2016 US Bike Share Activity Snapshot

Table of Contents

Introduction

> **Tip**: Quoted sections like this will provide helpful instructions on how to navigate and use a Jupyter notebook.

Over the past decade, bicycle-sharing systems have been growing in number and popularity in cities across the world. Bicycle-sharing systems allow users to rent bicycles for short trips, typically 30 minutes or less. Thanks to the rise in information technologies, it is easy for a user of the system to access a dock within the system to unlock or return bicycles. These technologies also provide a wealth of data that can be used to explore how these bike-sharing systems are used.

In this project, you will perform an exploratory analysis on data provided by Motivate, a bike-share system provider for many major cities in the United States. You will compare the system usage between three large cities: New York City, Chicago, and Washington, DC. You will also see if there are any differences within each system for those users that are registered, regular users and those users that are short-term, casual users.

Posing Questions

Before looking at the bike sharing data, you should start by asking questions you might want to understand about the bike share data. Consider, for example, if you were working for Motivate. What kinds of information would you want to know about in order to make smarter business decisions? If you were a user of the bike-share service, what factors might influence how you would want to use the service?

Question 1: Write at least two questions related to bike sharing that you think could be answered by data.

Answer:

- **Q1: which are the routes that are used most offeten at what point of the day?**
- **Q2: what is the average cycling distance covered by the usesrs during their trip?**

- **Q3: which stations and roots require larger no of cycles to be deployed?**
- **Q4: based on the users routes what are the best locations to establish new stations to increase the efficieny of the system?**

> **Tip**: If you double click on this cell, you will see the text change so that all of the formatting is removed. This allows you to edit this block of text. This block of text is written using [Markdown](), which is a way to format text using headers, links, italics, and many other options using a plain-text syntax. You will also use Markdown later in the Nanodegree program. Use **Shift** + **Enter** or **Shift** + **Return** to run the cell and show its rendered form.

Data Collection and Wrangling

Now it's time to collect and explore our data. In this project, we will focus on the record of individual trips taken in 2016 from our selected cities: New York City, Chicago, and Washington, DC. Each of these cities has a page where we can freely download the trip data.:

- New York City (Citi Bike): [Link]()
- Chicago (Divvy): [Link]()
- Washington, DC (Capital Bikeshare): [Link]()

If you visit these pages, you will notice that each city has a different way of delivering its data. Chicago updates with new data twice a year, Washington DC is quarterly, and New York City is monthly. **However, you do not need to download the data yourself.** The data has already been collected for you in the `/data/` folder of the project files. While the original data for 2016 is spread among multiple files for each city, the files in the `/data/` folder collect all of the trip data for the year into one file per city. Some data wrangling of inconsistencies in timestamp format within each city has already been performed for you. In addition, a random 2% sample of the original data is taken to make the exploration more manageable.

Question 2: However, there is still a lot of data for us to investigate, so it's a good idea to start off by looking at one entry from each of the cities we're going to analyze. Run the first code cell below to load some packages and functions that you'll be using in your analysis. Then, complete the second code cell to print out the first trip recorded from each of the cities (the second line of each data file).

> **Tip**: You can run a code cell like you formatted Markdown cells above by clicking on the cell and using the keyboard shortcut **Shift** + **Enter** or **Shift** + **Return**. Alternatively, a code cell can be executed using the **Play** button in the toolbar after selecting it. While the cell is running, you will see an asterisk in the message to the left of the cell, i.e. `In [*]:`. The asterisk will change into a number to show that execution has completed, e.g. `In [1]`. If there is output, it will show up as `Out [1]:`, with an appropriate number to match the "In" number.

`In [1]:`

```
## import all necessary packages and functions.
import csv # read and write csv files
from datetime import datetime # operations to parse dates
from pprint import pprint # use to print data structures like dictionaries
```

In [2]:

```python
def print_first_point(filename):
    """
    This function prints and returns the first data point (second row) from
    a csv file that includes a header row.
    """
    # print city name for reference
    city = filename.split('-')[0].split('/')[-1]
    print('\nCity: {}'.format(city))

    with open(filename, 'rt') as f_in:
        ## TODO: Use the csv library to set up a DictReader object. ##
        ## see https://docs.python.org/3/library/csv.html          ##


        trip_reader = csv.DictReader(f_in,skipinitialspace = True) ## Dict
reader object created


        ## TODO: Use a function on the DictReader object to read the    ##
        ## first trip from the data file and store it in a variable.    ##
        ## see https://docs.python.org/3/library/csv.html#reader-objects ##

        ## Creating a first trip empty dictionary to hold the value of the
first row of each data set
        first_trip = {}
        for data in trip_reader:
            first_trip = data                    ## Storing the value in fetch
d from the file to the local dictionary
            if(trip_reader.line_num == 2):    ## Boolean condition check to
print only the first records of each file
                break

        ## TODO: Use the pprint library to print the first trip. ##
        ## see https://docs.python.org/3/library/pprint.html     ##


        pprint(first_trip)                        ## Using the pprint library f
r printng the result back


    # output city name and first trip for later testing


    return (city, first_trip)                     ## returning the city name an
the first-trip dictionary contating the data

# list of files for each city
data_files = ['./data/NYC-CitiBike-2016.csv',
              './data/Chicago-Divvy-2016.csv',
              './data/Washington-CapitalBikeshare-2016.csv',]

# print the first trip from each file, store in dictionary
example_trips = {}
for data_file in data_files:
    city, first_trip = print_first_point(data_file)
```

```
        example_trips[city] = first_trip        ## Storing the information
in another dictionary called example_trips
```

```
City: NYC
OrderedDict([('tripduration', '839'),
            ('starttime', '1/1/2016 00:09:55'),
            ('stoptime', '1/1/2016 00:23:54'),
            ('start station id', '532'),
            ('start station name', 'S 5 Pl & S 4 St'),
            ('start station latitude', '40.710451'),
            ('start station longitude', '-73.960876'),
            ('end station id', '401'),
            ('end station name', 'Allen St & Rivington St'),
            ('end station latitude', '40.72019576'),
            ('end station longitude', '-73.98997825'),
            ('bikeid', '17109'),
            ('usertype', 'Customer'),
            ('birth year', ''),
            ('gender', '0')])

City: Chicago
OrderedDict([('trip_id', '9080545'),
            ('starttime', '3/31/2016 23:30'),
            ('stoptime', '3/31/2016 23:46'),
            ('bikeid', '2295'),
            ('tripduration', '926'),
            ('from_station_id', '156'),
            ('from_station_name', 'Clark St & Wellington Ave'),
            ('to_station_id', '166'),
            ('to_station_name', 'Ashland Ave & Wrightwood Ave'),
            ('usertype', 'Subscriber'),
            ('gender', 'Male'),
            ('birthyear', '1990')])

City: Washington
OrderedDict([('Duration (ms)', '427387'),
            ('Start date', '3/31/2016 22:57'),
            ('End date', '3/31/2016 23:04'),
            ('Start station number', '31602'),
            ('Start station', 'Park Rd & Holmead Pl NW'),
            ('End station number', '31207'),
            ('End station', 'Georgia Ave and Fairmont St NW'),
            ('Bike number', 'W20842'),
            ('Member Type', 'Registered')])
```

If everything has been filled out correctly, you should see below the printout of each city name (which has been parsed from the data file name) that the first trip has been parsed in the form of a dictionary. When you set up a `DictReader` object, the first row of the data file is normally interpreted as column names. Every other row in the data file will use those column names as keys, as a dictionary is generated for each row.

This will be useful since we can refer to quantities by an easily-understandable label instead of just a numeric index. For example, if we have a trip stored in the variable `row`, then we would rather get the trip duration from `row['duration']` instead of `row[0]`.

