Analysis of Ford GoBike System Data in the Bay Area

Yu Tao 2020/12/29

Investigation Overview

The goal of the investigation is to utilize the Ford GoBike system data and analyze the users' activities. Specifically, I would like to look at how each ride's duration and distance are affected by other features, such as the user's age, gender, type, the ride's hour of day, day of week and regions.

Dataset Overview

The Ford GoBike system data set (201902-Fordgobike-Tripdata.csv) includes information about individual rides made in a bike-sharing system covering the greater San Francisco Bay area. In the data set, there are 183412 observation of renting activities made in February 2019, including the duration, start/end time, start/end station name/position of the ride, and basic infomation about the rider (gender, birthyear, customer/subscriber). I performed data wrangling in an exploration part and the cleaned data can now be found in GoBike_Cleaned.csv.

In [15]:

```
# 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

# suppress warnings from final output
import warnings
warnings.simplefilter("ignore")
```

In [16]:

```
# load in the dataset into a pandas dataframe
df_bike = pd.read_csv('GoBike_Cleaned.csv')
df_bike.head()
```

Out[16]:

	duration_sec	start_time	end_time	start_station_id	start_station_name	start_station_latitude	start_station_longitude	e
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)	37.789625	-122.400811	
1	42521	2019-02-28 18:53:21.789	2019-03-01 06:42:03.056	23.0	The Embarcadero at Steuart St	37.791464	-122.391034	
2	61854	2019-02-28 12:13:13.218	2019-03-01 05:24:08.146	86.0	Market St at Dolores St	37.769305	-122.426826	
3	36490	2019-02-28 17:54:26.010	2019-03-01 04:02:36.842	375.0	Grove St at Masonic Ave	37.774836	-122.446546	
4	1585	2019-02-28 23:54:18.549	2019-03-01 00:20:44.074	7.0	Frank H Ogawa Plaza	37.804562	-122.271738	

5 rows × 23 columns

Ride Duration and Distance Distributions

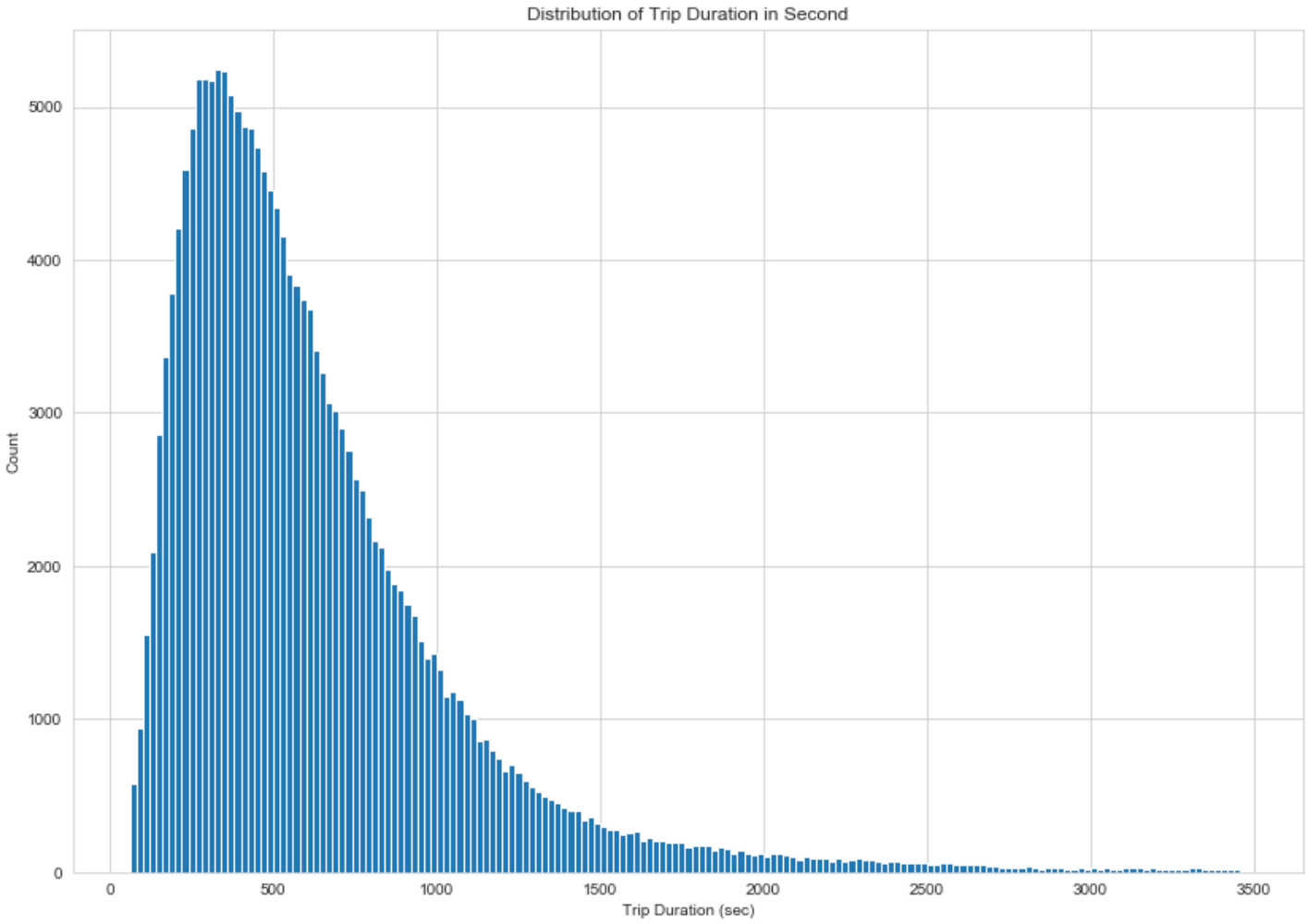
Ride Duration and Distance Distributions

The distribution of ride duration is positively(right)-skewed, with an average value of 621 seconds (10.35 min) on the 181382 cleaned observations. 75% of the riders finished their rides within 784 seconds (13.07 min).

The distribution of the trip distance is also positively(right)-skewed. The average distance of each trip is ~1.68 km. 75% of the trips are within 2.22 km.

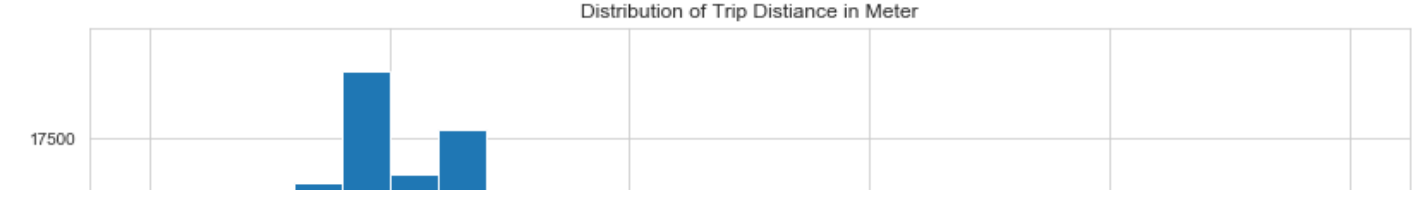
In [17]:

