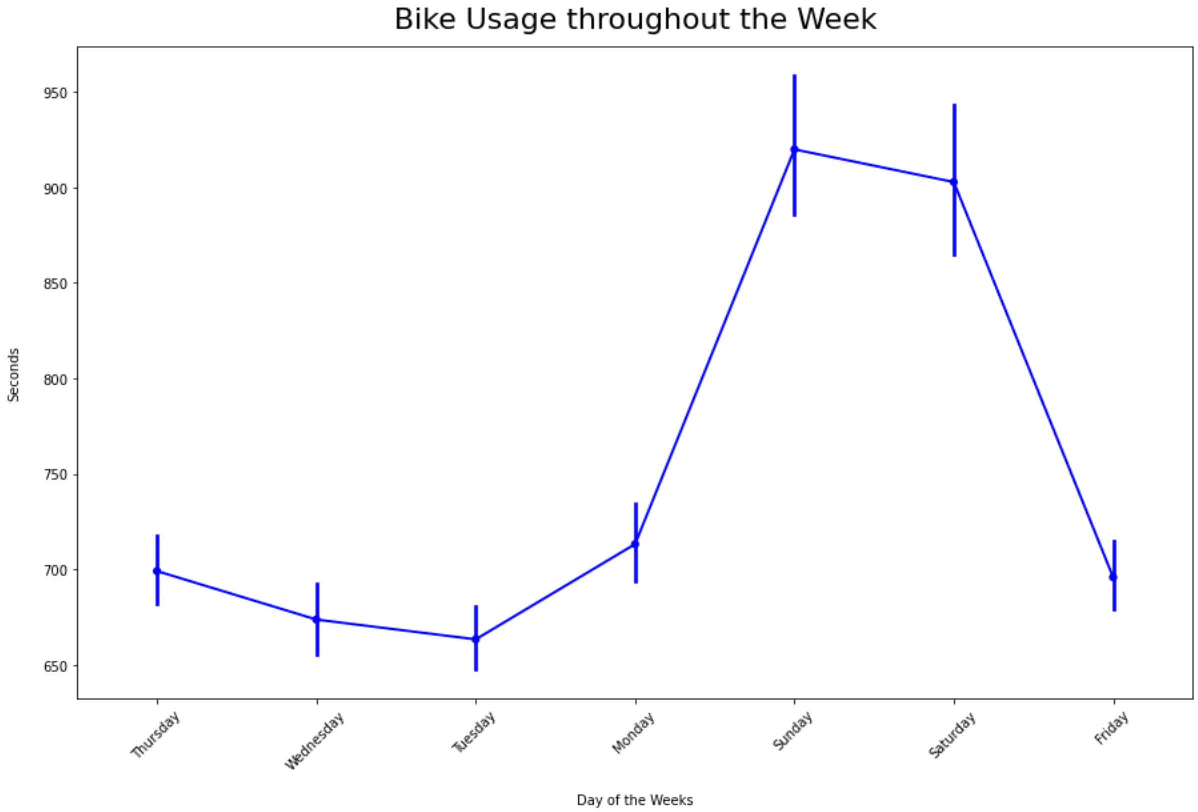
In [20]:
```
#Do Custmers use the bike or Subscibers.
sb.countplot(data = df, x = 'day', hue = 'user_type')
```

Out[20]: <matplotlib.axes._subplots.AxesSubplot at 0x7fd936d0b790>



Subscibers will always surpass customers in the number of time they will want to ride bikes.

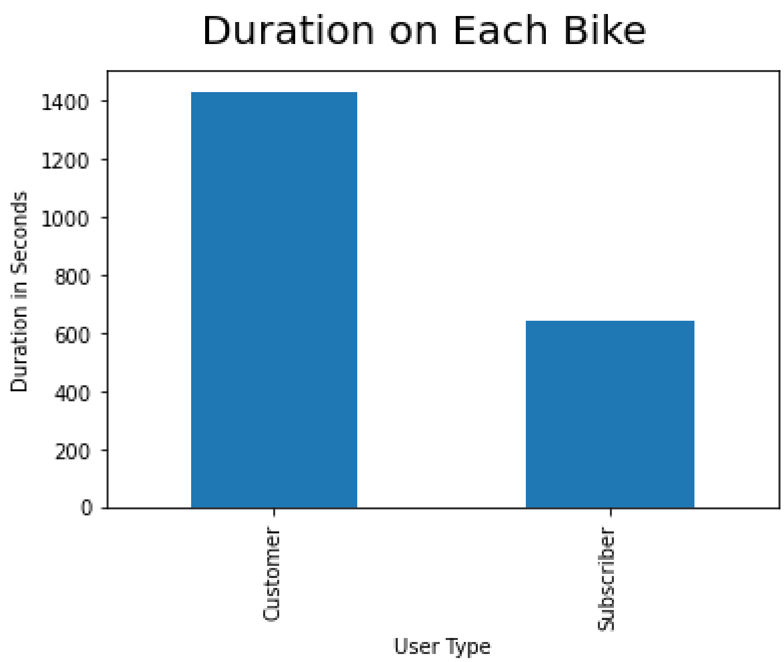
In [21]:
```
plt.figure(figsize=(15,9))
sb.pointplot(data = df, x='day', y='duration_sec', scale=.7, color='blu
e')
plt.title('Bike Usage throughout the Week', fontsize=22, y=1.015)
plt.xlabel('Day of the Weeks', labelpad=16)
plt.ylabel('Seconds', labelpad=16)
plt.xticks(rotation = 45);
```