Condensing the Trip Data

It should also be observable from the above printout that each city provides different information. Even where the information is the same, the column names and formats are sometimes different. To make things as simple as possible when we get to the actual exploration, we should trim and clean the data. Cleaning the data makes sure that the data formats across the cities are consistent, while trimming focuses only on the parts of the data we are most interested in to make the exploration easier to work with.

You will generate new data files with five values of interest for each trip: trip duration, starting month, starting hour, day of the week, and user type. Each of these may require additional wrangling depending on the city:

- **Duration**: This has been given to us in seconds (New York, Chicago) or milliseconds (Washington). A more natural unit of analysis will be if all the trip durations are given in terms of minutes.
- **Month**, **Hour**, **Day of Week**: Ridership volume is likely to change based on the season, time of day, and whether it is a weekday or weekend. Use the start time of the trip to obtain these values. The New York City data includes the seconds in their timestamps, while Washington and Chicago do not. The <u>datetime</u> package will be very useful here to make the needed conversions.
- **User Type**: It is possible that users who are subscribed to a bike-share system will have different patterns of use compared to users who only have temporary passes. Washington divides its users into two types: 'Registered' for users with annual, monthly, and other longer-term subscriptions, and 'Casual', for users with 24-hour, 3-day, and other short-term passes. The New York and Chicago data uses 'Subscriber' and 'Customer' for these groups, respectively. For consistency, you will convert the Washington labels to match the other two.

Question 3a: Complete the helper functions in the code cells below to address each of the cleaning tasks described above.

In [3]:

```python
def duration_in_mins(datum, city):
    """
    Takes as input a dictionary containing info about a single trip (datum) and
    its origin city (city) and returns the trip duration in units of minutes.

    Remember that Washington is in terms of milliseconds while Chicago and NYC
    are in terms of seconds.

    HINT: The csv module reads in all of the data as strings, including numeric
    values. You will need a function to convert the strings into an appropriate
    numeric type when making your transformations
    see https://docs.python.org/3/library/functions.html
    """
    # YOUR CODE HERE
    ## If condition checking the city for which the data is to be considered

    if(city == 'Washington'):
        datum['Duration (ms)'] = (float(datum['Duration (ms)'])/60000)
```

```
            duration = datum['Duration (ms)']

    elif(city == 'NYC' or city == 'Chicago'):
            datum['tripduration'] = (float(datum['tripduration']))/60
            duration = datum['tripduration']

    return duration


# Some tests to check that your code works. There should be no output if al
l of
# the assertions pass. The `example_trips` dictionary was obtained from whe
n
# you printed the first trip from each of the original data files.
tests = {'NYC': 13.9833,
         'Chicago': 15.4333,
         'Washington': 7.1231}

for city in tests:
    assert abs(duration_in_mins(example_trips[city], city) - tests[city])
< .001
```

In [4]:

```
def time_of_trip(datum, city):
    """
    Takes as input a dictionary containing info about a single trip (datum)
and
    its origin city (city) and returns the month, hour, and day of the
week in
    which the trip was made.

    Remember that NYC includes seconds, while Washington and Chicago do not
.

    HINT: You should use the datetime module to parse the original date
    strings into a format that is useful for extracting the desired informa
tion.
    see https://docs.python.org/3/library/datetime.html#strftime-and-strpti
me-behavior
    """

    # YOUR CODE HERE

    if(city == 'NYC'):                                          # c
alculating time and date for NYC
        date = datetime.strptime(datum['stoptime'],'%m/%d/%Y %H:%M:%S')

    elif(city == 'Washington'):                                 # c
alculating time and date for Washington
        date = datetime.strptime(datum['Start date'],'%m/%d/%Y %H:%M')

    elif(city == 'Chicago'):                                    # c
alculating time and date for Chicago
        date = datetime.strptime(datum['stoptime'],'%m/%d/%Y %H:%M')

    return (date.month, date.hour, date.strftime('%A'))


# Some tests to check that your code works. There should be no output if al
```

```
l of
# the assertions pass. The `example_trips` dictionary was obtained from whe
n
# you printed the first trip from each of the original data files.
tests = {'NYC': (1, 0, 'Friday'),
         'Chicago': (3, 23, 'Thursday'),
         'Washington': (3, 22, 'Thursday')}

for city in tests:
    assert time_of_trip(example_trips[city], city) == tests[city]
```

In [5]:

```
def type_of_user(datum, city):
    """
    Takes as input a dictionary containing info about a single trip (datum)
and
    its origin city (city) and returns the type of system user that made th
e
    trip.

    Remember that Washington has different category names compared to Chica
go
    and NYC.
    """

    # YOUR CODE HERE

    ## Here we are normalising the user category data in a uniform manner

    if(city == 'Washington'):
        if(datum['Member Type'] == 'Registered'):
            user_type = 'Subscriber'
        elif(datum['Member Type'] == 'Casual'):
            user_type = 'Customer'

    elif(city == 'Chicago' or city =='NYC'):
        user_type = datum['usertype']


    return user_type


# Some tests to check that your code works. There should be no output if al
l of
# the assertions pass. The `example_trips` dictionary was obtained from whe
n
# you printed the first trip from each of the original data files.
tests = {'NYC': 'Customer',
         'Chicago': 'Subscriber',
         'Washington': 'Subscriber'}

for city in tests:
    assert type_of_user(example_trips[city], city) == tests[city]
```

Question 3b: Now, use the helper functions you wrote above to create a condensed data file for each city consisting only of the data fields indicated above. In the /examples/ folder, you will see an example datafile from the Bay Area Bike Share before and after conversion. Make sure that your

output is formatted to be consistent with the example file.

In [6]:

```python
def condense_data(in_file, out_file, city):
    """
    This function takes full data from the specified input file
    and writes the condensed data to a specified output file. The city
    argument determines how the input file will be parsed.

    HINT: See the cell below to see how the arguments are structured!
    """

    with open(out_file, 'w') as f_out, open(in_file, 'r') as f_in:
        # set up csv DictWriter object - writer requires column names for the
        # first row as the "fieldnames" argument
        out_colnames = ['duration', 'month', 'hour', 'day_of_week', 'user_type']
        trip_writer = csv.DictWriter(f_out, fieldnames = out_colnames)
        trip_writer.writeheader()

        ## TODO: set up csv DictReader object ##

        trip_reader = csv.DictReader(f_in,skipinitialspace = True)

        # collect data from and process each row
        for row in trip_reader:
            # set up a dictionary to hold the values for the cleaned and trimmed
            # data point
            new_point = {}

            ## TODO: use the helper functions to get the cleaned data from ##
            ## the original data dictionaries. ##
            ## Note that the keys for the new_point dictionary should match ##
            ## the column names set in the DictWriter object above. ##

            new_point['duration'] = duration_in_mins(row, city)
            month , hour , day_of_week = time_of_trip(row, city)
            new_point['month'] = month
            new_point['hour'] = hour
            new_point['day_of_week'] = day_of_week
            new_point['user_type'] = type_of_user(row, city)


            ## TODO: write the processed information to the output file. ##
            ## see https://docs.python.org/3/library/csv.html#writer-objects ##

            trip_writer.writerow(new_point)
```

```
# Run this cell to check your work
city_info = {'Washington': {'in_file': './data/Washington-CapitalBikeshare-
2016.csv',
                            'out_file': './data/Washington-2016-Summary.csv
},
             'Chicago': {'in_file': './data/Chicago-Divvy-2016.csv',
                         'out_file': './data/Chicago-2016-Summary.csv'},
             'NYC': {'in_file': './data/NYC-CitiBike-2016.csv',
                     'out_file': './data/NYC-2016-Summary.csv'}}

for city, filenames in city_info.items():
    condense_data(filenames['in_file'], filenames['out_file'], city)
    print_first_point(filenames['out_file'])
```

```
City: Washington
OrderedDict([('duration', '7.123116666666666'),
             ('month', '3'),
             ('hour', '22'),
             ('day_of_week', 'Thursday'),
             ('user_type', 'Subscriber')])

City: Chicago
OrderedDict([('duration', '15.433333333333334'),
             ('month', '3'),
             ('hour', '23'),
             ('day_of_week', 'Thursday'),
             ('user_type', 'Subscriber')])

City: NYC
OrderedDict([('duration', '13.983333333333333'),
             ('month', '1'),
             ('hour', '0'),
             ('day_of_week', 'Friday'),
             ('user_type', 'Customer')])
```

> **Tip**: If you save a jupyter Notebook, the output from running code blocks will also be saved. However, the state of your workspace will be reset once a new session is started. Make sure that you run all of the necessary code blocks from your previous session to reestablish variables and functions before picking up where you last left off.