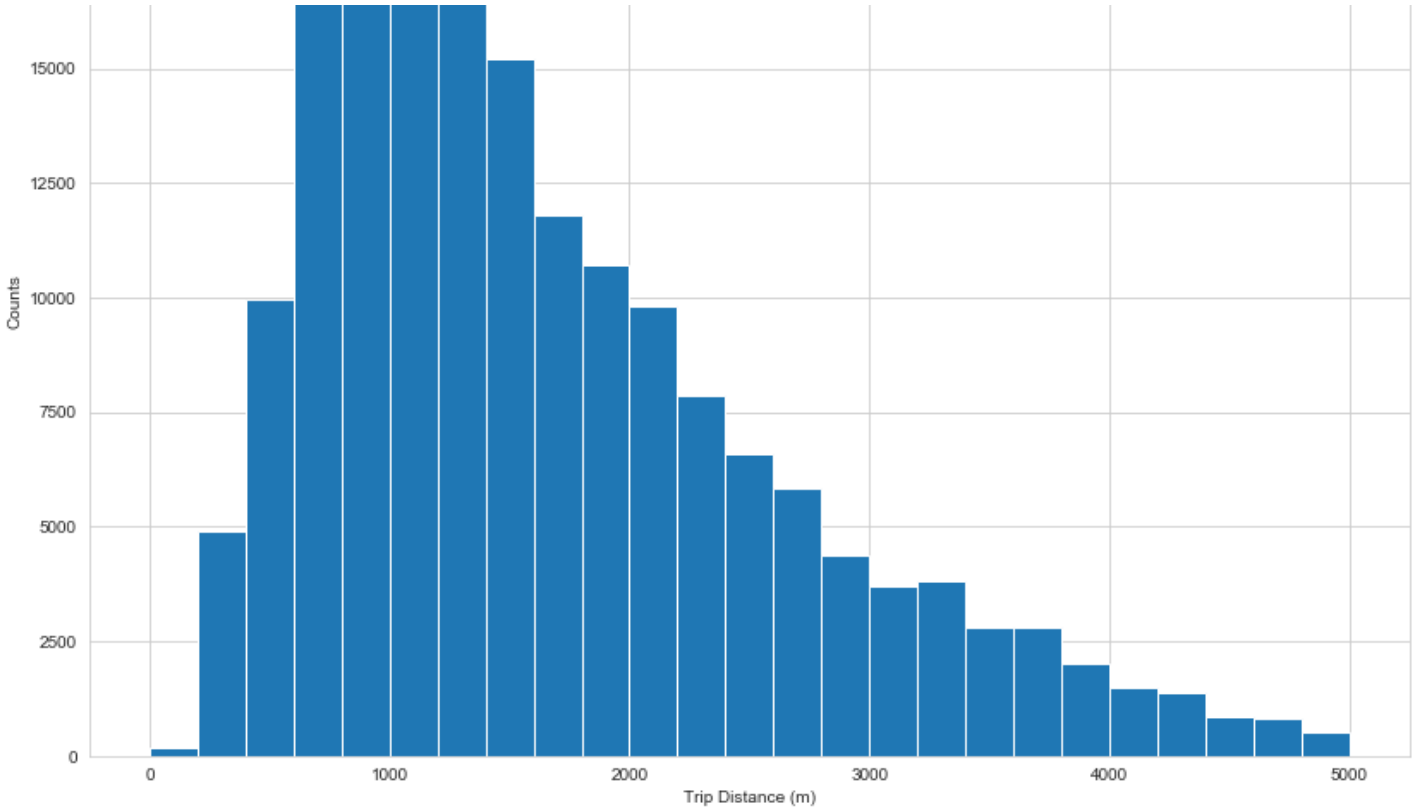
```
# distribution of trip duration
sb.set_style("whitegrid")
plt.figure(figsize=[11.69, 8.27])
plt.hist(data = df_bike.query('duration_sec < 3457'), x = 'duration_sec', bins = np.aran
ge(60, 3500, 20))
plt.xlabel('Trip Duration (sec)')
plt.ylabel('Count')
plt.title('Distribution of Trip Duration in Second')
plt.tight_layout();
```



In [18]:

```
# distribution of trip distance
sb.set_style("whitegrid")
plt.figure(figsize=[11.69, 8.27])
plt.hist(data = df_bike.query('(distance > 1) & (distance < 5070)'), x = 'distance', bin
s = np.arange(0, 5070, 200))
plt.xlabel('Trip Distance (m)')
plt.ylabel('Counts')
plt.title('Distribution of Trip Distiance in Meter')
plt.tight_layout();
```



