Exploratory Data Analysis:

```
There are 983648 data points in the dataset.
The average duration of trips is 16.98 minutes.
The median trip duration is 8.63 minutes.
25% of trips are shorter than 5.78 minutes.
25% of trips are longer than 12.47 minutes.
```

1. What are the most popular start and end station pairings?

```
 start_station_name                        end_station_name
code_count
 Harry Bridges Plaza (Ferry Building)     Embarcadero at Sansome
9150
 San Francisco Caltrain 2 (330 Townsend)  Townsend at 7th
8508
 2nd at Townsend                          Harry Bridges Plaza (Fer
ry Building)          7620
 Harry Bridges Plaza (Ferry Building)     2nd at Townsend
6888
 Embarcadero at Sansome                   Steuart at Market
6874
```

2. Which bikes have been ridden the most?

```
 bike_number       count
 389               2872
 392               2853
 524               2853
 503               2845
 328               2827
```

3. How long was the average ride length each year?

   (a): YEARLY TRIPS:

```
year    yearly_trips
2013    100563
2014    326339
2015    346252
2016    210494
```

   (b): AVERAGE RIDE LENGTH PER YEAR:

```
year    avg_duration_min
2013    21.971096
2014    18.866123
2015    15.682313
2016    13.816306
```

4. How many bikes were there yearly?

```
year    bike_count
2013    689
2014    687
2015    655
2016    608
```

1. Who uses the bike sharing?

   Subscribers: People with subscriptions frequently use the bike sharing.((Graph is provided below))

1. What day is the bike share used?

   on Analysis, we got to know that on weekdays, sharing is mostly used as compared to weekends. (Graph is provided below)
2. At what time, the sharing is used?

   During the daytime, mostly used by the subscriber.

CONCLUSION:

```
1. Most of the trips are when trip duration is 10-15 min. Hance, People m
ostly use it for short distances.
2. Bikes must be provided nearby offices, stations, eatery places and tou
rist places in order to increase the use. It will
   increase customers along with the subscribers.
3. The bike count is also decreased continuously over the period of four
 years, the reason could be theft or damaged bikes,
   which has dereased the average duration of riding.
```

Import the Libraries

In [10]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
```

In [11]:

```
import matplotlib
matplotlib.style.use('ggplot')
```

Import the dataset

In [12]:

```
trip_data = pd.read_csv('./bikeshare_trips.csv')
```

The first step is to look at the structure of the data and printing out the first few rows of the data.

In [13]:

```
trip_data.head(3)
```

Out[13]:

	trip_id	duration_sec	start_date	start_station_name	start_station_id	end_date	e
0	944732	2618	2015-09-24 17:22:00 UTC	Mezes	83	2015-09-24 18:06:00 UTC	N
1	984595	5957	2015-10-25 18:12:00 UTC	Mezes	83	2015-10-25 19:51:00 UTC	N
2	984596	5913	2015-10-25 18:13:00 UTC	Mezes	83	2015-10-25 19:51:00 UTC	N

In [14]:

```
trip_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 983648 entries, 0 to 983647
Data columns (total 11 columns):
trip_id               983648 non-null int64
duration_sec          983648 non-null int64
start_date            983648 non-null object
start_station_name    983648 non-null object
start_station_id      983648 non-null int64
end_date              983648 non-null object
end_station_name      983648 non-null object
end_station_id        983648 non-null int64
bike_number           983648 non-null int64
zip_code              976838 non-null object
subscriber_type       983648 non-null object
dtypes: int64(5), object(6)
memory usage: 82.6+ MB
```

Convert the duration_sec to duration(min) in order to hav the better understand for ow much time is the sharing is done.

In [15]:

```
#Convert the durantion_sec from seconds to minutes and rename the name of column

trip_data['duration_sec'] = trip_data['duration_sec'] /60
trip_data = trip_data.rename(index=str, columns={"duration_sec": "duration"})
```

In [16]:

```
trip_data.head(1)
```

Out[16]:

	trip_id	duration	start_date	start_station_name	start_station_id	end_date	end_
0	944732	43.633333	2015-09-24 17:22:00 UTC	Mezes	83	2015-09-24 18:06:00 UTC	Mez

DataSet Summary

In [17]:

```
n = trip_data.shape[0]
duration_mean = trip_data['duration'].mean()
duration_qtiles = trip_data['duration'].quantile([.25, .5, .75]).as_matrix()
```

```
C:\ProgramData\Anaconda3\lib\site-packages\ipykernel_launcher.py:3:
FutureWarning: Method .as_matrix will be removed in a future versio
n. Use .values instead.
  This is separate from the ipykernel package so we can avoid doing
imports until
```

In [9]:

```
print('There are {:d} data points in the dataset.'.format(n))
print('The average duration of trips is {:.2f} minutes.'.format(duration_mean))
print('The median trip duration is {:.2f} minutes.'.format(duration_qtiles[1]))
print('25% of trips are shorter than {:.2f} minutes.'.format(duration_qtiles[0
]))
print('25% of trips are longer than {:.2f} minutes.'.format(duration_qtiles[2]))
```

```
There are 983648 data points in the dataset.
The average duration of trips is 16.98 minutes.
The median trip duration is 8.63 minutes.
25% of trips are shorter than 5.78 minutes.
25% of trips are longer than 12.47 minutes.
```

Most Popular Start-End Station Pairs:

In [44]:

```
df_top_freq = trip_data.groupby(['start_station_name', 'end_station_name'])['sub
scriber_type'].agg(
    {"code_count": len}).sort_values(
    "code_count", ascending=False).head(5).reset_index()
```

```
C:\ProgramData\Anaconda3\lib\site-packages\ipykernel_launcher.py:2:
FutureWarning: using a dict on a Series for aggregation
is deprecated and will be removed in a future version
```

In [45]:

```
df_top_freq
```

Out[45]:

	start_station_name	end_station_name	code_count
0	Harry Bridges Plaza (Ferry Building)	Embarcadero at Sansome	9150
1	San Francisco Caltrain 2 (330 Townsend)	Townsend at 7th	8508
2	2nd at Townsend	Harry Bridges Plaza (Ferry Building)	7620
3	Harry Bridges Plaza (Ferry Building)	2nd at Townsend	6888
4	Embarcadero at Sansome	Steuart at Market	6874

Which Bike have ridden the most

In [46]:

```
trip_data['bike_number'].mode()
```

Out[46]:

```
0    389
dtype: int64
```

In [47]:

```
df_top_bike = trip_data.groupby(['bike_number'])['subscriber_type'].agg(
    {"code_count": len}).sort_values(
    "code_count", ascending=False).head(5).reset_index()
```

```
C:\ProgramData\Anaconda3\lib\site-packages\ipykernel_launcher.py:2:
FutureWarning: using a dict on a Series for aggregation
is deprecated and will be removed in a future version
```

In [48]:

```
df_top_bike
```

Out[48]:

	bike_number	code_count
0	389	2872
1	392	2853
2	524	2853
3	503	2845
4	328	2827

How long was the average ride length each year?

In [49]:

```
#Exract Year from date
trip_data['year'] = trip_data.start_date.str.extract(r'([0-9][0-9][0-9][0-9])',
expand=True)
```

In [50]:

```
yearly_trip =trip_data.pivot_table(index=['year'],values=["trip_id"],aggfunc=[le
n]).reset_index()
```

In [51]:

```
yearly_trip.columns =['year','yearly_trips']
```

In [52]:

```
yearly_trip
```

Out[52]:

	year	yearly_trips
0	2013	100563
1	2014	326339
2	2015	346252
3	2016	210494

In [53]:

```
duration_table =trip_data.pivot_table(index=['year'],values=["duration"],aggfunc
=[np.mean]).reset_index()
```

In [54]:

```
duration_table.columns = ['year','avg_duration_min']
```

In [55]:

```
duration_table
```

Out[55]:

	year	avg_duration_min
0	2013	21.971096
1	2014	18.866123
2	2015	15.682313
3	2016	13.816306

In [56]:

```
result = pd.merge(duration_table,
                  yearly_trip[['year','yearly_trips']],
                  on='year').sort_values('yearly_trips',ascending=False)
```

In [57]:

```
result
```

Out[57]:

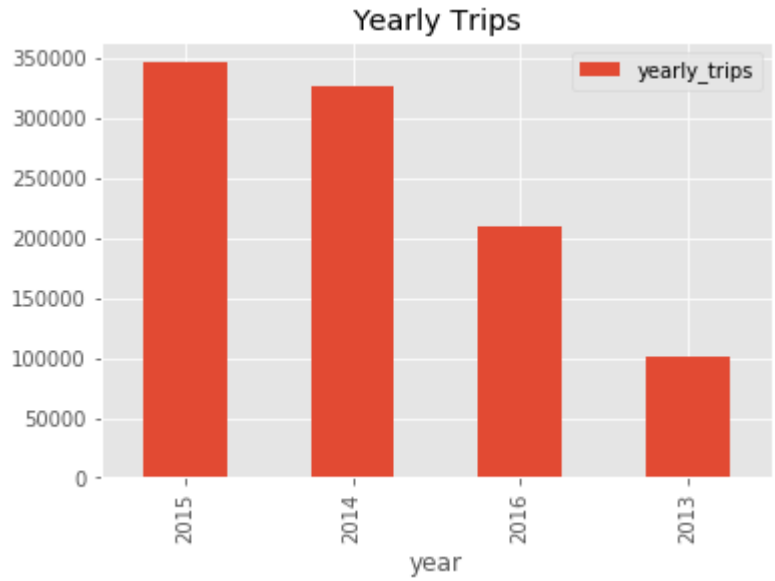
	year	avg_duration_min	yearly_trips
2	2015	15.682313	346252
1	2014	18.866123	326339
3	2016	13.816306	210494
0	2013	21.971096	100563

In [58]:

```
result.plot(x='year',y=['yearly_trips'],kind="bar")
plt.title("Yearly Trips")
```

Out[58]:

```
Text(0.5,1,'Yearly Trips')
```

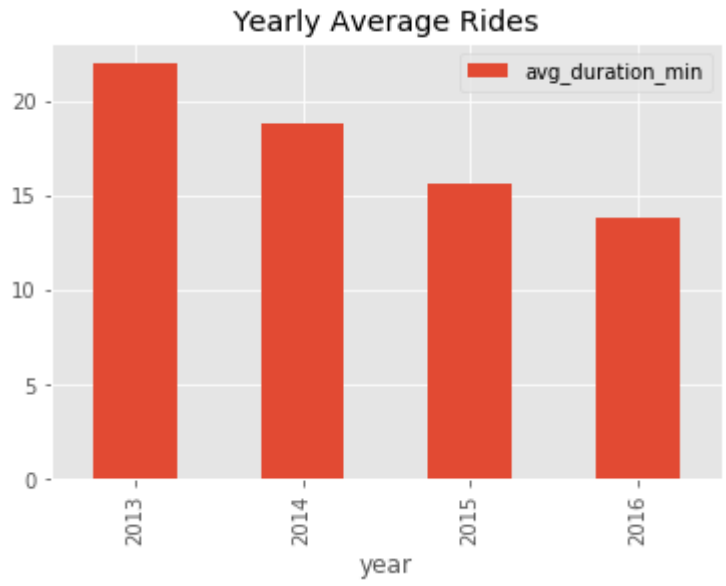


In [59]:

```
result.sort_values('avg_duration_min',ascending=False).plot(x='year',y=['avg_dur
ation_min'],kind="bar")
plt.title("Yearly Average Rides")
```

Out[59]:

```
Text(0.5,1,'Yearly Average Rides')
```



How many bikes were there yearly ?

In [60]:

```
func = lambda x: x.nunique()
number_of_bikes =trip_data.pivot_table(index=['year'],values=["bike_number"],agg
func=func).reset_index()
number_of_bikes.sort_values('bike_number',ascending=False)
```

Out[60]:

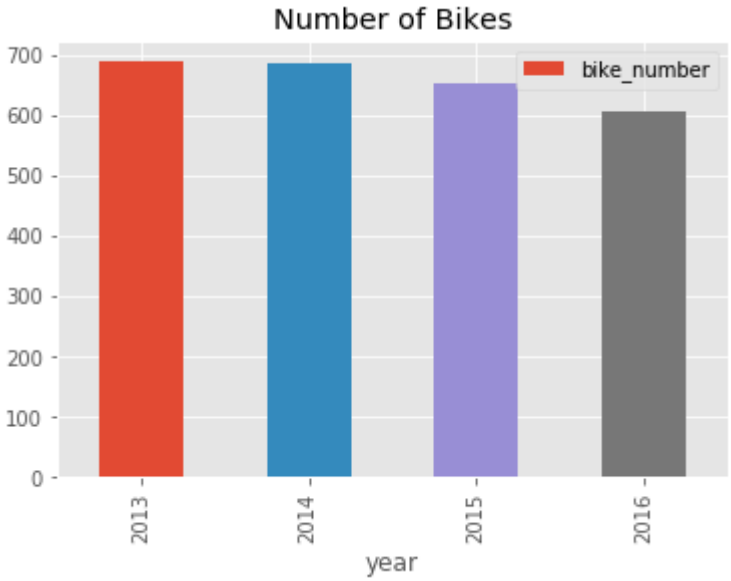
	year	bike_number
0	2013	689
1	2014	687
2	2015	655
3	2016	608

In [61]:

```
number_of_bikes.plot(x='year',y='bike_number',kind='bar')
plt.title("Number of Bikes")
```

Out[61]:

Text(0.5,1,'Number of Bikes')



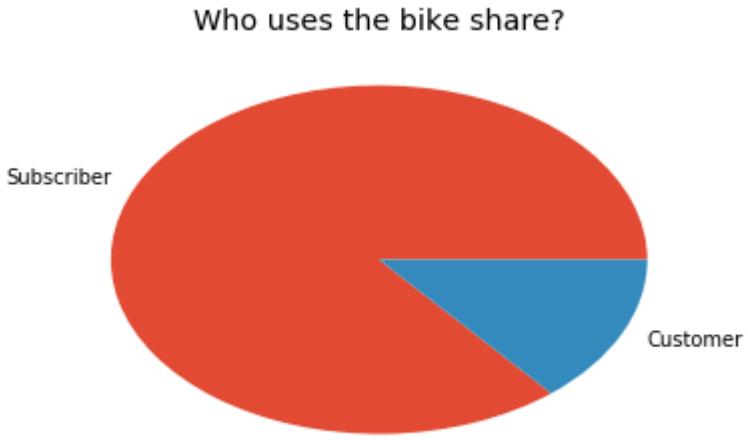
Who uses bike the most?

In [67]:

```
trip_data['subscriber_type'].value_counts().plot(kind='pie')
plt.ylabel('')
plt.title("Who uses the bike share?")
```

Out[67]:

Text(0.5,1,'Who uses the bike share?')



What day is the bike share used?

In [63]:

```
#Exract Year, date, time from start_date column
trip_data['year'] = trip_data.start_date.str.extract(r'([0-9][0-9][0-9][0-9])',
expand=True)
trip_data['date'] =  trip_data.start_date.str.extract(r'([0-9][0-9][0-9][0-9]-[0
-9][0-9]-[0-9][0-9])', expand=True)
trip_data['time'] = trip_data.start_date.str.extract(r'([0-9][0-9]:[0-9][0-9]:[0
-9][0-9])', expand=True)
```

In [64]:

```
# Get the day of week from date column

trip_data['date'] = pd.to_datetime(trip_data['date'])
trip_data['day_of_week'] = trip_data['date'].dt.dayofweek

days = {0:'Mon',1:'Tues',2:'Weds',3:'Thurs',4:'Fri',5:'Sat',6:'Sun'}

trip_data['day_of_week'] = trip_data['day_of_week'].apply(lambda x: days[x])
```

In [65]:

```
func = lambda x: x.nunique()
frequent_day =trip_data.pivot_table(index=['day_of_week'],values=["trip_id"],agg
func=func).reset_index()
frequent_day
```

Out[65]:

	day_of_week	trip_id
0	Fri	159977
1	Mon	169937
2	Sat	60279
3	Sun	51375
4	Thurs	176908
5	Tues	184405
6	Weds	180767

In [66]:

```
frequent_day.sort_values('trip_id',ascending=False).plot(x='day_of_week',y='trip
_id',kind='bar')
plt.title("What day is the bike share used?")
```

Out[66]:

```
Text(0.5,1,'What day is the bike share used?')
```