Exploratory Data Analysis

Now that you have the data collected and wrangled, you're ready to start exploring the data. In this section you will write some code to compute descriptive statistics from the data. You will also be introduced to the `matplotlib` library to create some basic histograms of the data.

Statistics

First, let's compute some basic counts. The first cell below contains a function that uses the csv module to iterate through a provided data file, returning the number of trips made by subscribers and customers. The second cell runs this function on the example Bay Area data in the `/examples/` folder. Modify the cells to answer the question below.

Question 4a: Which city has the highest number of trips? Which city has the highest proportion of trips made by subscribers? Which city has the highest proportion of trips made by short-term customers?

Answer:

- **Answer 1** :- Based on the Statistics generated from data **New York City (NYC)** Has the highest number of trips followed by **Chicago** and then **Washington**

- **Answer 2** :- Based on the Statistics generated from data **New York City (NYC)** Has the highest proportion of trips made by subscribers followed by **Washington** and then **Chicago**

- **Answer 3** :- Based on the Statistics generated from data **Chicago** Has the highest proportion of trips made by short-term customers followed by **Washington** and then **New York City (NYC)**

In [8]:

```python
def number_of_trips(filename):
    """
    This function reads in a file with trip data and reports the number of
    trips made by subscribers, customers, and total overall.
    """
    with open(filename, 'r') as f_in:
        # set up csv reader object
        reader = csv.DictReader(f_in)

        # initialize count variables
        n_subscribers = 0
        n_customers = 0

        # tally up ride types
        for row in reader:
            if row['user_type'] == 'Subscriber':
                n_subscribers += 1
            else:
                n_customers += 1

        # compute total number of rides
        n_total = n_subscribers + n_customers

        # return tallies as a tuple
        return(n_subscribers, n_customers, n_total)
```

In [9]:

```python
## Modify this and the previous cell to answer Question 4a. Remember to run ##
## the function on the cleaned data files you created from Question 3. ##

## Creating a key to select the data files
location= {'Washington':'./data/Washington-2016-Summary.csv','Chicago':'./data/Chicago-2016-Summary.csv','NYC':'./data/NYC-2016-Summary.csv'}

## Statistics for Washington
```

```python
data_file = location['Washington']
no_subscribers, no_customers, np_total = number_of_trips(data_file) ## Stor
ing data in the variables
print("Washington \nTotal No of trips = {}\n Trips by Subscriber = {} % \n
Trips by customers = {} % \n".format(np_total,((no_subscribers*100)/np_tota
l),((no_customers*100)/np_total)))

## Statistics for Chicago


data_file = location['Chicago']
no_subscribers, no_customers, np_total = number_of_trips(data_file)  ## Sto
ring data in the variables
print("Chicago \nTotal No of trips = {}\n Trips by Subscriber = {} % \n Tri
ps by customers = {} % \n".format(np_total,((no_subscribers*100)/np_total),
((no_customers*100)/np_total)))


## Statistics for NYC

data_file = location['NYC']
no_subscribers, no_customers, np_total = number_of_trips(data_file)  ## Sto
ring data in the variables
print("NYC \nTotal No of trips = {}\n Trips by Subscriber = {} %\n Trips by
customers = {} % \n".format(np_total,((no_subscribers*100)/np_total),((no_c
ustomers*100)/np_total)))
```

```
Washington
Total No of trips = 66326
 Trips by Subscriber = 78.0282242257938 %
 Trips by customers = 21.971775774206193 %

Chicago
Total No of trips = 72131
 Trips by Subscriber = 76.22520136973007 %
 Trips by customers = 23.774798630269924 %

NYC
Total No of trips = 276798
 Trips by Subscriber = 88.83590199351151 %
 Trips by customers = 11.164098006488485 %
```

> **Tip**: In order to add additional cells to a notebook, you can use the "Insert Cell Above"
> and "Insert Cell Below" options from the menu bar above. There is also an icon in the
> toolbar for adding new cells, with additional icons for moving the cells up and down
> the document. By default, new cells are of the code type; you can also specify the cell
> type (e.g. Code or Markdown) of selected cells from the Cell menu or the dropdown in
> the toolbar.

Now, you will write your own code to continue investigating properties of the data.

Question 4b: Bike-share systems are designed for riders to take short trips. Most of the time, users
are allowed to take trips of 30 minutes or less with no additional charges, with overage charges made
for trips of longer than that duration. What is the average trip length for each city? What proportion of

rides made in each city are longer than 30 minutes?

Answer:

- **Answer 1** :- Based on the Statistics generated from data **Washington** Has the highest average trip length (18.4) followed by **Chicago** (16.0) and then **New York City (NYC)** (15.3)

- **Answer 2** :- Based on the Statistics generated from data **Washington** (10.2%) Has the highest proportion of rides made in each city longer than 30 minutes followed by **Chicago** (7.6%) and then **New York City (NYC)** (6.6%)

In [10]:

```python
## Use this and additional cells to answer Question 4b.            ##
##                                                                  ##
## HINT: The csv module reads in all of the data as strings, including  ##
## numeric values. You will need a function to convert the strings     ##
## into an appropriate numeric type before you aggregate data.         ##
## TIP: For the Bay Area example, the average trip length is 14 minutes ##
## and 3.5% of trips are longer than 30 minutes.                    ##

## Function to calculate the average trip length of each city

def Average_trip_length(filename):
    with open(filename, 'r') as f_in:

        # seting up csv reader object
        trip_reader = csv.DictReader(f_in)

        # initialize count variables
        no_of_the_trips = 0
        sum_of_the_trip = 0
        more_than_30_mins_trip = 0

        # tally up trip values to calculate the given parameters from the d
ata
        for row in trip_reader:
            if int(float(row['duration'])) > 30:
                more_than_30_mins_trip += 1
            sum_of_the_trip += int(float(row['duration']))
            no_of_the_trips += 1

        # computing the average trip length
        average_trip_length = (sum_of_the_trip/no_of_the_trips)
        ratio_of_rides_more_than_30_mins = (more_than_30_mins_trip*100)/no_
of_the_trips

        # return tallies as a tuple
        return(ratio_of_rides_more_than_30_mins,average_trip_length)
```

In [11]:

```python
## Displaying the result after processing from the sumary of data generated
before

## Creating a key to select the data files
location= {'Washington':'./data/Washington-2016-Summary.csv','Chicago':'./d
ata/Chicago-2016-Summary.csv', 'NYC':'./data/NYC-2016-Summary.csv'}
```

```
ata/CHICagO-2016-Summary.csv', NIC : './data/NIC-2016-Summary.csv' }

## Statistics for Washington


data_file = location['Washington']
more_than_30_min , average_trip = Average_trip_length(data_file) ## Storing
data in the variables
print("Washington \n Average_trip = {}\n Trips of longer than 30 mins = {}
% \n".format(average_trip,more_than_30_min))

## Statistics for Chicago


data_file = location['Chicago']
more_than_30_min , average_trip = Average_trip_length(data_file) ## Storing
data in the variables
print("Chicago \n Average_trip = {}\n Trips of longer than 30 mins = {} % \
n".format(average_trip,more_than_30_min))


## Statistics for NYC

data_file = location['NYC']
more_than_30_min , average_trip = Average_trip_length(data_file) ## Storing
data in the variables
print("NYC \n Average_trip = {}\n Trips of longer than 30 mins = {} % \n".f
ormat(average_trip,more_than_30_min))
```

```
Washington
 Average_trip = 18.4320929952055
 Trips of longer than 30 mins = 10.241835780840093 %

Chicago
 Average_trip = 16.071286963996062
 Trips of longer than 30 mins = 7.662447491369869 %

NYC
 Average_trip = 15.322249438218485
 Trips of longer than 30 mins = 6.614570914529729 %
```

Question 4c: Dig deeper into the question of trip duration based on ridership. Choose one city. Within that city, which type of user takes longer rides on average: Subscribers or Customers?