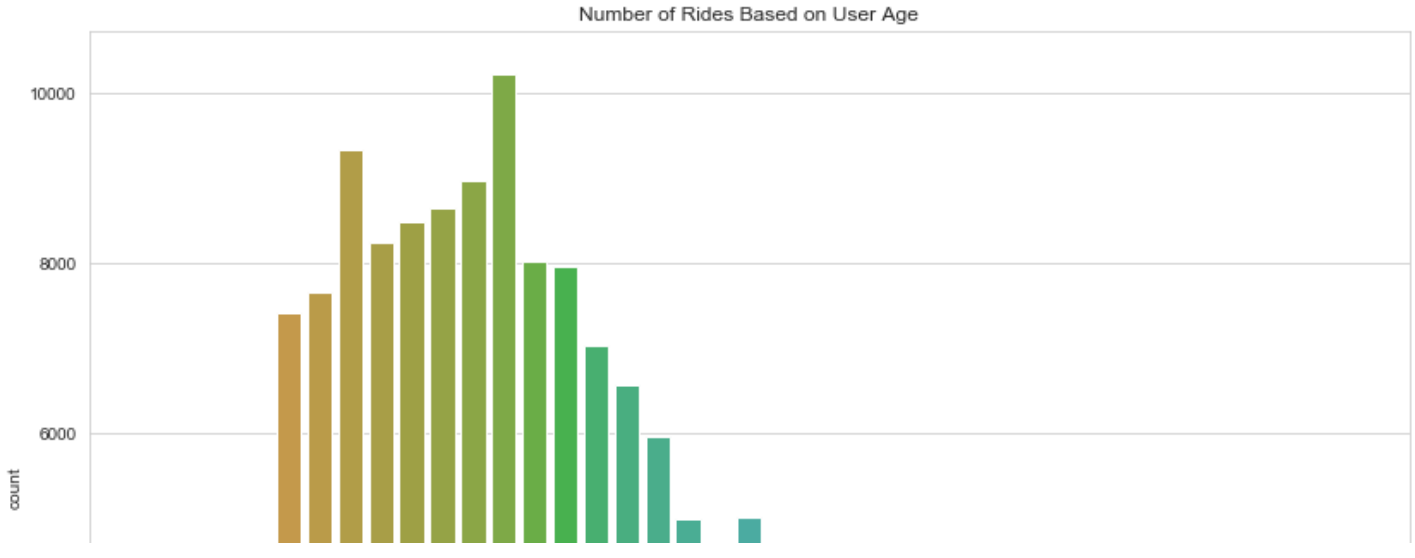


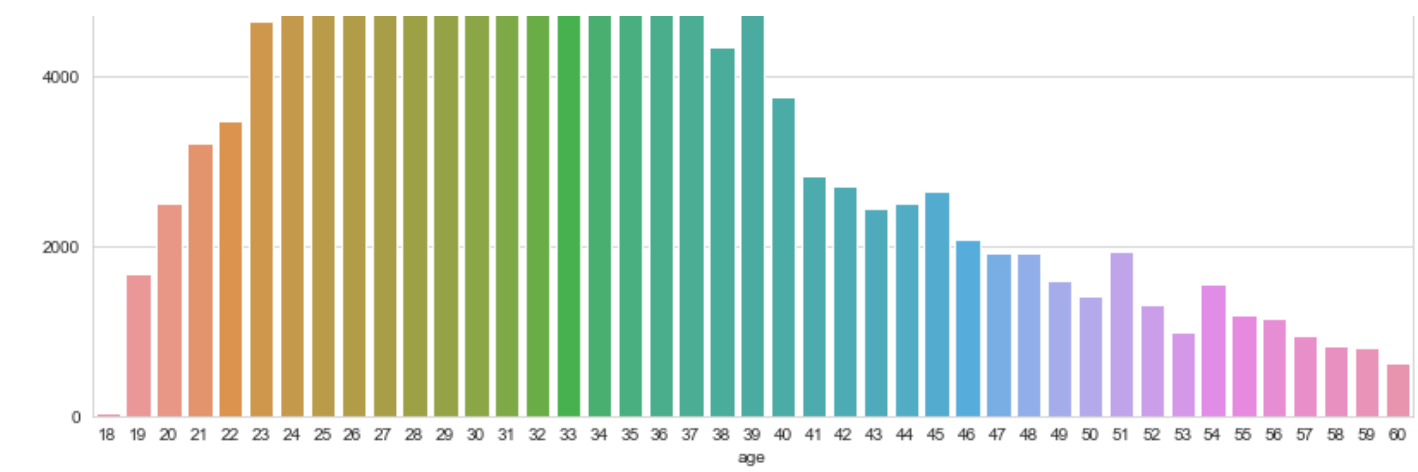
Rentals on Age, Gender & User Type

- **To rent a bike, the user needs to be at least 18 years old, 75% of the users are younger than 39 years old. The highest renting count (more than 10000 times) is from 31-year-old age group, The GoBike is much more popular among young people.**
- **Male users have used the GoBike the most frequently, and they recored ~130000 times of rides. In comparison, female users rided with GoBike for ~40000 times. A couple of thousands rides go to users with other genders.**
- **There are two types of users in the data set. From some research, the users can either become a "Subscriber" by paying monthly or yearly membership fees, or become a "Customer" by paying for single trip or for single day. It turns out that more than 89.2% of rental bike records are made by subscribers, 10.9% of rental bike records are made by customers.**

In [19]:

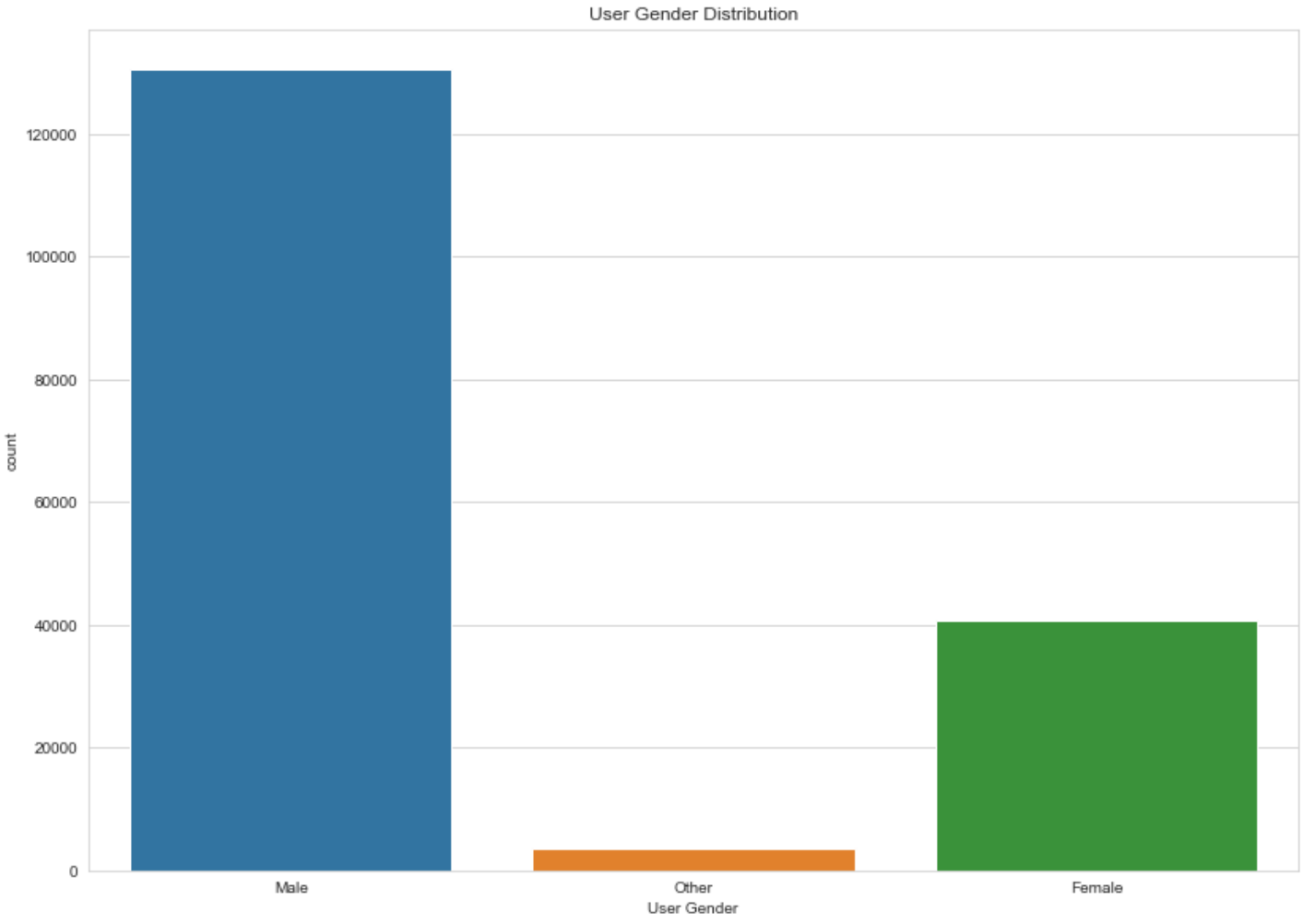
```
# user age distribution
sb.set_style("whitegrid")
fig, ax = plt.subplots(figsize = (11.69, 8.27))
ax = sb.countplot(x = "age", data = df_bike.query('age <= 60'))
ax.set_xticklabels(range(18, 61, 1))
ax.set_title('Number of Rides Based on User Age')
plt.tight_layout();
```





In [20]:

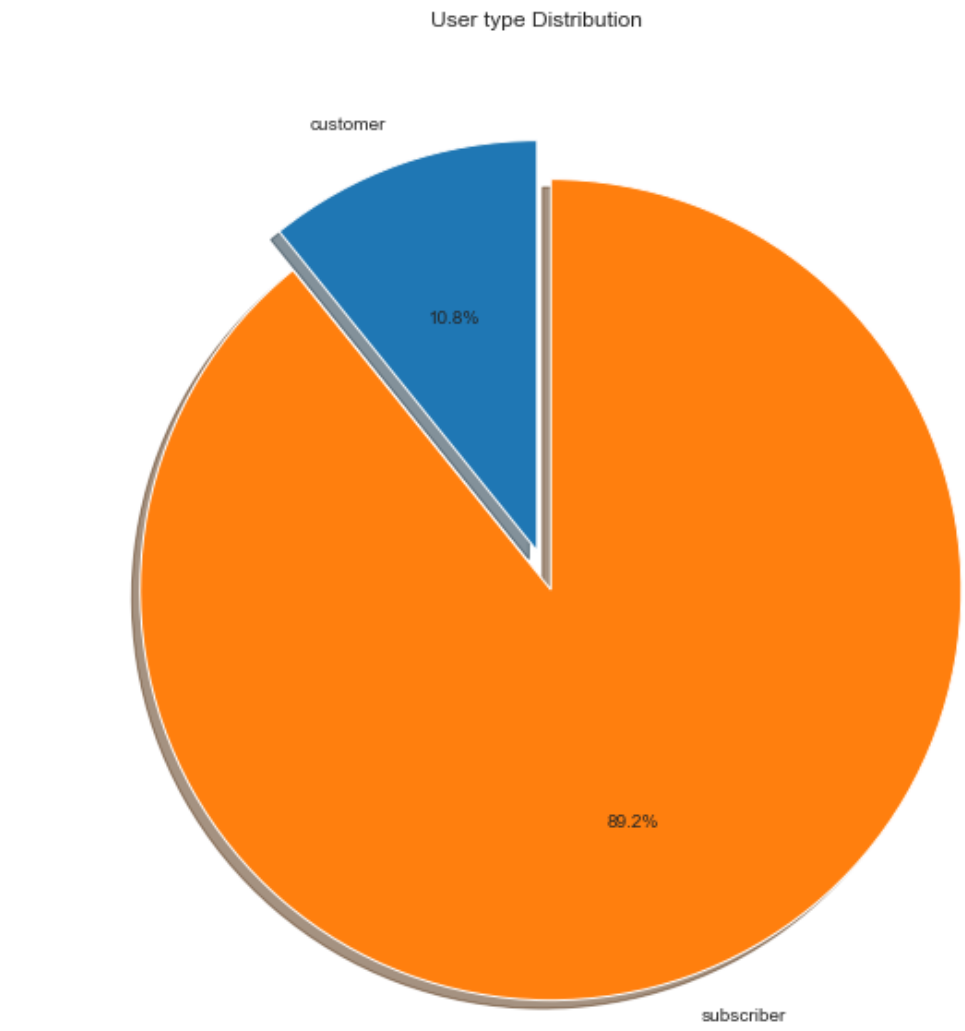
```
# plot ride counts on genders
sb.set_style("whitegrid")
fig, ax = plt.subplots(figsize = (11.69,8.27))
ax = sb.countplot(x = "member_gender", data = df_bike)
ax.set_title('User Gender Distribution')
ax.set_xlabel('User Gender')
plt.tight_layout();
```



In [21]:

```
# user type ratio
customer_ratio = df_bike.query('user_type == "Customer"').shape[0] / df_bike.shape[0]
subscriber_ratio = df_bike.query('user_type == "Subscriber"').shape[0] / df_bike.shape[0
]
# pie chart of user type
sb.set_style("whitegrid")
pie_data = [customer_ratio, subscriber_ratio]
fig, ax = plt.subplots(figsize = (11.69,8.27))
plt.pie(pie_data, explode=[0, 0.1], labels=['customer', 'subscriber'], startangle=90, au
topct='%1.1f%%', shadow=True)
```

```
plt.title('User type Distribution')
plt.tight_layout();
```

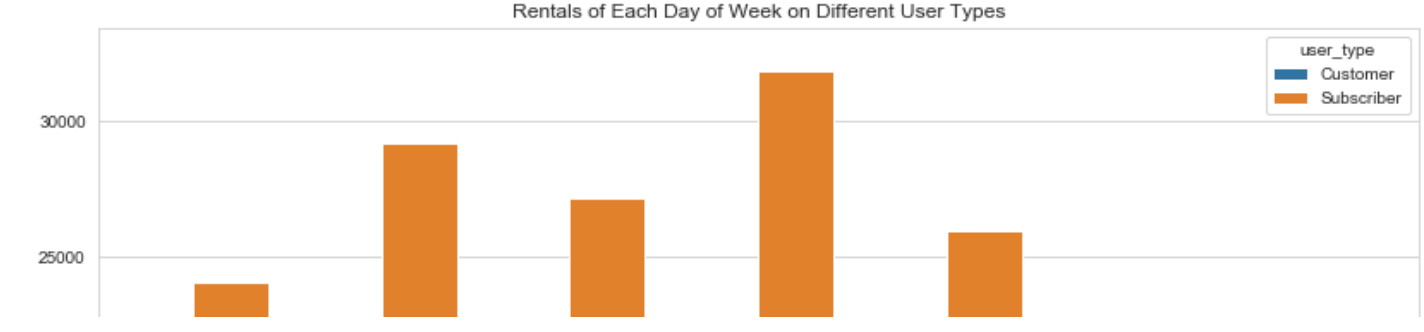


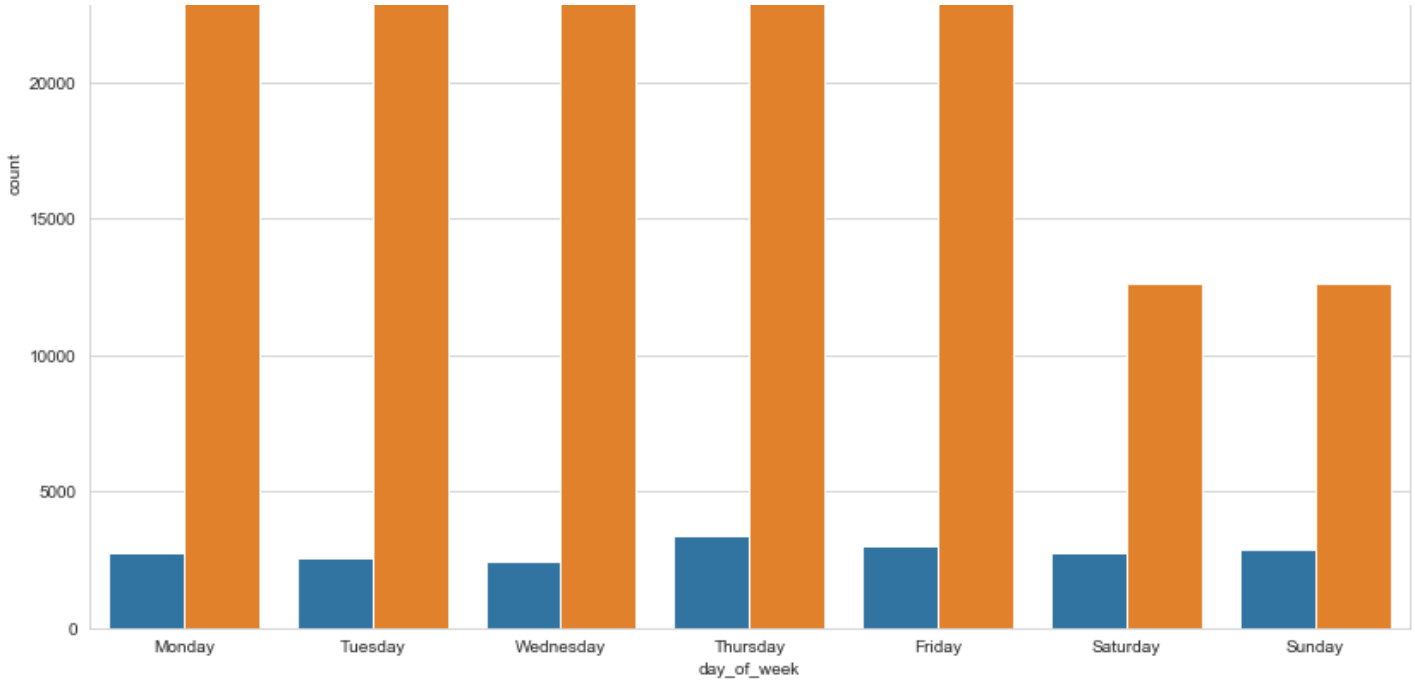
Distribution of rentals on day of week by user type

Most of the rentals are made by subscribers compared to customers on each day of the week. For subscribers, we see an interesting pattern that many more rentals were made during weekdays than weekends (>10000 more rentals in total on Weekdays), it's likely that they use GoBike for commuting. Meanwhile for customers, they have fewer (3000 to 4000 in total From Monday to Sunday) rental counts throughout the week than subscribers, and the numbers are pretty stable on each day, e.g., these rentals might partly come from tourists in the Bay area.