People will tend to have longer rides during the weekend, people are not commuting as much they are mostly riding recreationally.

In [22]:
```
#Customers tend to spend longer times on the bike then subscribers
df.groupby('user_type').duration_sec.mean().plot(kind = 'bar');
plt.ylabel('Duration in Seconds', fontsize = 10)
plt.xlabel('User Type',fontsize = 10 )
plt.suptitle('Duration on Each Bike ', fontsize = 20)
```

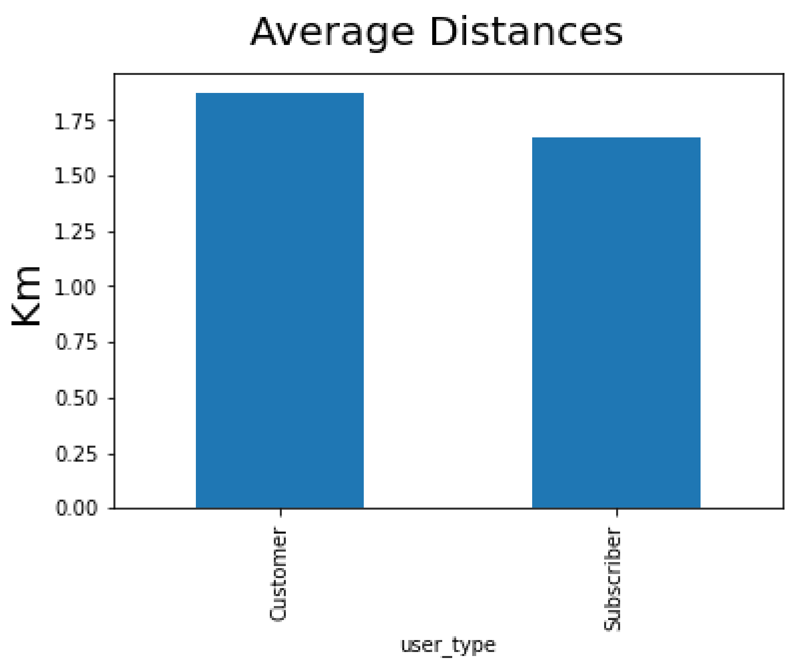
Out[22]:
```
Text(0.5, 0.98, 'Duration on Each Bike ')
```



I did not expect that it would be customers on average that have longer lasting rides but it might make sense since when they do decide to ride it is most often for recreation then commuting where Subscribers tend to know they use the bikes a lot and dont spend long amounts of time on each ride.

In [23]:
```
#Customers on average go farther distances but there isnt a huge differe
nce.

df.groupby('user_type').Distance.mean().plot(kind = 'bar');
plt.ylabel('Km', fontsize = 20)

plt.suptitle('Average Distances', fontsize = 20)
```

Out[23]: Text(0.5, 0.98, 'Average Distances')



Its is also shocking to know that customers on average spend longer times and travel the farthest distance.

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?

It was interesting to see that customer are definetly more casual then subscibers when they choose to ride bikes, they tend to ride them for longer times and for farther distances then the subscribers do. It can also be expected that subscirbers are more regular with how often they us the bikes as they ride them during the week which might mean that they use them to commute to work.