Answer:

- **Answer 1** :-Based on the Statistics generated from data in almost all the three cities
  Subscriber tend to take longer rides as compared to the customers by a huge margin

In [12]:

```
## Use this and additional cells to answer Question 4c. If you have    ##
## not done so yet, consider revising some of your previous code to     ##
## make use of functions for reusability.                               ##
##                                                                      ##
## TIP: For the Bay Area example data, you should find the average      ##
## Subscriber trip duration to be 9.5 minutes and the average Customer ##
```

```python
## trip duration to be 54.6 minutes. Do the other cities have this     ##
## level of difference?                                                ##

def Average_trip_duration_user(filename):
    with open(filename, 'r') as f_in:

        # seting up csv reader object
        trip_reader = csv.DictReader(f_in)

        # initialize count variables
        no_of_the_trips_by_customer = 0
        no_of_the_trip_by_subscriber = 0
        total_no_of_trips = 0

        # tally up trip values to calculate the given parameters from the d
ata
        for row in trip_reader:
            if row['user_type'] == 'Customer':
                no_of_the_trips_by_customer += 1

            total_no_of_trips += 1
        no_of_the_trip_by_subscriber = (total_no_of_trips -
no_of_the_trips_by_customer)

        # computing the average trip length for users
        average_trip_length_customer = ((no_of_the_trips_by_customer*100)/t
otal_no_of_trips)
        average_trip_length_subscriber = ((no_of_the_trip_by_subscriber*100
)/total_no_of_trips)

        # return tallies as a tuple
        return(average_trip_length_customer,average_trip_length_subscriber)
```

In [ ]:

```python
## Displaying the result after processing from the sumary of data generated
before

## Creating a key to select the data files
location= {'Washington':'./data/Washington-2016-Summary.csv','Chicago':'./d
ata/Chicago-2016-Summary.csv','NYC':'./data/NYC-2016-Summary.csv'}

## Statistics for Washington


data_file = location['Washington']
average_trip_length_customer , average_trip_length_subscriber = Average_tri
p_duration_user(data_file) ## Storing data in the variables
print("Washington \n Average_trip_Customert = {}\n Average_trip_Subscriber
= {} %
\n".format(average_trip_length_customer,average_trip_length_subscriber))

## Statistics for Chicago


data_file = location['Chicago']
average_trip_length_customer , average_trip_length_subscriber = Average_tri
p_duration_user(data_file) ## Storing data in the variables
print("Chicago \n Average_trip_Customert = {}\n Average_trip_Subscriber = {
} % \n".format(average_trip_length_customer,average_trip_length_subscriber)
```

```
)


## Statistics for NYC

data_file = location['NYC']
average_trip_length_customer , average_trip_length_subscriber = Average_tri
p_duration_user(data_file) ## Storing data in the variables
print("NYC \n Average_trip_Customert = {}\n Average_trip_Subscriber = {} %
\n".format(average_trip_length_customer,average_trip_length_subscriber))
```

```
Washington
 Average_trip_Customert = 21.971775774206193
 Average_trip_Subscriber = 78.0282242257938 %
```

Visualizations

The last set of values that you computed should have pulled up an interesting result. While the mean trip time for Subscribers is well under 30 minutes, the mean trip time for Customers is actually *above* 30 minutes! It will be interesting for us to look at how the trip times are distributed. In order to do this, a new library will be introduced here, `matplotlib`. Run the cell below to load the library and to generate an example plot.

In [296]:

```python
# load library
import matplotlib.pyplot as plt

# this is a 'magic word' that allows for plots to be displayed
# inline with the notebook. If you want to know more, see:
# http://ipython.readthedocs.io/en/stable/interactive/magics.html
%matplotlib inline

# example histogram, data taken from bay area sample
data = [ 7.65,  8.92,  7.42,  5.50, 16.17,  4.20,  8.98,  9.62, 11.48, 14.3
,
        19.02, 21.53,  3.90,  7.97,  2.62,  2.67,  3.08, 14.40, 12.90,  7.8
,
        25.12,  8.30,  4.93, 12.43, 10.60,  6.17, 10.88,  4.78, 15.15,  3.5
,
         9.43, 13.32, 11.72,  9.85,  5.22, 15.10,  3.95,  3.17,  8.78,  1.8
,
         4.55, 12.68, 12.38,  9.78,  7.63,  6.45, 17.38, 11.90, 11.52,  8.6
,]
plt.hist(data)
plt.title('Distribution of Trip Durations')
plt.xlabel('Duration (m)')
plt.show()
```
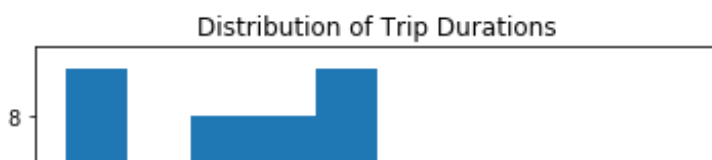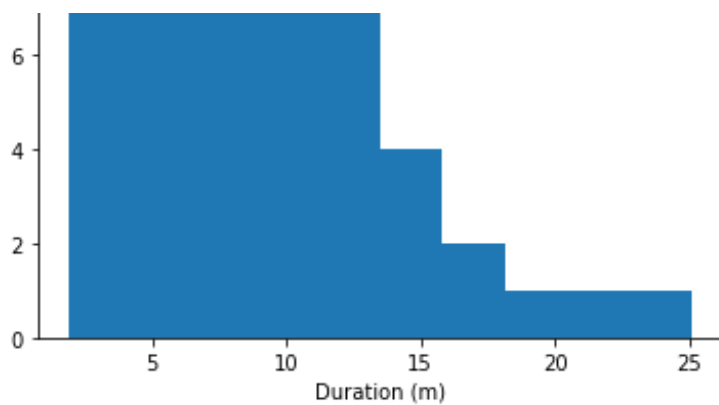
In the above cell, we collected fifty trip times in a list, and passed this list as the first argument to the `.hist()` function. This function performs the computations and creates plotting objects for generating a histogram, but the plot is actually not rendered until the `.show()` function is executed. The `.title()` and `.xlabel()` functions provide some labeling for plot context.

You will now use these functions to create a histogram of the trip times for the city you selected in question 4c. Don't separate the Subscribers and Customers for now: just collect all of the trip times and plot them.