In [22]:

```
# rentals of each day of week on different user types
sb.set_style("whitegrid")
fig, ax = plt.subplots(figsize = (11.69,8.27))
ax = sb.countplot(x = "day_of_week", hue = 'user_type', data = df_bike,
                  order = ['Monday','Tuesday','Wednesday','Thursday','Friday','Saturday'
,'Sunday'])
ax.set_title('Rentals of Each Day of Week on Different User Types')
plt.tight_layout();
```





Average ride distance and duration by day of month

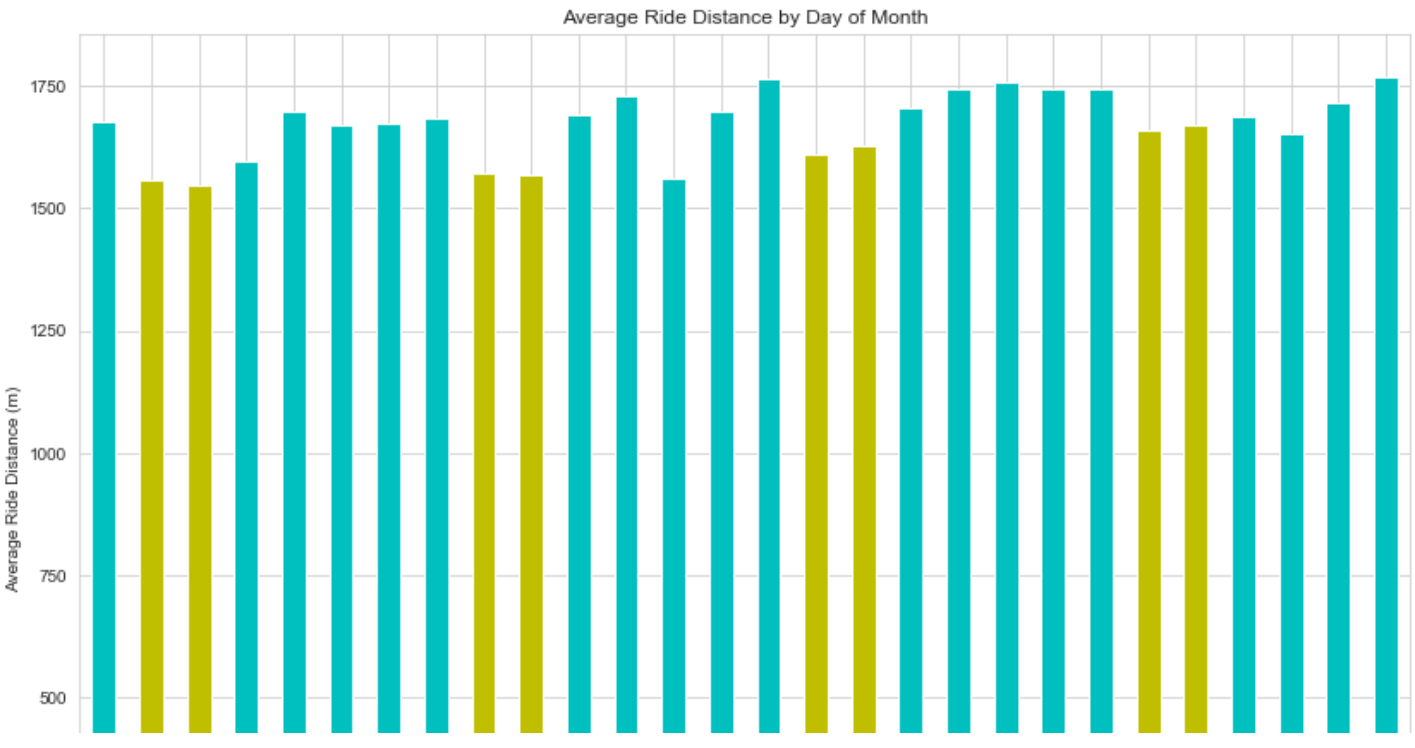
On average, the trip distance is longer on weekedays (averaged between 1.7 to 1.8 km) than weekends (averaged between 1.6 to 1.7 km). However, it's opposite for trip duration, on weekdays the users are going on shorter trips that takes 10 to 12 min, while the average duration on the weekend rises to 13 min to 18 min.

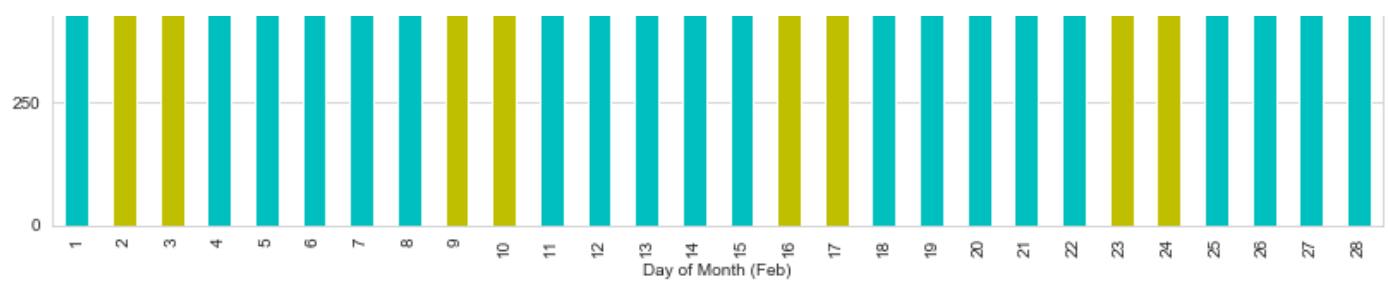
In [23]:

```
# plotting color: 'y' on weekends, 'c'on weekdays for Feb 2019
colors = ('c','y','y','c','c','c','c','c', 'y','y','c','c','c','c','c',
          'y','y','c','c','c','c','c', 'y','y','c','c','c','c')

# groupby data by day and find the average value
df_avg_distance = df_bike.groupby('day')['distance'].mean()

# visualize using a bar plot
sb.set_style("whitegrid")
df_avg_distance.plot.bar(figsize = (11.69, 8.27), color = colors)
plt.title('Average Ride Distance by Day of Month')
plt.xlabel('Day of Month (Feb)')
plt.ylabel('Average Ride Distance (m)')
plt.tight_layout();
```