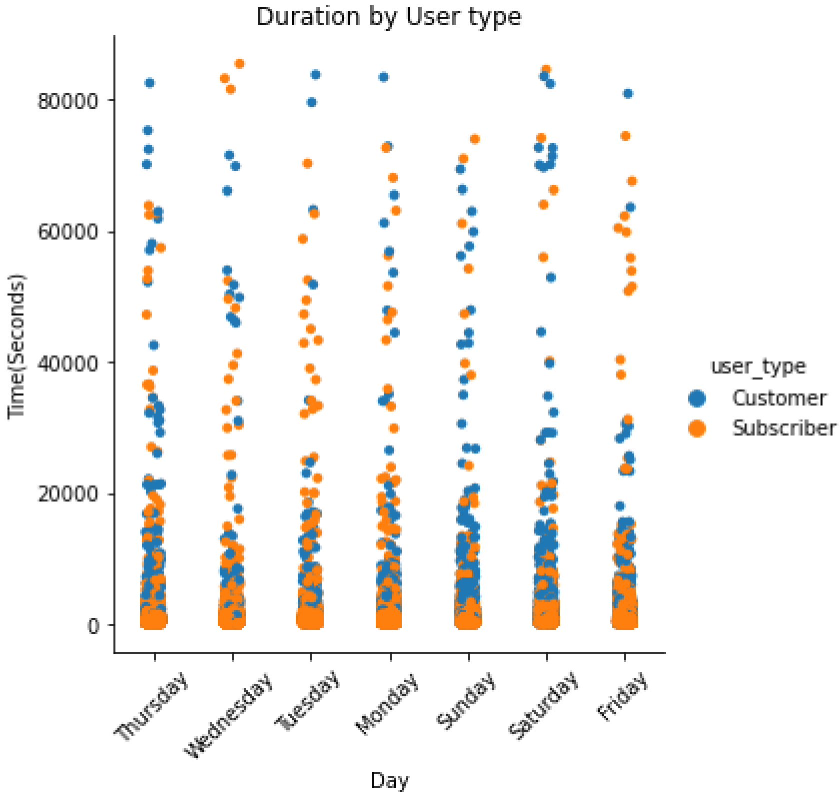
Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

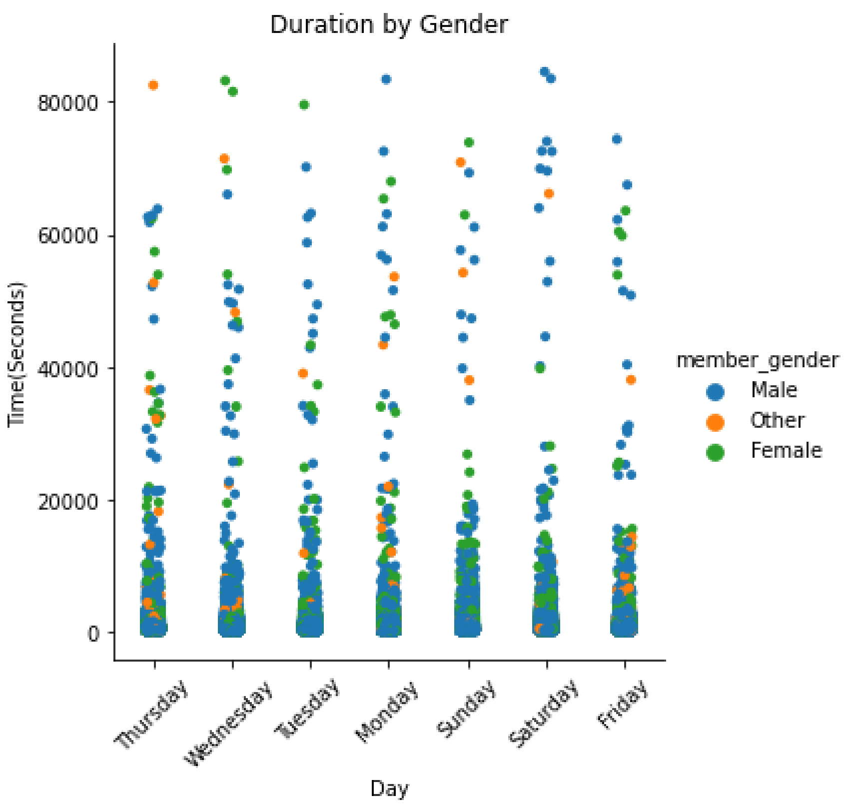
Men tend to ride bikes more than any other gender.

Multivariate Exploration

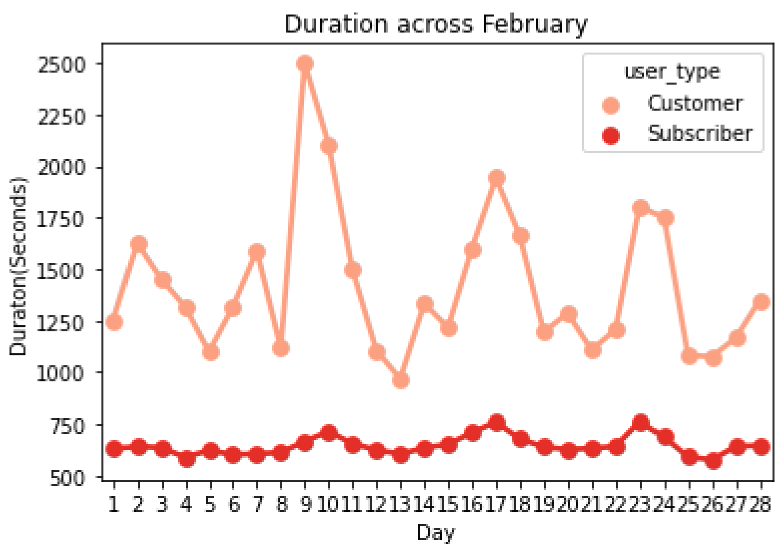
First, I will analyze the distributions between customer and subscriber, based on gender, and on weekday and monthly usage.