In [297]:

```python
## Use this and additional cells to collect all of the trip times as a list
##
## and then use pyplot functions to generate a histogram of trip times.
##

%matplotlib inline
def plota(filename):

    data = []

    with open(filename, 'r') as f_in:
        # seting up csv reader object
        trip_duration_reader = csv.DictReader(f_in)
        for row in trip_duration_reader:
            data.append( float( row['duration'] ) )
    return data


## Washington city plot

plt.hist(plota(location['Washington']),range=(0,75), bins=15)
plt.title('Washington Trip-Times Distribution')
plt.xlabel('Duration (m)')
plt.show()

## Chicago city plot

plt.hist(plota(location['Chicago']),range=(0,75), bins=15)
plt.title('Chicago Trip-Times Distribution')
plt.xlabel('Duration (m)')
plt.show()

## NYC city plot

plt.hist(plota(location['NYC']),range=(0,75), bins=15)
```

```
plt.title('NYC Trip-Times Distribution')
plt.xlabel('Duration (m)')
plt.show()
```


Washington Trip-Times Distribution

Chicago Trip-Times Distribution

NYC Trip-Times Distribution
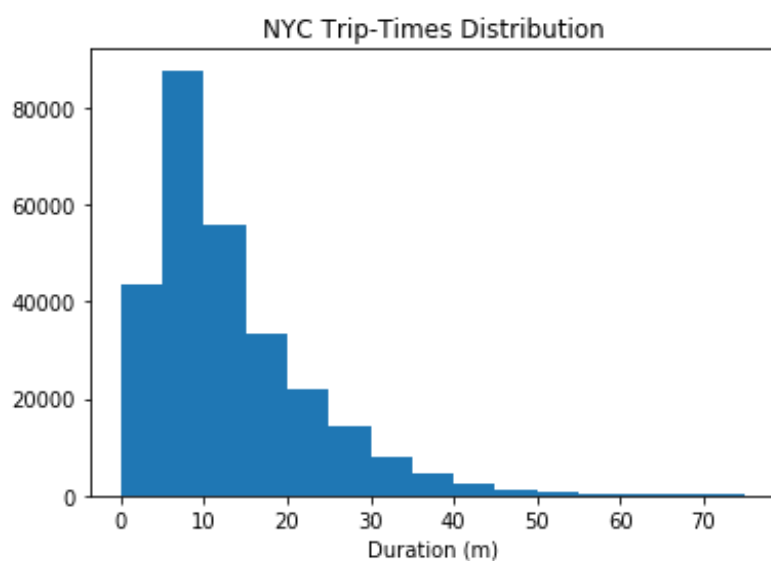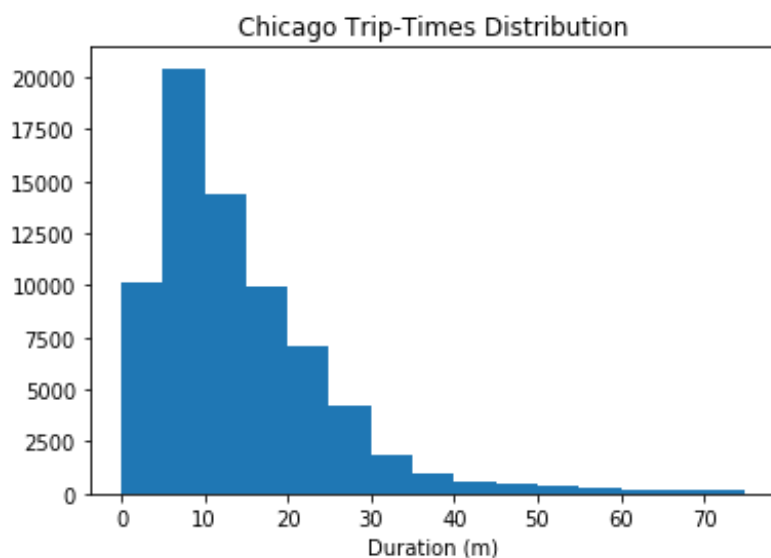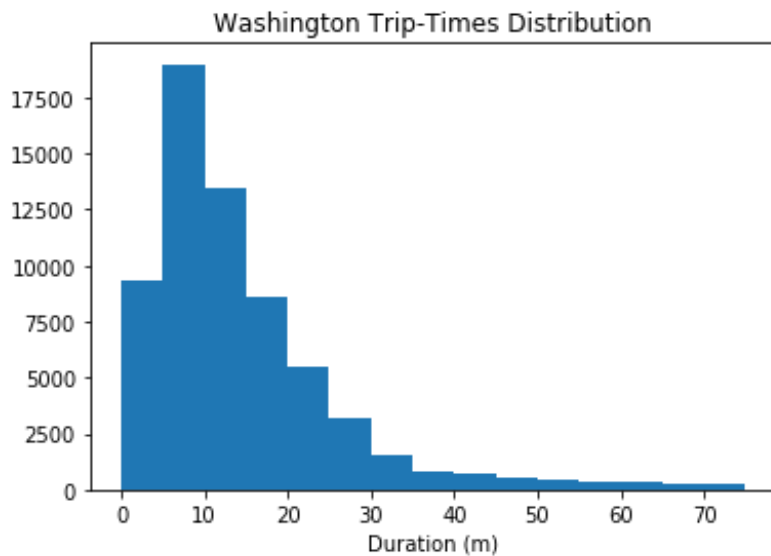
In [298]:

```
## Creating a function to plot a graph

def create_histogram(filename,city,label,data_label,bin_no):
```

```
    %matplotlib inline

    with open(filename, 'r') as f_in:

        # seting up csv reader object
        trip_duration_reader = csv.DictReader(f_in)
        dataC = []
        dataS = []
        ## collecting the data for the plot for different user and cities a
nd ploting the histogram


        for row in trip_duration_reader:
            if row['user_type'] == "Customer":
                dataC.append( float( row[data_label] ) )
            else :
                dataS.append( float( row[data_label] ) )
    if(label == 'Customer'):
        plt.hist(dataC,range=(0,75), bins = bin_no)
    else:
        plt.hist(dataS,range=(0,75), bins = bin_no)
    plt.title(city + " " + label + " Trip-Times Distribution" )
    plt.xlabel('Duration (m)')
    plt.show()
```

If you followed the use of the `.hist()` and `.show()` functions exactly like in the example, you're probably looking at a plot that's completely unexpected. The plot consists of one extremely tall bar on the left, maybe a very short second bar, and a whole lot of empty space in the center and right. Take a look at the duration values on the x-axis. This suggests that there are some highly infrequent outliers in the data. Instead of reprocessing the data, you will use additional parameters with the `.hist()` function to limit the range of data that is plotted. Documentation for the function can be found [here].

Question 5: Use the parameters of the `.hist()` function to plot the distribution of trip times for the Subscribers in your selected city. Do the same thing for only the Customers. Add limits to the plots so that only trips of duration less than 75 minutes are plotted. As a bonus, set the plots up so that bars are in five-minute wide intervals. For each group, where is the peak of each distribution? How would you describe the shape of each distribution?

Answer:

- **Answer 1**:- For the **Subscribers in all the cities the peak values stays in between 0-10 range** in the histogram while for the **Customer** in case of **washington city the peak is in the range of 10-20** while for the **reamianing city it is in the range 20-30** of the histogram

- **Answer 2**:- The shape of the curve for both **Subscribers and Customer is positively skewed** while the **tail of the customer plots is thicker as compared to the Subscribers plots which is much thinner**

```
In [299]:
```

```
## Use this and additional cells to answer Question 5. ##

location= {'Washington':'./data/Washington-2016-Summary.csv','Chicago':'./d
```

```
ata/Chicago-2016-Summary.csv','NYC':'./data/NYC-2016-Summary.csv'}

## creating the plots for the customer and subscriber in washington
create_histogram(location['Washington'],'Washington','Customer','duration'
,15)
create_histogram(location['Washington'],'Washington','Subscriber','duration
' ,15)

## creating the plots for the customer and subscriber in Chicago
create_histogram(location['Chicago'],'Chicago','Customer','duration' ,15)
create_histogram(location['Chicago'],'Chicago','Subscriber','duration' ,15)

## creating the plots for the customer and subscriber in NYC
create_histogram(location['NYC'],'NYC','Customer','duration' ,15)
create_histogram(location['NYC'],'NYC','Subscriber','duration',15)
```
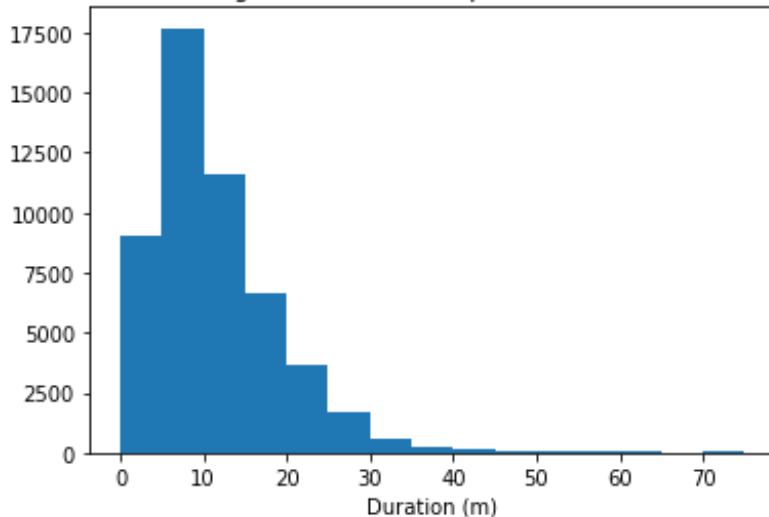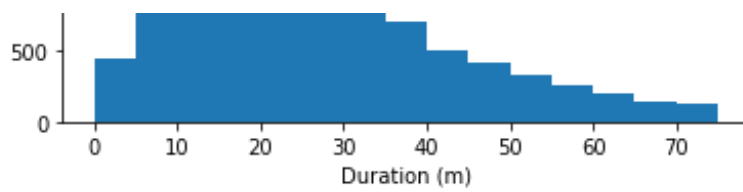

Washington Customer Trip-Times Distribution
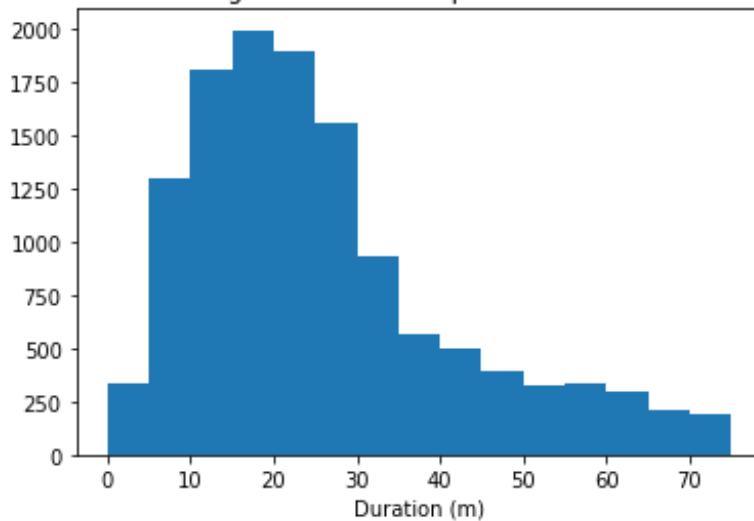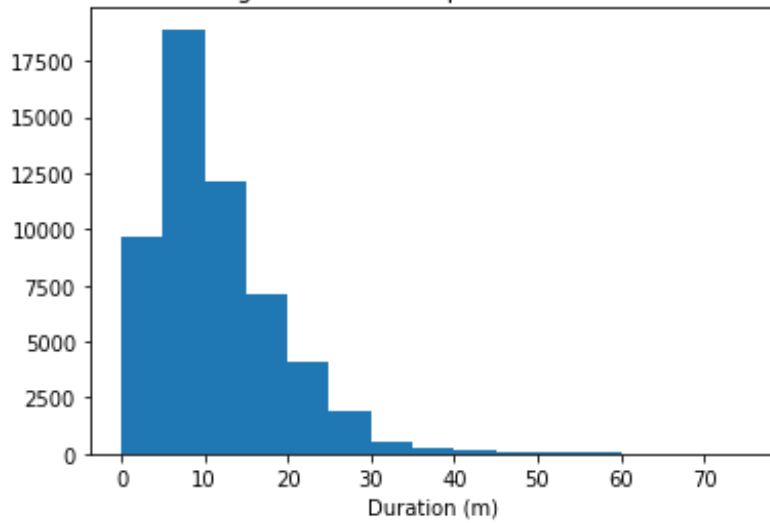

Washington Subscriber Trip-Times Distribution

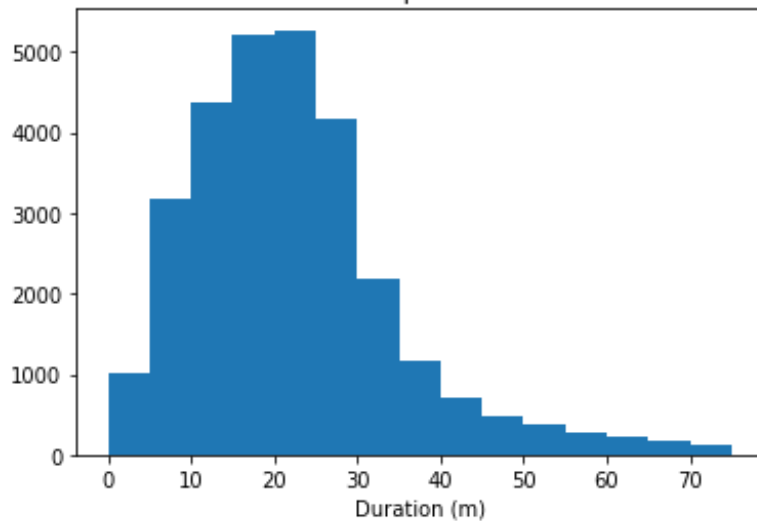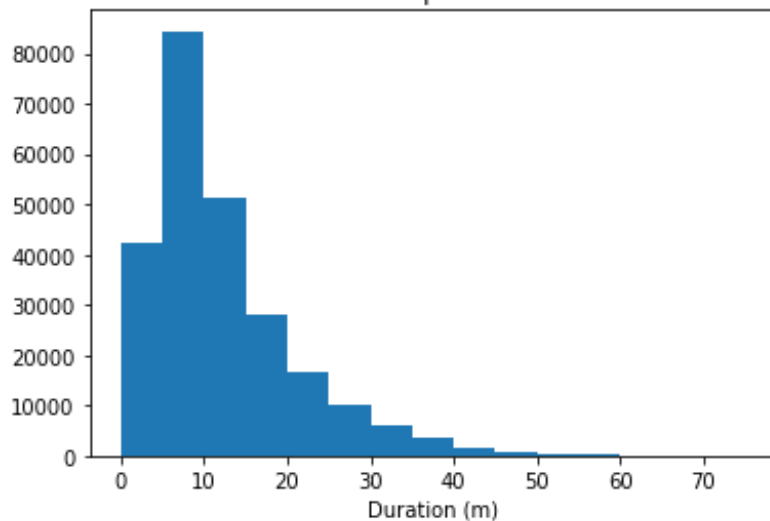Chicago Customer Trip-Times Distribution
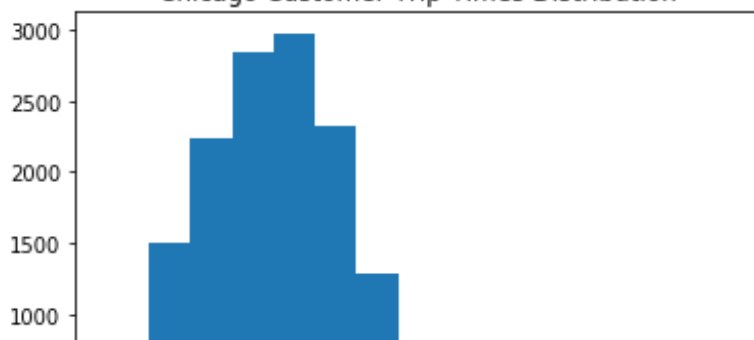
Chicago Subscriber Trip-Times Distribution

NYC Customer Trip-Times Distribution

NYC Subscriber Trip-Times Distribution

Performing Your Own Analysis

Performing Your Own Analysis

So far, you've performed an initial exploration into the data available. You have compared the relative volume of trips made between three U.S. cities and the ratio of trips made by Subscribers and Customers. For one of these cities, you have investigated differences between Subscribers and Customers in terms of how long a typical trip lasts. Now it is your turn to continue the exploration in a direction that you choose. Here are a few suggestions for questions to explore:

- How does ridership differ by month or season? Which month / season has the highest ridership? Does the ratio of Subscriber trips to Customer trips change depending on the month or season?
- Is the pattern of ridership different on the weekends versus weekdays? On what days are Subscribers most likely to use the system? What about Customers? Does the average duration of rides change depending on the day of the week?
- During what time of day is the system used the most? Is there a difference in usage patterns for Subscribers and Customers?

If any of the questions you posed in your answer to question 1 align with the bullet points above, this is a good opportunity to investigate one of them. As part of your investigation, you will need to create a visualization. If you want to create something other than a histogram, then you might want to consult the Pyplot documentation. In particular, if you are plotting values across a categorical variable (e.g. city, user type), a bar chart will be useful. The documentation page for `.bar()` includes links at the bottom of the page with examples for you to build off of for your own use.

Question 6: Continue the investigation by exploring another question that could be answered by the data available. Document the question you want to explore below. Your investigation should involve at least two variables and should compare at least two groups. You should also use at least one visualization as part of your explorations.

Answer:

Q:- Which days of the week experince high usage in different cities and by which type of users?

- **Answer**

- For the City **Washington** Here based on the stats we obeserver that on **saturday the customer rush is very high** while on **wedenesday , tuesday , thursday appear to be very high for the Subscriber***

- For the City **Chicago** Here based on the stats we obeserver that on **sunday and saturday the customer rush is very high** while on **monday , tuesday appear to be very high for the Subscriber**

- For the City **NYC** Here based on the stats we obeserver that on **saturday and sunday the customer rush is very high** while on **wedenesday , thursday appear to be very high for the Subscriber**

In [300]:

```
## plotting the bar graphs for the data collected above
days_of_week = ['Sun', 'Mon', 'Tues', 'Wed', 'Thurs', 'Fri', 'Sat']
def plot_bargraph(user,city,data):
    %matplotlib inline
```

```
    plt.bar(range(len(days_of_week)),data)
    plt.xticks(range(len(days_of_week)),days_of_week)
    plt.ylabel('Trip Count')
    plt.xlabel('Days')
    plt.title('Total Trips made by ' + user + ' in ' + city + ' for seven da
ys of the week')
    plt.show()
```

In [301]:

```
## Use this and additional cells to continue to explore the dataset. ##
## Once you have performed your exploration, document your findings  ##
## in the Markdown cell above. ##


def no_of_trip_by_user(filename):
    with open(filename, 'r') as f_in:
        day = ['Sunday', 'Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Frid
ay', 'Saturday']
        # seting up csv reader object
        trip_reader = csv.DictReader(f_in)

        # initialize count variables
        data_of_no_of_the_trips_by_customer = [0,0,0,0,0,0,0]
        data_of_no_of_the_trip_by_subscriber = [0,0,0,0,0,0,0]

        # tally up trip values to calculate the given parameters from the d
ata
        for row in trip_reader:
            if row['user_type'] == 'Customer':
                for count in range(7):
                    if row['day_of_week'] == day[count]:
                        data_of_no_of_the_trips_by_customer[count] += 1
            elif row['user_type'] == 'Subscriber':
                for count in range(7):
                    if row['day_of_week'] == day[count]:
                        data_of_no_of_the_trip_by_subscriber[count] += 1

        # return tallies as a tuple

return(data_of_no_of_the_trips_by_customer,data_of_no_of_the_trip_by_subscri
ber)
```

In [302]:

```
## Displaying the Data now

location= {'Washington':'./data/Washington-2016-Summary.csv','Chicago':'./d
ata/Chicago-2016-Summary.csv','NYC':'./data/NYC-2016-Summary.csv'}

## creating the plots for the customer and subscriber in washington
cust_trip , subs_trip = no_of_trip_by_user(location['Washington'])

plot_bargraph('Customer','Washington',cust_trip)
plot_bargraph('Subscriber','Washington',subs_trip)


## creating the plots for the customer and subscriber in Chicago
```

```
## creating the plots for the customer and subscriber in Chicago
cust_trip , subs_trip = no_of_trip_by_user(location['Chicago'])

plot_bargraph('Customer','Chicago',cust_trip)
plot_bargraph('Subscriber','Chicago',subs_trip)

## creating the plots for the customer and subscriber in NYC
cust_trip , subs_trip = no_of_trip_by_user(location['NYC'])

plot_bargraph('Customer','NYC',cust_trip)
plot_bargraph('Subscriber','NYC',subs_trip)
```
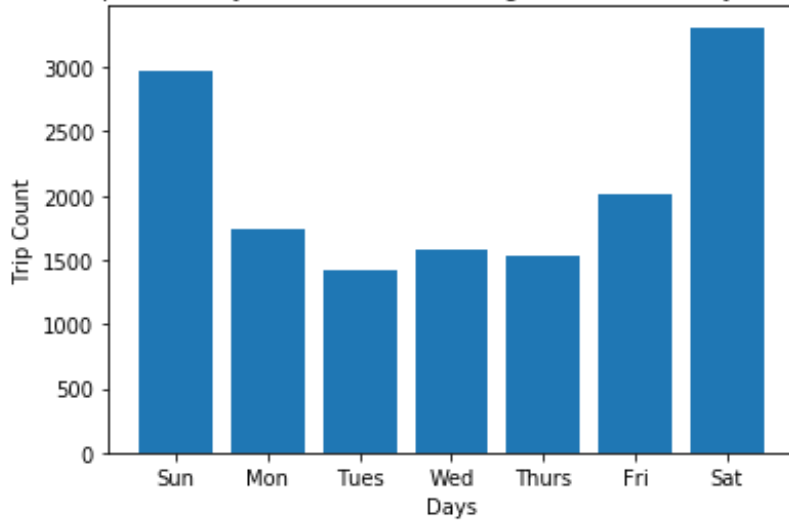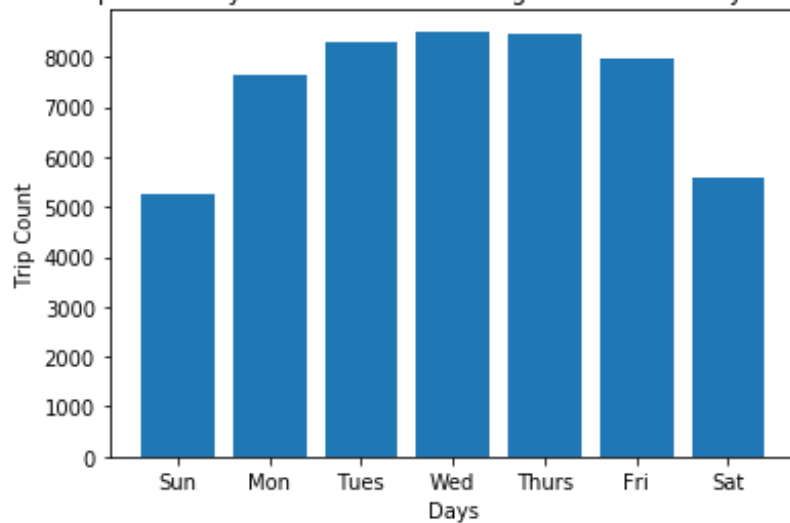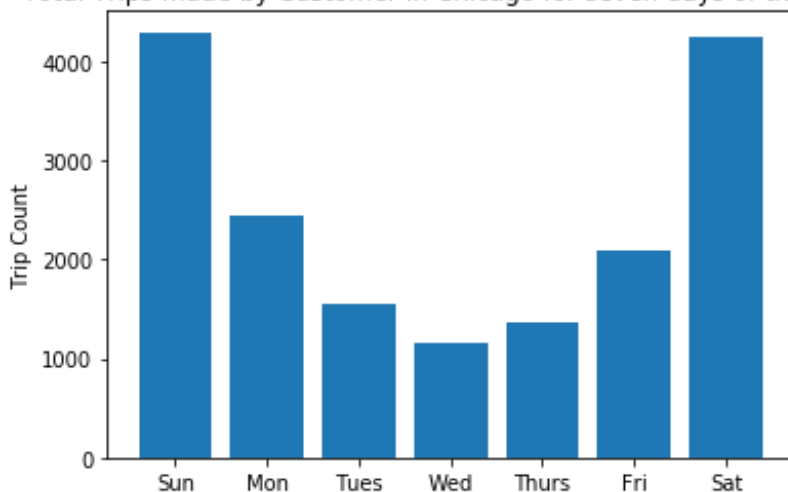
Total Trips made by Customer in Washington for seven days of the week