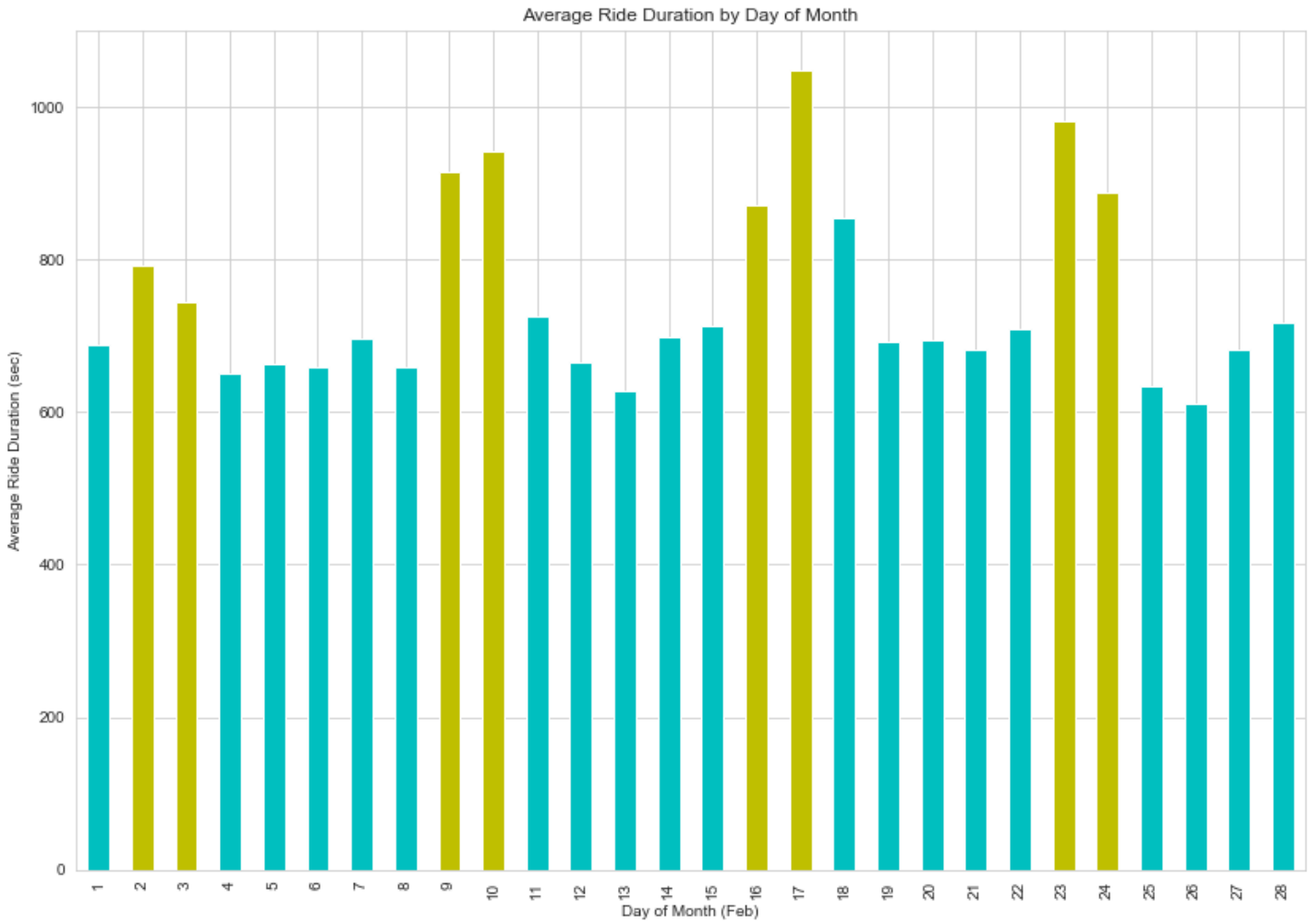




In [24]:

```
# groupby data by day and find the average value
df_avg_distance = df_bike.groupby('day')['duration_sec'].mean()

# visualize using a bar plot
sb.set_style("whitegrid")
df_avg_distance.plot.bar(figsize = (11.69, 8.27), color = colors)
plt.title('Average Ride Duration by Day of Month')
plt.xlabel('Day of Month (Feb)')
plt.ylabel('Average Ride Duration (sec)')
plt.tight_layout();
```



Number of rides during day of week per age group

During weekdays, the rentals come from people between 30-39 the most, followed by people between 20-29 who rented 2000-4000 times fewer. Whereas in the weekend, 20-29 group takes over and makes the most rentals, they only lead 30-39 group by a few hundred.

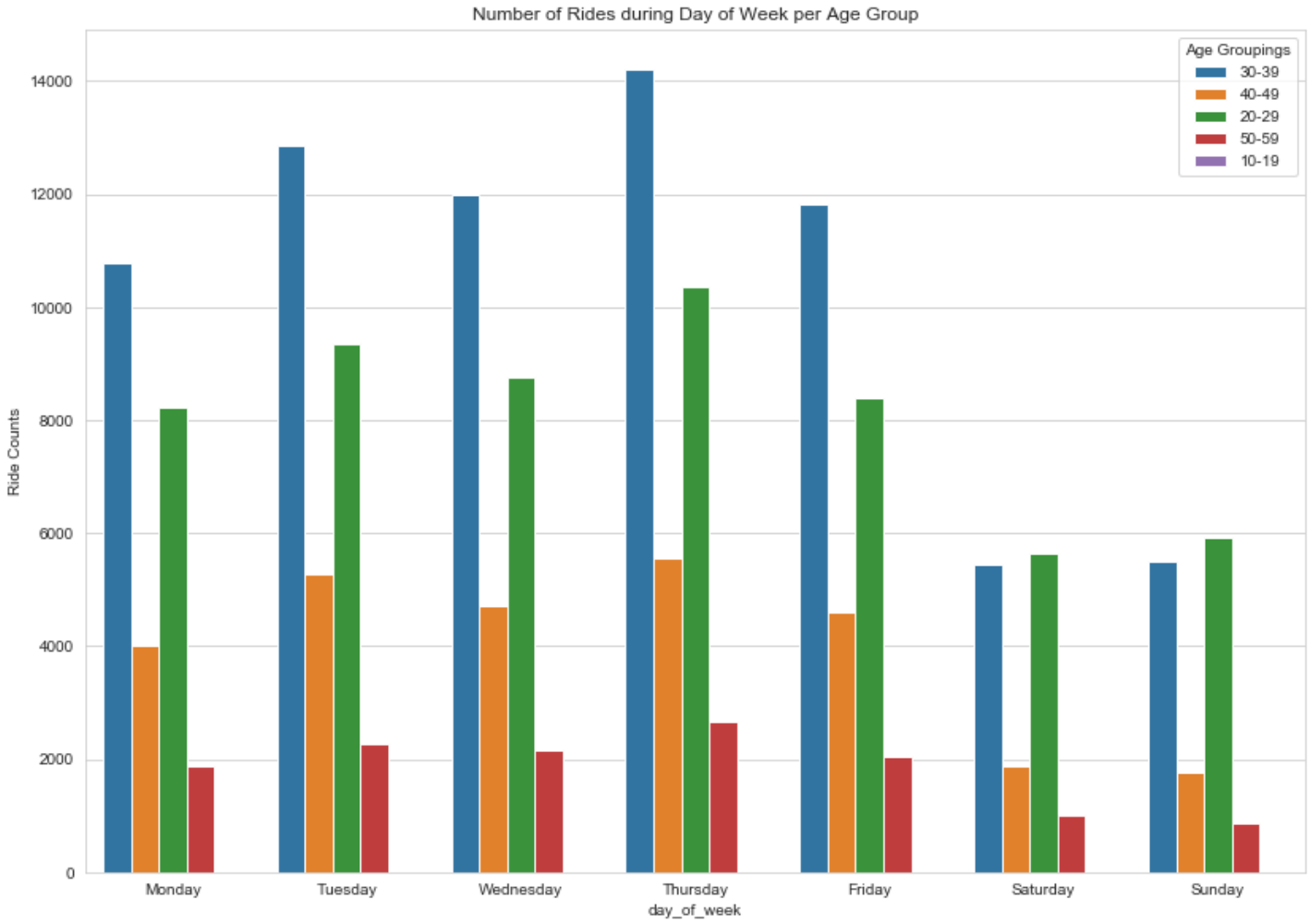
In [25]:

```
# rides counts on age group and day of week, save in df_age_dow
df_age_dow = df_bike.loc[df_bike['age_group'].isin(['10-19', '20-29', '30-39', '40-49', '
50-59'])]

# visualization
```

```
plt.figure(figsize = (11.69, 8.27))
sb.set_style("whitegrid")
ax = sb.countplot(data = df_age_dow, x = 'day_of_week', hue = 'age_group',
                  order=['Monday','Tuesday','Wednesday','Thursday','Friday','Saturday','
Sunday'])

plt.title('Number of Rides during Day of Week per Age Group')
plt.legend(title = 'Age Groupings', loc = 'upper right')
plt.ylabel('Ride Counts')
plt.tight_layout();
```