In [24]:
```
#which Gender and user type spends the most time on the bike.
g = sb.catplot(x="day", y="duration_sec", hue="user_type", data=df);
plt.xticks(rotation=45);
plt.title('Duration by User type');
plt.ylabel('Time(Seconds)');
plt.xlabel('Day');

g = sb.catplot(x="day", y="duration_sec", hue="member_gender", data=df);
plt.xticks(rotation=45);
plt.title('Duration by Gender');
plt.ylabel('Time(Seconds)');
plt.xlabel('Day');
```


Duration by User type
Time(Seconds)
0
20000
40000
60000
80000
Thursday
Wednesday
Tuesday
Monday
Sunday
Saturday
Friday
Day
user_type
Customer
Subscriber

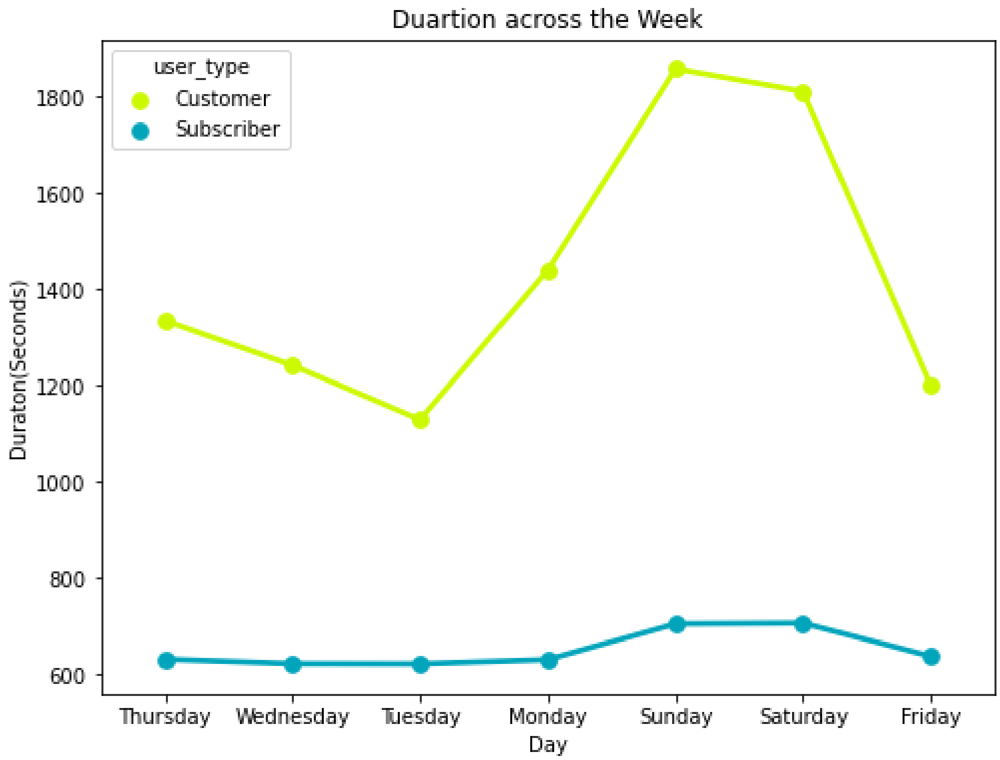

Duration by Gender
Time(Seconds)
0
20000
40000
60000
80000
Thursday
Wednesday
Tuesday
Monday
Sunday
Saturday
Friday
Day
member_gender
Male
Other
Female

In [25]:
```
# Which USER_TYPE( Customer / Subscriber) spends longer times on the bik
e.

sb.pointplot(data = df, x = 'Day', y = 'duration_sec', hue = 'user_type'
, palette = 'Reds', ci=None)
plt.title('Duration across February')
plt.ylabel('Duraton(Seconds)')
plt.xlabel('Day')
plt.show();
```