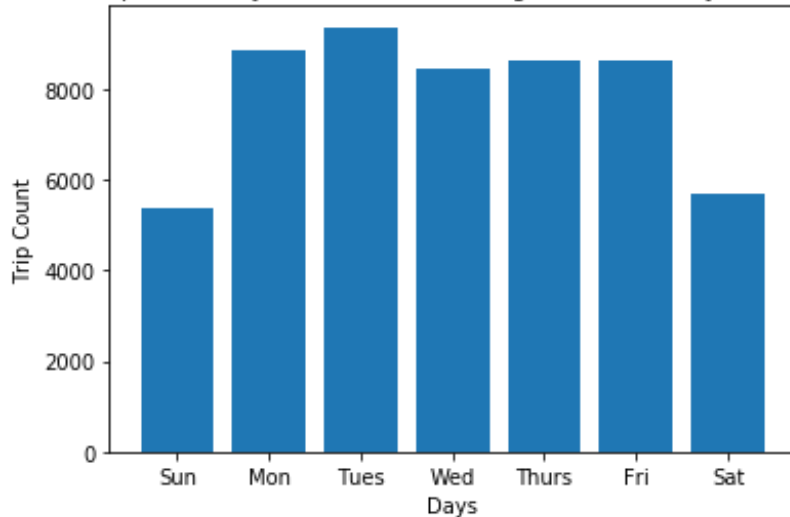Total Trips made by Subscriber in Washington for seven days of the week

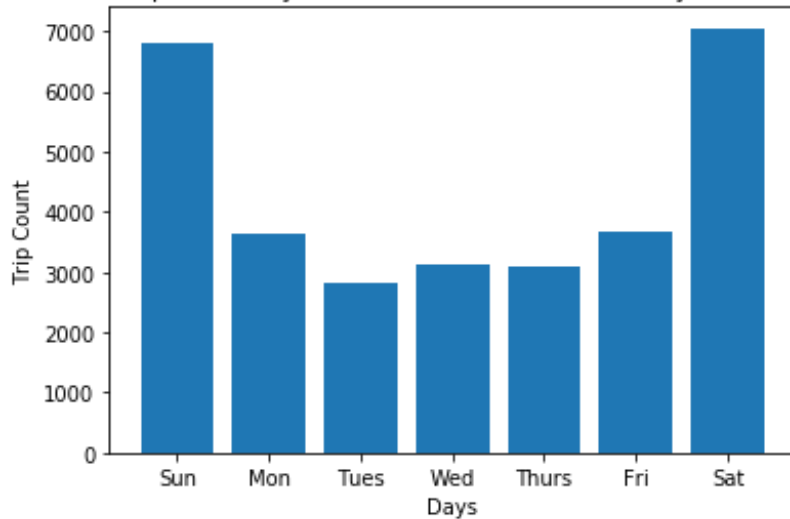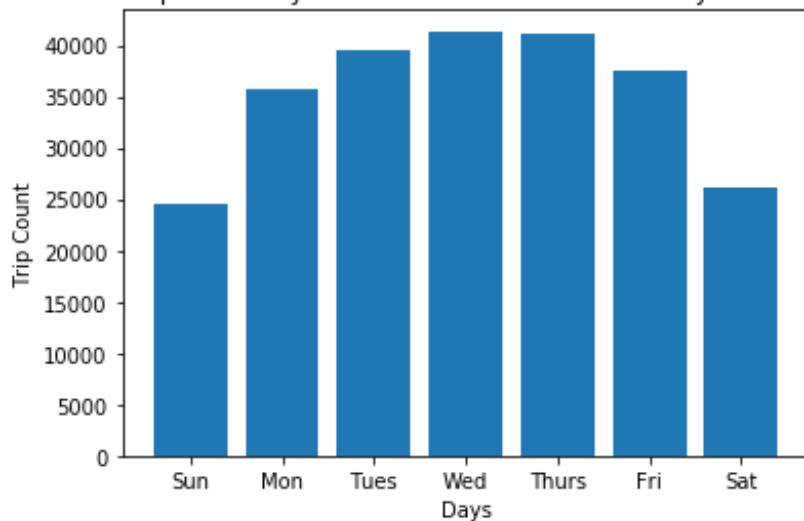Total Trips made by Customer in Chicago for seven days of the week

Total Trips made by Subscriber in Chicago for seven days of the week

Total Trips made by Customer in NYC for seven days of the week

Total Trips made by Subscriber in NYC for seven days of the week

Conclusions

Congratulations on completing the project! This is only a sampling of the data analysis process: from generating questions, wrangling the data, and to exploring the data. Normally, at this point in the data analysis process, you might want to draw conclusions about the data by performing a statistical test or

analysis process, you might want to draw conclusions about the data by performing a statistical test or fitting the data to a model for making predictions. There are also a lot of potential analyses that could be performed on the data which are not possible with only the data provided. For example, detailed location data has not been investigated. Where are the most commonly used docks? What are the most common routes? As another example, weather has potential to have a large impact on daily ridership. How much is ridership impacted when there is rain or snow? Are subscribers or customers affected more by changes in weather?

Question 7: Putting the bike share data aside, think of a topic or field of interest where you would like to be able to apply the techniques of data science. What would you like to be able to learn from your chosen subject?

Answer:

- **I am very much intrested in Cybersecurity systems and this is the area where i consider my self applying these nobel machine learning or data analysis techniques to build more reliant and better defence mechanism to prevent the attacks of future which might also involve ML Techniques to crack the present systems.**
- **I want to apply the power of data analytics via machine learning to build smart systems that are aware of their surroundngs and are well equiped to handle the 0 day attacks like ransomware etc**
- **I am very intrested to dive in furthure in the subject to understand the working of data analysis more better and deeper**

> **Tip**: If we want to share the results of our analysis with others, we aren't limited to giving them a copy of the jupyter Notebook (.ipynb) file. We can also export the Notebook output in a form that can be opened even for those without Python installed. From the **File** menu in the upper left, go to the **Download as** submenu. You can then choose a different format that can be viewed more generally, such as HTML (.html) or PDF (.pdf). You may need additional packages or software to perform these exports.
>
> If you are working on this project via the Project Notebook page in the classroom, you can also submit this project directly from the workspace. **Before you do that**, you should save an HTML copy of the completed project to the workspace by running the code cell below. If it worked correctly, the output code should be a 0, and if you click on the jupyter icon in the upper left, you should see your .html document in the workspace directory. Alternatively, you can download the .html copy of your report following the steps in the previous paragraph, then *upload* the report to the directory (by clicking the jupyter icon).
>
> Either way, once you've gotten the .html report in your workspace, you can complete your submission by clicking on the "Submit Project" button to the lower-right hand side of the workspace.

In [303]:

```python
from subprocess import call
call(['python', '-m', 'nbconvert', 'Bike_Share_Analysis.ipynb'])
```

Out[303]:

0

Resource

- python documentation : Link
- Links shared in the project instructions