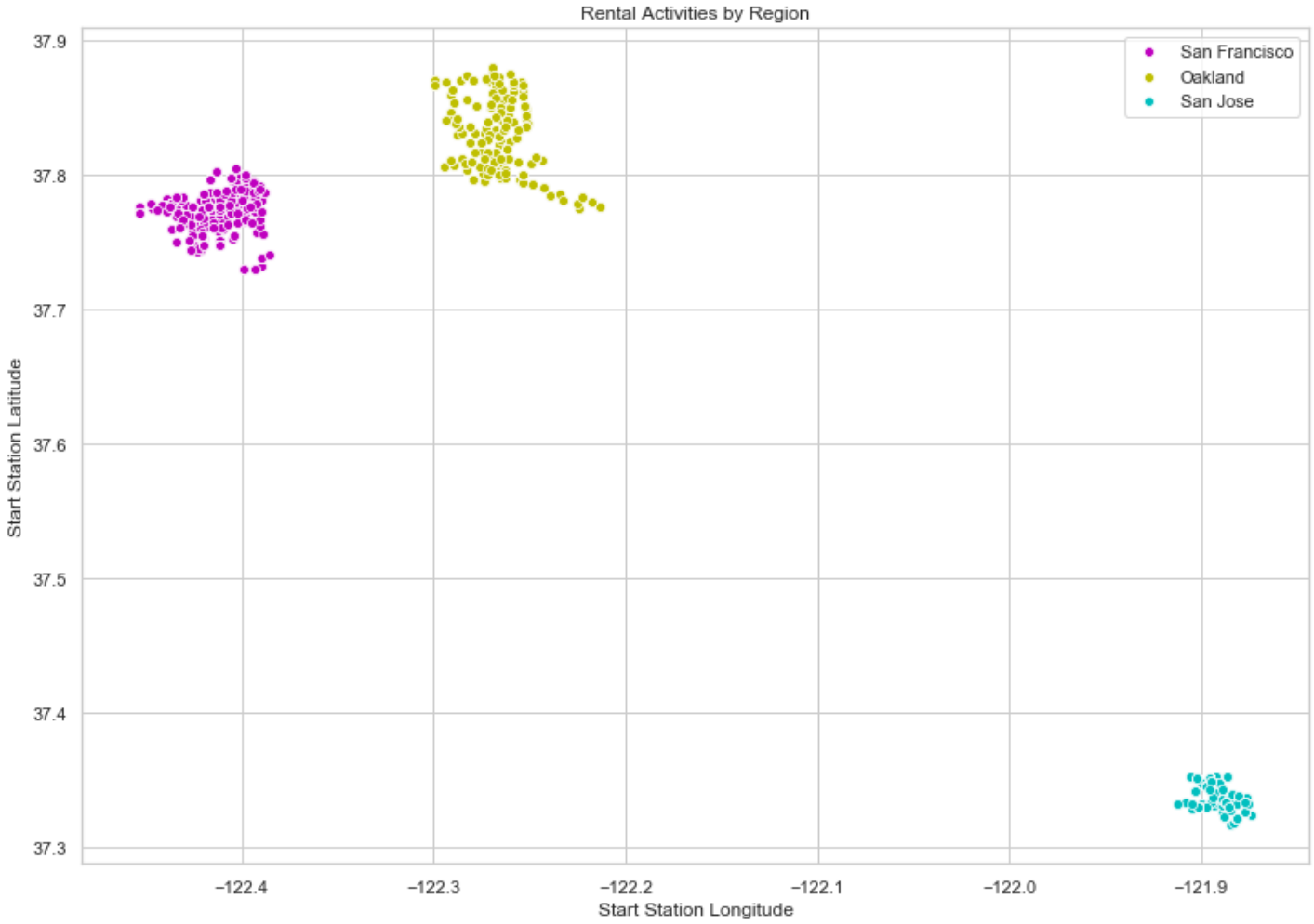
Regional Average Ride Duration and Distance on each Hour of Day

- **We have 133708 rentals from San Francisco area, 41434 rentals from Oakland and 8073 rentals from San Jose.**
- **Compared between the regions of San Francisco, Oakland and San Jose, on average the ride duration over hour in each region follows the similar pattern and in contrast to the ride distance, peak is located at noon. San Francisco riders took the longest trip (over 10 min), Oakland, San Jose are the second (~9.2 min) and the third (~8.3 min) (again we expect this because SF is a larger city with more populations).**
- **Compared between the regions of San Francisco, Oakland and San Jose, we can see that on average the ride distance over hour in each region follows similar rush-hour pattern as peaks are observed at ~8 am and ~5 pm. San Francisco riders took the longest trip, averaged over 1.5 km, Oakland, San Jose are the second (~1.4 km) and the third (~1.2 km) (we expect this because SF is a larger city with more populations).**

In [26]:

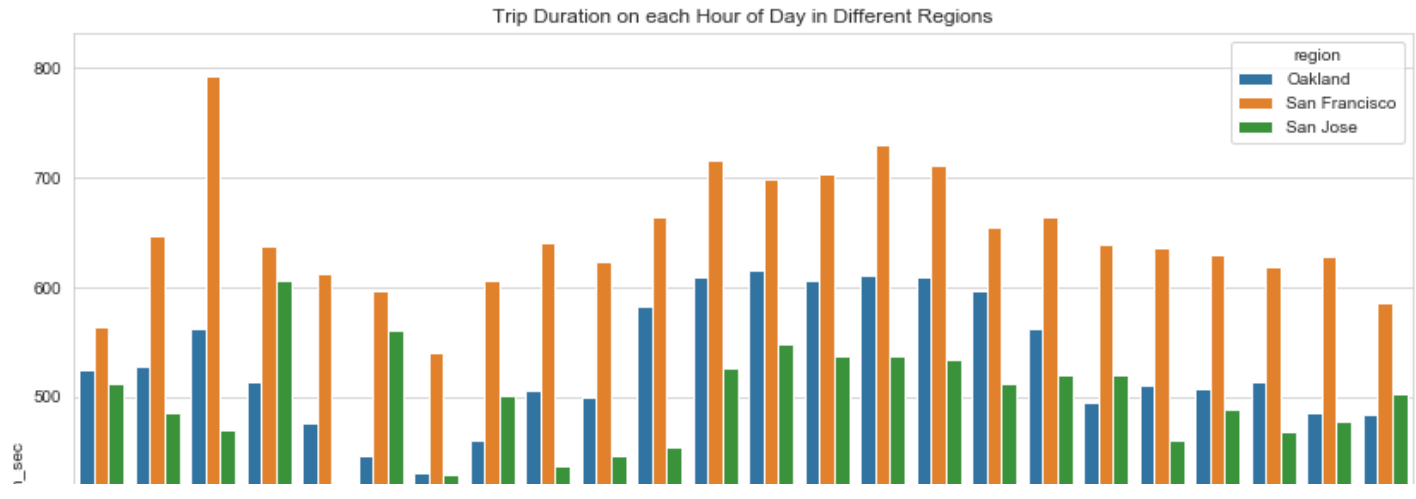
```
# visualize different regions
sb.set_style("whitegrid")
with sb.plotting_context("notebook"):
    # create a matplotlib figure
    fig, ax = plt.subplots(figsize = (11.69,8.27))
    sb.scatterplot(y = "start_station_latitude", x = "start_station_longitude", data = d
f_bike.query("region == 'San Francisco'"),
                   label = 'San Francisco', color = 'm',)
```