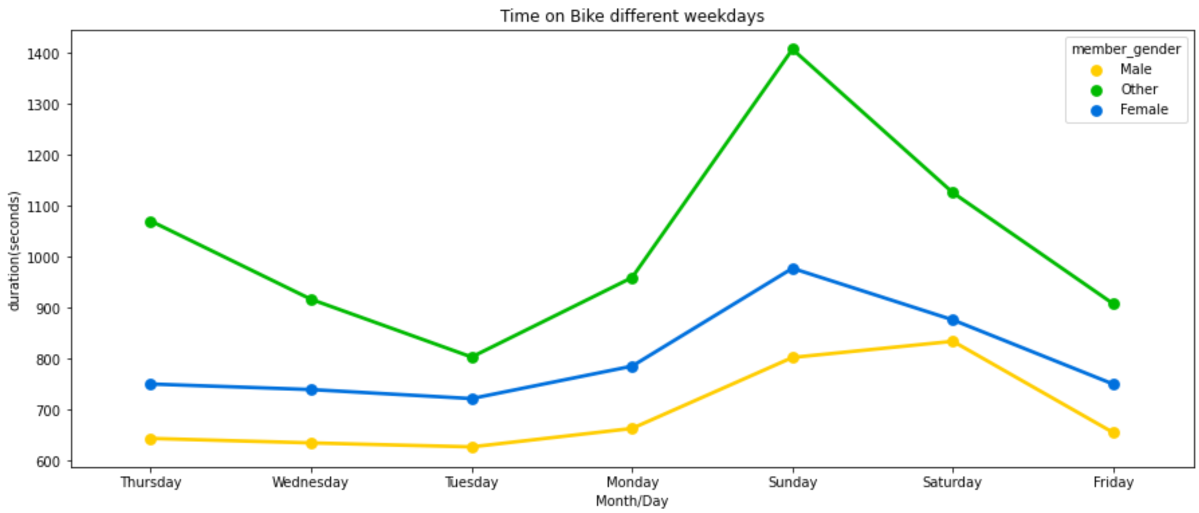
There is never a moment through out the month were a customer does not outride a subscriber. Ther alway spend a significant longer time on the bike

In [26]:
```
#Deeper look into the TIME the user has on the bike
fig = plt.figure(figsize = [8,6])
sb.pointplot(data = df, x = 'day', y = 'duration_sec', hue = 'user_type'
, palette = 'nipy_spectral_r', ci=None)
plt.title('Duartion across the Week')
plt.ylabel('Duraton(Seconds)')
plt.xlabel('Day')
plt.show();
```

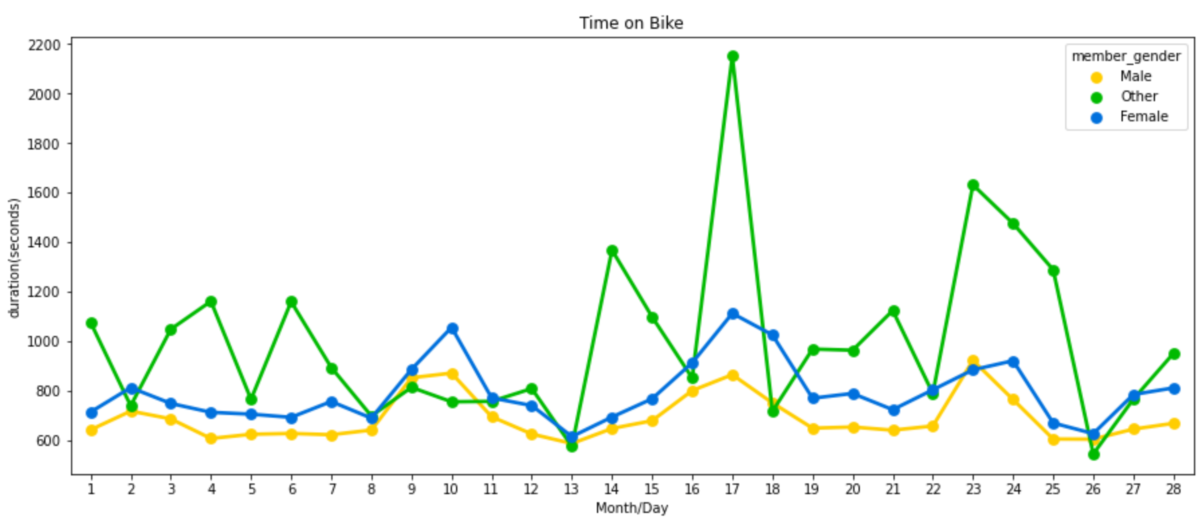


This graph makes it obvious that Customers use the bike more recreatonally as they mostly use them during the weekend. There is also a spike in usage for Subsribers but not as significant as Customers.

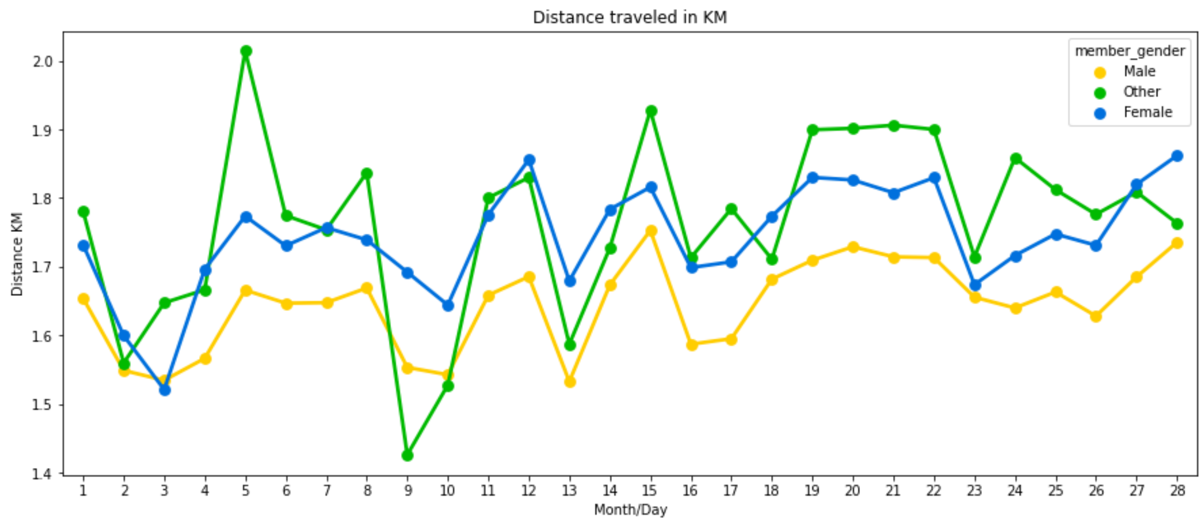
In [27]:
```
#Duration across the week from different genders.
fig = plt.figure(figsize = [15,6])
sb.pointplot(data = df, x = 'day', y = 'duration_sec', hue = 'member_gen
der', palette = 'nipy_spectral_r', ci=None)
plt.title('Time on Bike different weekdays')
plt.ylabel('duration(seconds)')
plt.xlabel('Month/Day')
plt.show();
```



In [28]:
```
#Duration across the week from different genders.
fig = plt.figure(figsize = [15,6])
sb.pointplot(data = df, x = 'Day', y = 'duration_sec', hue = 'member_gen
der', palette = 'nipy_spectral_r', ci=None)
plt.title('Time on Bike')
plt.ylabel('duration(seconds)')
plt.xlabel('Month/Day')
plt.show();
```

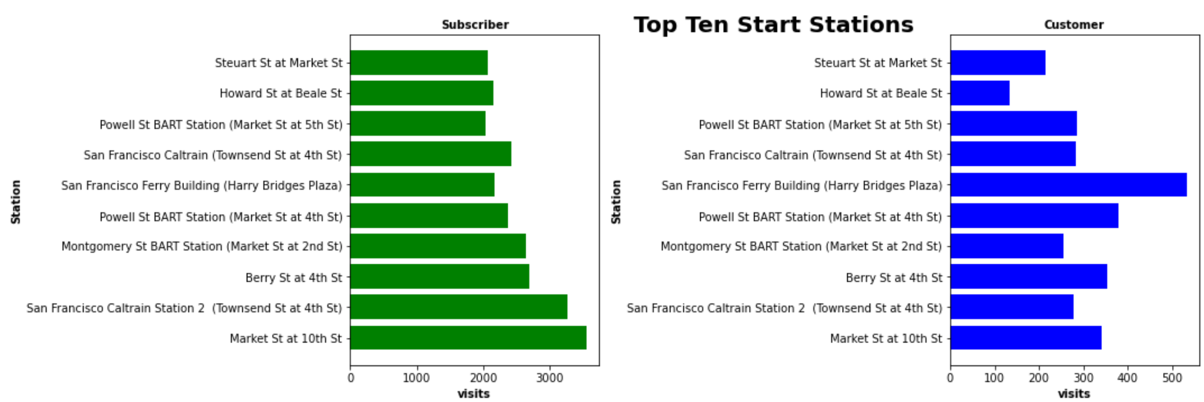


In [29]:
```
#Distance across the month from different genders.
fig = plt.figure(figsize = [15,6])
sb.pointplot(data = df, x = 'Day', y = 'Distance', hue = 'member_gender'
, palette = 'nipy_spectral_r', ci=None)
plt.title('Distance traveled in KM')
plt.ylabel('Distance KM')
plt.xlabel('Month/Day')
plt.show();
```



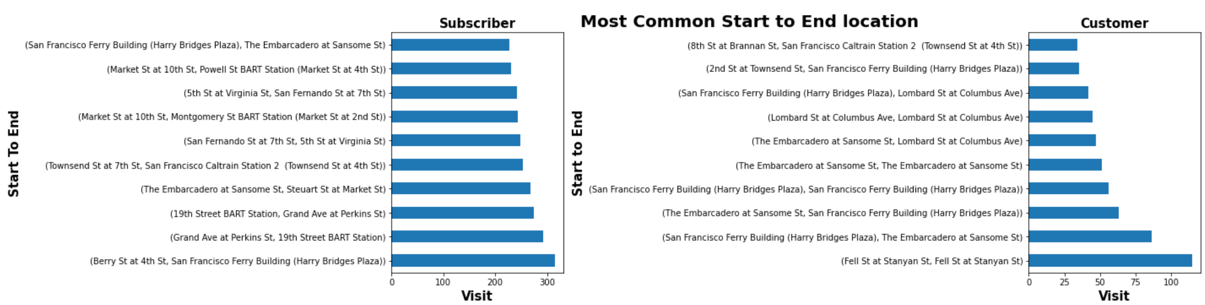
It is interesting that woman tend to ride longer and farther the men on most day in February.

In [30]:
```
# Top ten start stations by different users

df_sub = df.query('user_type == "Subscriber"').start_station_name.value_
counts(ascending=True)
q = df.start_station_name.value_counts().nlargest(10)
q
q = q.index[0:10]
q = q.astype('category')



plt.figure(figsize=(15,5))

plt.subplot(1, 2, 1)
plt.barh(q,df_sub[q],color='Green',  align='center');
plt.title('Subscriber', fontweight='semibold', fontsize=10)
plt.ylabel('Station', fontweight='semibold', fontsize=10)
plt.xlabel('visits', fontweight='semibold', fontsize=10)

plt.subplot(1, 2, 2)
df_Cus = df.query('user_type == "Customer"').start_station_name.value_co
unts(ascending=True)
plt.barh(q,df_Cus[q],color='blue', align='center');
plt.title('Customer', fontweight='semibold', fontsize=10)
plt.ylabel('Station', fontweight='semibold', fontsize=10)
plt.xlabel('visits', fontweight='semibold', fontsize=10)

plt.suptitle('                                                      Top Ten Start S
tations', fontsize=20, fontweight='semibold')

plt.tight_layout()
```



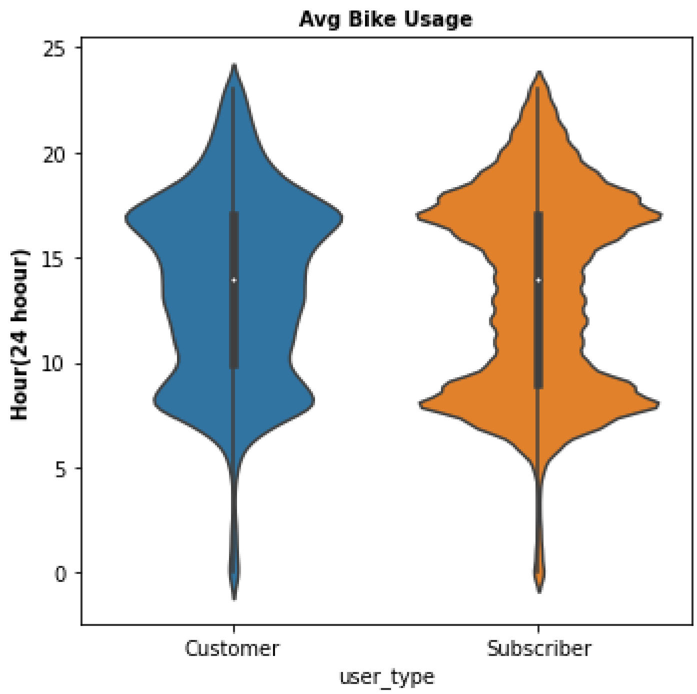
It is proabably good to know that these are the top Ten station for reasons that there need to be more bikes placed there. It is also interesting at that both customers and Subscibers have the same start location, this probably means that they are optimally placed since subscirbers and Customers often have different reason to ride, these bikes are placed to suite both their needs which makes the company efficient.

In [31]:
```
# What is a Customers / Subscribers usual start and end location.

plt.figure(figsize=(20,5))

plt.suptitle('                                                                Most Common Start to End location ', fontsize=20, fontweight='semibold')

plt.subplot(1, 2, 1)
count_series = df.query('user_type == "Subscriber"').groupby(['start_station_name', 'end_station_name']).size().nlargest(10)
count_series.plot(kind = 'barh')
plt.title('Subscriber', fontweight='semibold', fontsize=15)
plt.ylabel('Start To End',fontweight='semibold', fontsize=15)
plt.xlabel('Visit',fontweight='semibold', fontsize=15)

plt.subplot(1, 2, 2)
count_se = df.query('user_type == "Customer"').groupby(['start_station_name', 'end_station_name']).size().nlargest(10)
count_se.plot(kind = 'barh')
plt.title('Customer', fontweight='semibold', fontsize=15)
plt.ylabel('Start to End', fontweight='semibold', fontsize=15)
plt.xlabel('Visit', fontweight='semibold', fontsize=15)



plt.tight_layout()
```



While we can see that the Custoemrs have lower visiting numbers than Subscibers , We can also see that they have entirely different rout probably because of the commute vs recreational use.

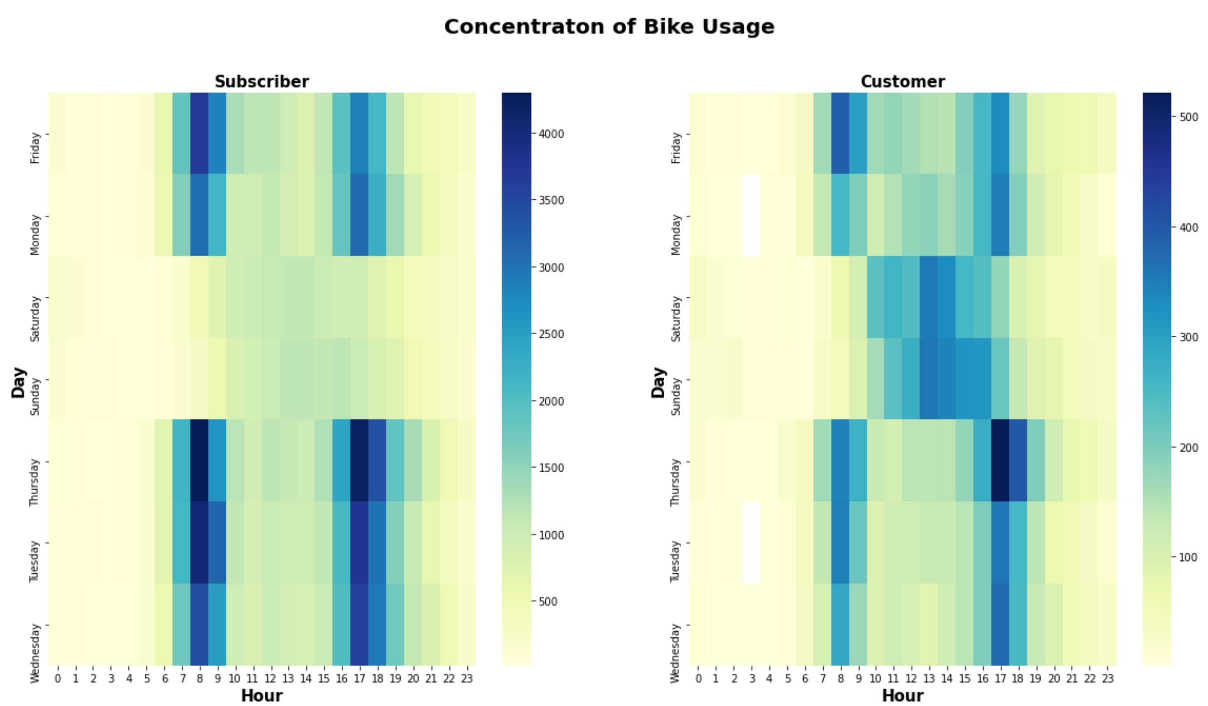
In [32]:
```
# Here were can see where on any average the bike usage hours pick up an
d slow down.

plt.figure(figsize=(5,5))
sb.violinplot(data=df, x='user_type', y='Hour');

plt.title('Avg Bike Usage', fontweight='semibold', fontsize=10)
plt.ylabel('Hour(24 hoour)', fontweight='semibold', fontsize=10)



plt.tight_layout()
```



This graph helps us see on average the busiest hours of the day for Bike riding. It looks liek its either early in the mornign or later in the afternoon.

In [33]:
```
#Here we can see the concentraton of BIKE USAGE during the HOURS of the
 WEEKDAYS.

df_sub = df.query('user_type == "Subscriber"').groupby(["day", "Hour"])[
"bike_id"].size().reset_index();
df_cust = df.query('user_type == "Customer"').groupby(["day", "Hour"])[
"bike_id"].size().reset_index();

plt.figure(figsize=(20,10));

plt.suptitle('Concentraton of Bike Usage  ', fontsize=20, fontweight='se
mibold')


df_sub = df_sub.pivot("day", "Hour", "bike_id");
df_cust = df_cust.pivot("day", "Hour", "bike_id");
plt.subplot(1, 2, 1);
sb.heatmap(df_sub, cmap="YlGnBu");
plt.title('Subscriber', fontweight='semibold', fontsize=15)
plt.ylabel('Day', fontweight='semibold', fontsize=15)
plt.xlabel('Hour', fontweight='semibold', fontsize=15)


plt.subplot(1, 2, 2);

sb.heatmap(df_cust, cmap="YlGnBu");
plt.title('Customer', fontweight='semibold', fontsize=15)
plt.ylabel('Day', fontweight='semibold', fontsize=15)
plt.xlabel('Hour', fontweight='semibold', fontsize=15)
```

Out[33]:
```
Text(0.5, 69.0, 'Hour')
```



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?

The multivariate charts above have the same observations made in previous charts. The Customers prefer to ride for longer perods of time and travel father. Customers also prefer to bike during the weeekend while the Subscribers are more consistent and prefer to ride during the week days . We can also see the subscribers and customers have very differnt routs. But while they have different routes we can also see that both subscribers and customers have simialar start locations so we can conclude that the company is being effecient with how they place their bike because its suiting the needs of two different types of customers.

Were there any interesting or surprising interactions between features?

Gender is also interesting because while there are more men biking, Women in February tend to ride for longer periods of time and bike farther.