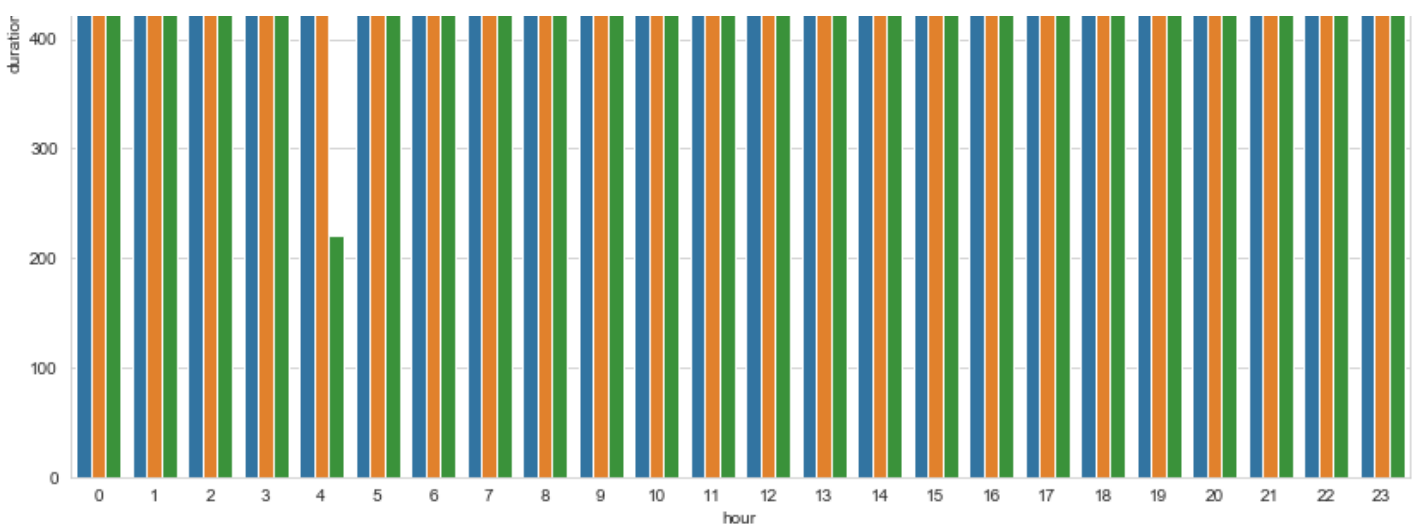
```
    sb.scatterplot(y = "start_station_latitude", x = "start_station_longitude", data = df_bike.query("region == 'Oakland'"),
                   label = 'Oakland', color = 'y')
    sb.scatterplot(y = "start_station_latitude", x = "start_station_longitude", data = df_bike.query("region == 'San Jose'"),
                   label = 'San Jose', color = 'c')
    plt.title('Rental Activities by Region')
    plt.ylabel('Start Station Latitude')
    plt.xlabel('Start Station Longitude')
    plt.legend()
    plt.tight_layout();
```



In [27]:

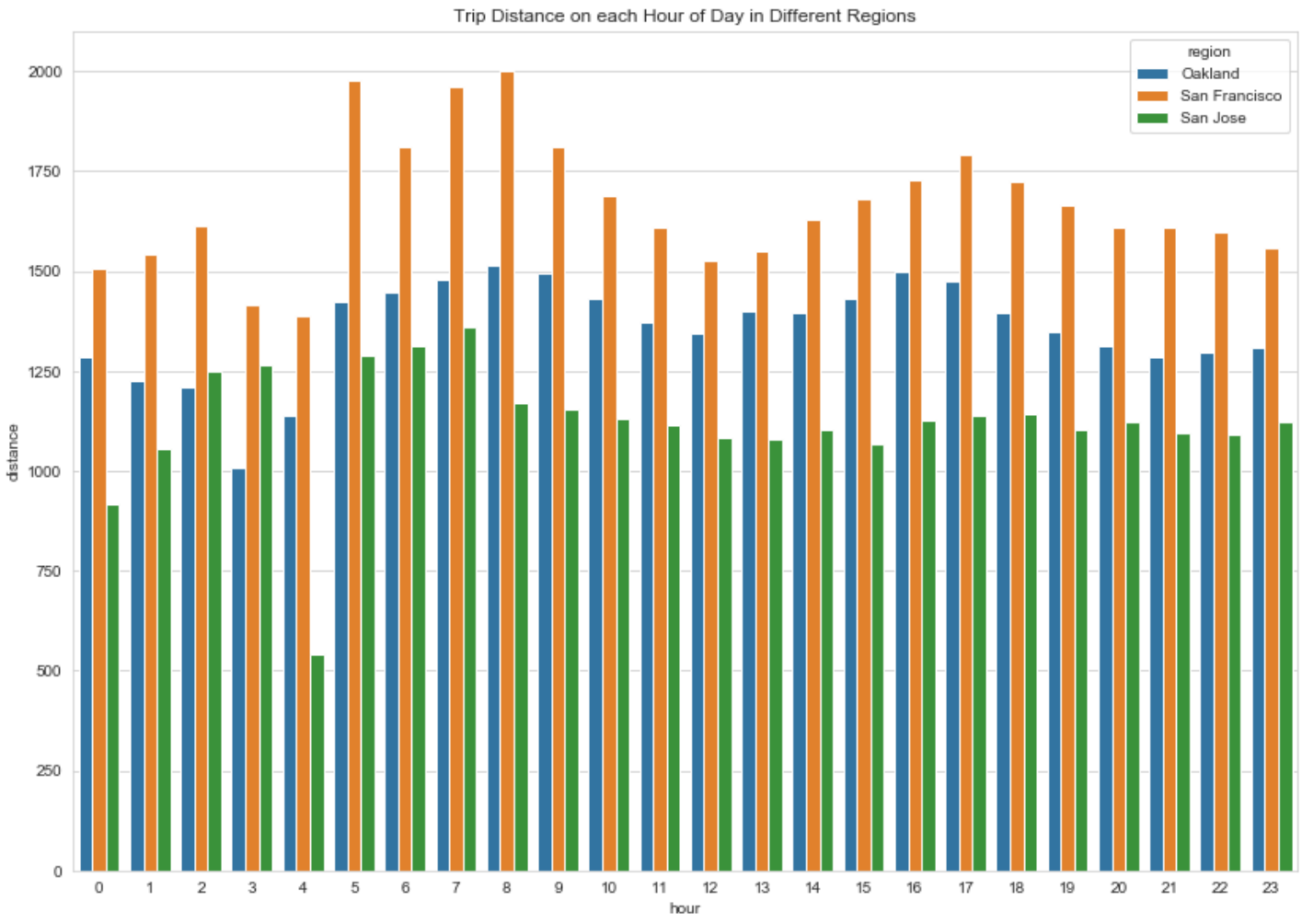
```
# plot trip duration on each hour of day in different regions
# cut off at .99 percentile, 3457 sec
sb.set_style("whitegrid")
fig, ax = plt.subplots(figsize = (11.69,8.27))
sb.barplot(x = 'hour', y = 'duration_sec', data = df_bike.query('duration_sec < 3457').groupby(['hour', 'region'], as_index = False).mean(), hue = 'region')
plt.title('Trip Duration on each Hour of Day in Different Regions')
plt.tight_layout();
```





In [28]:

```
# plot trip distance on each hour of day in different regions
# cut off at .99 percentile, 5070 meter
sb.set_style("whitegrid")
fig, ax = plt.subplots(figsize = (11.69,8.27))
sb.barplot(x = 'hour', y = 'distance', data = df_bike.query('distance < 5070').groupby([
'hour', 'region'], as_index = False).mean(), hue = 'region')
plt.title('Trip Distance on each Hour of Day in Different Regions')
plt.tight_layout();
```



In [ ]:

```
!jupyter nbconvert "slide_deck.ipynb" --to slides --post serve --template output_toggle
```

In [ ]: