

Bike_Share_Analysis

July 9, 2018

1 2016 US Bike Share Activity Snapshot

1.1 Table of Contents

- Section ??
- Section ??
- Section ??
 - Section ??
- Section ??
 - Section ??
 - Section ??
- Section ??
- Section ??

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: a) Which city has minimum riderships? b) Trips made by males in Chicago city.

Answer: a) **Washington** has minimum riderships(66326). b) **41194** males made trips in Chicago city.

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.

```
In [11]: import pandas as pd
```

```
In [12]: data_file1 = pd.read_csv('./data/Chicago-Divvy-2016.csv')
         print(pd.value_counts(data_file1['gender']))
```

```
Male      41194
```

```
Female    13783
```

```
Name: gender, dtype: int64
```

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 [13]: ## 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 [14]: 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('/')[0]
    print('\nCity: {}'.format(city))

    with open(filename, 'r') as f_in:
        #set up a DictReader object.

        trip_reader = csv.DictReader(f_in)

        #read first_trip of data file
        first_trip = next(trip_reader)

        #converts list of tuples in dictionary.
        first_trip = dict(first_trip)

        # output city name and first trip for later testing
        pprint(first_trip)

    return (city, first_trip)

# 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
```

```
City: NYC
{'bikeid': '17109',
 'birth year': '',
 'end station id': '401',
```

```
'end station latitude': '40.72019576',
'end station longitude': '-73.98997825',
'end station name': 'Allen St & Rivington St',
'gender': '0',
'start station id': '532',
'start station latitude': '40.710451',
'start station longitude': '-73.960876',
'start station name': 'S 5 Pl & S 4 St',
'starttime': '1/1/2016 00:09:55',
'stoptime': '1/1/2016 00:23:54',
'tripduration': '839',
'usertype': 'Customer'}
```

City: Chicago

```
{'bikeid': '2295',
'birthyear': '1990',
'from_station_id': '156',
'from_station_name': 'Clark St & Wellington Ave',
'gender': 'Male',
'starttime': '3/31/2016 23:30',
'stoptime': '3/31/2016 23:46',
'to_station_id': '166',
'to_station_name': 'Ashland Ave & Wrightwood Ave',
'trip_id': '9080545',
'tripduration': '926',
'usertype': 'Subscriber'}
```

City: Washington

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

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 `datetime` 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 [15]: 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.
```

```
#import date time using from date time  
from datetime import datetime  
if city == 'NYC':  
    #convert string tripduration into float  
    duration = float(datum['tripduration'])  
  
    #convert seconds into minutes and store in duration  
    duration = duration / 60  
  
elif city == 'Chicago':  
    #convert string tripduration into float  
    duration = float(datum['tripduration'])  
    #convert string tripduration into float
```

```

        duration = duration / 60

    elif city == 'Washington':
        #convert string tripduration into float
        duration = float(datum['Duration (ms)'])
        #convert milliseconds into minutes and store in duration
        duration = duration / 60000

    return duration

# the assertions pass. The `example_trips` dictionary was obtained from when
# 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 [16]:

```

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.

    #import datetime using from datetime
    from datetime import datetime
    if city == 'NYC':
        #store datetime into date
        date = datum['starttime']
        #function strptime returns date value as struct_time
        date = datetime.strptime(date, '%m/%d/%Y %H:%M:%S')
        #return month by accessing month from struct_time
        month = date.month
        #return hour by accessing hour from struct_time
        hour = date.hour
        #strftime function returns week day as name
        day_of_week = date.strftime('%A')

    elif city == 'Chicago':
        date = datum['starttime']
        date = datetime.strptime(date, '%m/%d/%Y %H:%M')
        month = date.month
        hour = date.hour
        day_of_week = date.strftime('%A')

```

```

elif city == 'Washington':
    date = datum['Start date']
    date = datetime.strptime(date, '%m/%d/%Y %H:%M')
    month = date.month
    hour = date.hour
    day_of_week = date.strftime('%A')

return (month, hour, day_of_week)

# Some tests to check that your code works. There should be no output if all of
# the assertions pass. The `example_trips` dictionary was obtained from when
# 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 [17]: 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 the trip.

         if city == 'NYC':
             user_type = datum['usertype']

         elif city == 'Chicago':
             user_type = datum['usertype']

         elif city == 'Washington':
             #Registered for anually, monthly Subscribers
             if datum['Member Type'] == 'Registered':
                 user_type = 'Subscriber'
             #Casual for short-term Customers.
             elif datum['Member Type'] == 'Casual':
                 user_type = 'Customer'

         return user_type

         # Some tests to check that your code works. There should be no output if all of
         # the assertions pass. The `example_trips` dictionary was obtained from when

```

```

# 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 [18]: def condense_data(in_file, out_file, city):

    #open old file for read and write.

    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()

        #set up csv DictReader object
        trip_reader = csv.DictReader(f_in)

        for row in trip_reader:

            #clean and trim unwanted data from old data and put useful data into a Dict
            new_point = {'duration':duration_in_mins(row, city),
                        'month':time_of_trip(row, city)[0],
                        'hour':time_of_trip(row, city)[1],
                        'day_of_week':time_of_trip(row, city)[2],
                        'user_type':type_of_user(row, city)
                        }

            # write the processed information to the output file.

            trip_writer.writerow(new_point)

In [19]: # 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
{'day_of_week': 'Thursday',
 'duration': '7.1231166666666666',
 'hour': '22',
 'month': '3',
 'user_type': 'Subscriber'}

City: Chicago
{'day_of_week': 'Thursday',
 'duration': '15.433333333333334',
 'hour': '23',
 'month': '3',
 'user_type': 'Subscriber'}

City: NYC
{'day_of_week': 'Friday',
 'duration': '13.983333333333333',
 'hour': '0',
 'month': '1',
 '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: NYC has the highest number of trips. NYC has the highest proportion of trips made by **subscribers**. Chicago has the highest proportion of trips made by **short-term customers**.

```

In [20]: 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

            # calculate proportion of subscribers and store into subscribers_proportion.
            subscribers_proportion = n_subscribers / n_total
            # calculate proportion of customers and store into customers_proportion.
            customers_proportion = n_customers / n_total

            # return tallies as a tuple
            return(n_subscribers, n_customers, n_total, subscribers_proportion, customers_p

In [21]: ## 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.      ##

        data_file = ['./data/Washington-2016-Summary.csv',
                      './data/Chicago-2016-Summary.csv',
                      './data/NYC-2016-Summary.csv']
        for data_file in data_file:
            print(number_of_trips(data_file))

(51753, 14573, 66326, 0.7802822422579381, 0.21971775774206193)
(54982, 17149, 72131, 0.7622520136973008, 0.23774798630269925)
(245896, 30902, 276798, 0.8883590199351151, 0.11164098006488486)

```

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: Average trip length in **Washington** is **18.93** and proportion of rides made longer than 30 minutes is **0.108**. Average trip length in **Chicago** is **16.56** and proportion of rides made longer than 30 minutes is **0.083**. Average trip length in **NYC** is **15.81** and proportion of rides made longer than 30 minutes is **0.073**.

```
In [22]: def len_of_trip(filename):
        with open(filename, 'r') as f_in:

            reader = csv.DictReader(f_in)
            #initialize variables with zero.
            trip_sum = 0
            trip_len = 0
            count = 0

            for row in reader:
                #total trips
                trip_len+= 1
                trip_duration = float(row['duration'])
                #total sum of all trip duration
                trip_sum+= trip_duration

                #trip length of more 30 minutes duration.
                if trip_duration > 30:
                    count+= 1

                #calculate average trip duration and store into avg_trip_duration
                avg_trip_duration = trip_sum / trip_len
                #calculate proportion of trip made more than 30 minutes.
                proportion = count / trip_len
            return avg_trip_duration, proportion

In [23]: data_file = ['./data/Washington-2016-Summary.csv',
                    './data/Chicago-2016-Summary.csv',
                    './data/NYC-2016-Summary.csv']
        for data_file in data_file:
            print(len_of_trip(data_file))

(18.93287355913721, 0.10838886711093688)
(16.563629368787335, 0.08332062497400562)
(15.81259299802294, 0.07302437156337835)
```

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: **Customers** take longer rides on average in **Chicago** city.

```
In [24]: def len_trip(filename):
    with open(filename, 'r') as f_in:

        reader = csv.DictReader(f_in)

        #initialize variables with zero to calculate some statistics for subscribers.
        trip_sum1 = 0
        trip_len1 = 0

        #initialize variables with zero to calculate some statistics for customers.
        trip_sum2 = 0
        trip_len2 = 0

        for row in reader:
            trip_duration = float(row['duration'])

            if row['user_type'] == 'Subscriber':
                #trip length of trip made by subscribers.
                trip_len1+= 1
                #sum of all trip duration made by subscribers.
                trip_sum1+= trip_duration

            elif row['user_type'] == 'Customer':
                #trip length of trip made by customers
                trip_len2+= 1
                #sum of all trip duration made by customers
                trip_sum2+= trip_duration
            #calculate average of trip duration made by subscribers.
            avg_trip_duration1 = trip_sum1 / trip_len1
            #calculate average of trip duration made by customers.
            avg_trip_duration2 = trip_sum2 / trip_len2

        return (avg_trip_duration1, avg_trip_duration2)

In [25]: data_file = ['./data/Chicago-2016-Summary.csv']
    for data_file in data_file:
        print(len_trip(data_file))

(12.067201690250076, 30.979781133982506)
```

Visualizations

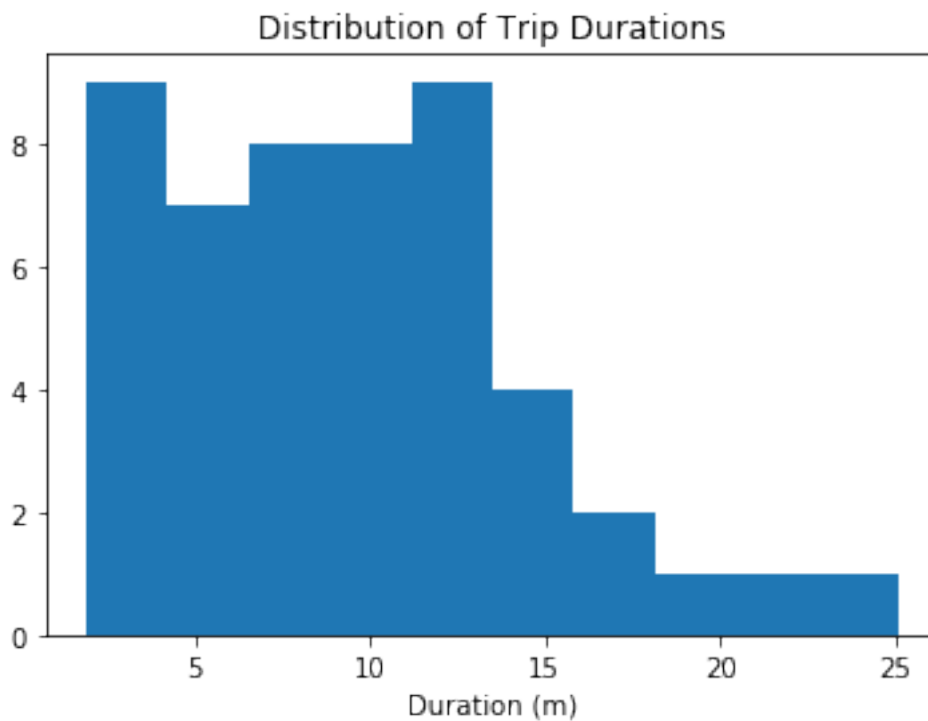
The last set of values that you computed should have pulled up an interesting result. While the mean trip time for Subscribers is well under 30 minutes, the mean trip time for Customers is actually *above* 30 minutes! It will be interesting for us to look at how the trip times are distributed. In order 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 [26]: # load library
import matplotlib.pyplot as plt

%matplotlib inline
import numpy as np
import pandas as pd

# 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, 14.33,
        19.02, 21.53,  3.90,  7.97,  2.62,  2.67,  3.08, 14.40, 12.90,  7.83,
        25.12,  8.30,  4.93, 12.43, 10.60,  6.17, 10.88,  4.78, 15.15,  3.53,
         9.43, 13.32, 11.72,  9.85,  5.22, 15.10,  3.95,  3.17,  8.78,  1.88,
         4.55, 12.68, 12.38,  9.78,  7.63,  6.45, 17.38, 11.90, 11.52,  8.63,]

plt.hist(data)
plt.title('Distribution of Trip Durations')
plt.xlabel('Duration (m)')
plt.show()
```

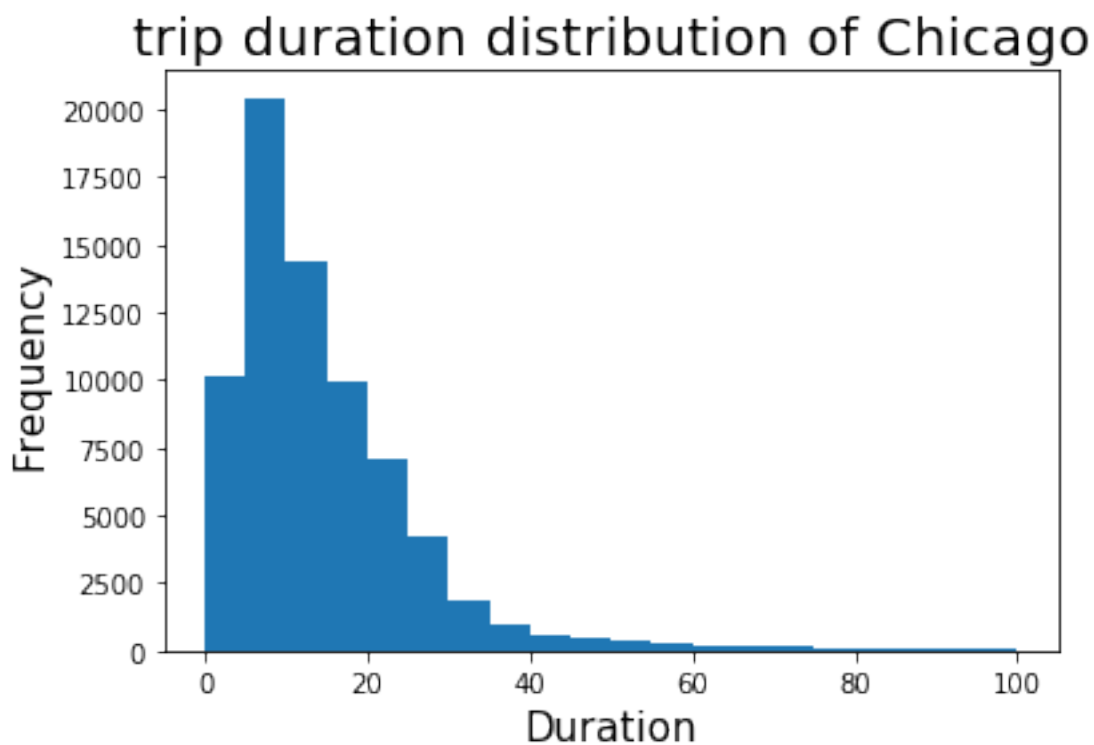


In the above cell, we collected fifty trip times in a list, and passed this list as the first argument to the `.hist()` function. This function performs the computations and creates plotting objects for generating a histogram, but the plot is actually not rendered until the `.show()` function is executed. 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 [27]: #trip duration distribution histogram of Chicago city.
data_file = pd.read_csv('./data/Chicago-2016-Summary.csv')
data_file = data_file['duration']

plt.hist(data_file, bins=20, range=(0, 100))
plt.title('trip duration distribution of Chicago', fontsize=20)
plt.xlabel('Duration', fontsize=15)
plt.ylabel('Frequency', fontsize=15)
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

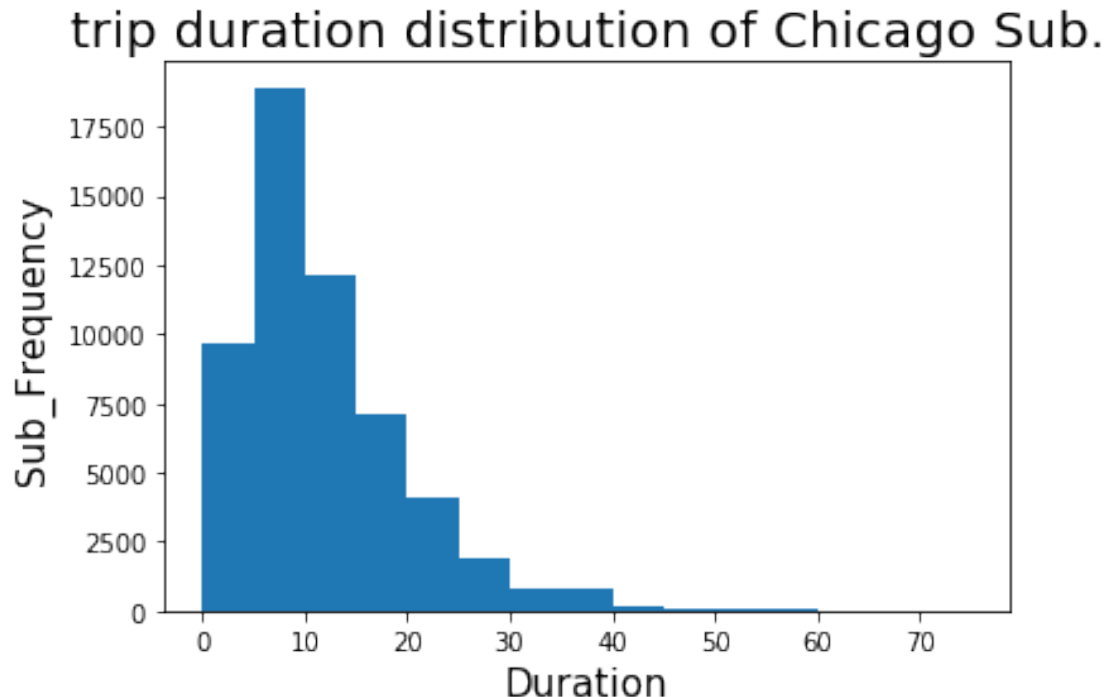
parameters with the `.hist()` function to limit the range of data that is plotted. Documentation for the function can be found [\[here\]](#).

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: For **Subscribers** histogram, the pick is between 5 to 10 minutes. For **Customers** histogram, the pick is between 20 to 25 minutes. **Shape:-** Subscribers distribution is **Right skewed(positive skewness)**. Customers distribution looks like normal distribution but is also skew positively.

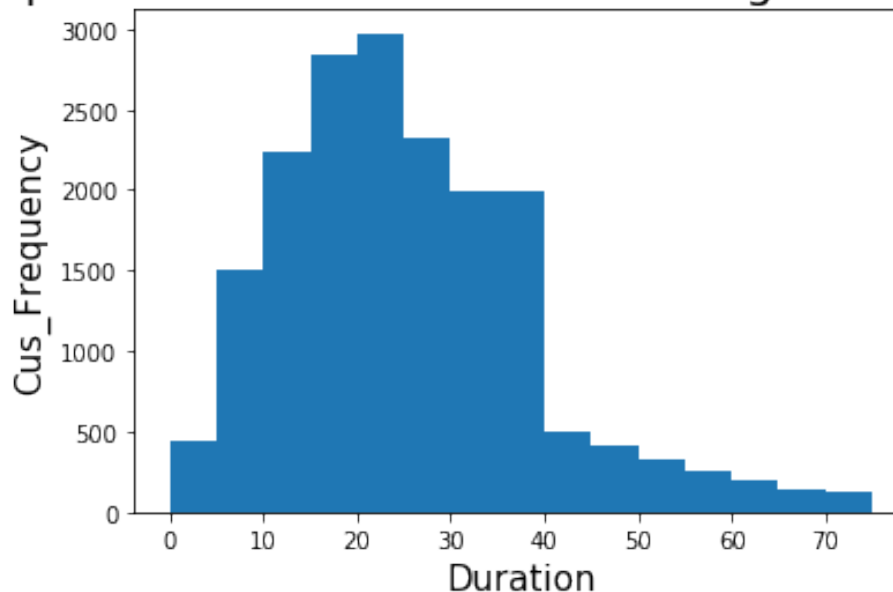
```
In [28]: data_file = pd.read_csv('./data/Chicago-2016-Summary.csv')
         #print(data_file)

In [29]: #trip duration distribution of trip made by subscribers in Chicago city.
         d = data_file[data_file.user_type == 'Subscriber'].duration
         bin_edges = [0, 5, 10, 15, 20, 25, 30, 40, 45, 50, 55, 60, 65, 70, 75]
         plt.hist(d,bin_edges)
         plt.title('trip duration distribution of Chicago Sub.', fontsize = 20)
         plt.xlabel('Duration', fontsize = 15)
         plt.ylabel('Sub_Frequency', fontsize = 15)
         plt.show()
```



```
In [57]: #trip duration distribution of trip made by customers in Chicago city.
d1 = data_file[data_file.user_type == 'Customer'].duration
bin_edges = [0, 5, 10, 15, 20, 25, 30, 40, 45, 50, 55, 60, 65, 70, 75]
plt.hist(d1, bins = bin_edges)
plt.title('trip duration distribution of Chicago Customers', fontsize = 20)
plt.xlabel('Duration', fontsize = 15)
plt.ylabel('Cus_Frequency', fontsize = 15)
plt.show()
```

trip duration distribution of Chicago Customers



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.

Q6. This question includes four question:- **1.**In which day of week maximum trips made in each city? **2.**How day of week influence trip duration of Subscribers and Customers in each city? **3.**Which city has highest average duration? **4.**In which month maximum trips made in each city?

```
In [30]: data_file1 = pd.read_csv('./data/Washington-2016-Summary.csv')

        data_file2 = pd.read_csv('./data/Chicago-2016-Summary.csv')

        data_file3 = pd.read_csv('./data/NYC-2016-Summary.csv')
```

1.In which day of week maximum trips made in each city?
Washington.

```
In [31]: print(pd.value_counts(data_file1['day_of_week']).head(1))

Wednesday    10103
Name: day_of_week, dtype: int64
```

Chicago.

```
In [32]: print(pd.value_counts(data_file2['day_of_week']).head(1))

Monday       11286
Name: day_of_week, dtype: int64
```

NYC.

```
In [33]: print(pd.value_counts(data_file3['day_of_week']).head(1))

Wednesday    44629
Name: day_of_week, dtype: int64
```

Answer(1):- In **Washington** maximum trips made in **Wednesday**. In **Chicago** maximum trips made in **Monday**. In **NYC** maximum trips made in **Wednesday**.

2.How day of week influence trip duration of Subscribers and Customers in each city?
Washington.

```
In [34]: data_file1.groupby(['user_type', 'day_of_week'])['duration'].mean()
```

```
Out[34]: user_type  day_of_week
        Customer  Friday      40.715875
           Monday      40.752832
           Saturday     44.589926
           Sunday      43.429348
           Thursday     39.685053
           Tuesday      41.777629
           Wednesday     36.370252
        Subscriber Friday      12.171476
           Monday      12.306999
           Saturday     13.094494
           Sunday      12.951062
           Thursday     12.523436
           Tuesday      12.392428
           Wednesday     12.564929
        Name: duration, dtype: float64
```

Chicago.

```
In [35]: data_file2.groupby(['user_type', 'day_of_week'])['duration'].mean()
```

```
Out[35]: user_type  day_of_week
        Customer  Friday      31.502309
           Monday      30.510664
           Saturday     30.573563
           Sunday      31.606878
           Thursday     29.475214
           Tuesday      31.029496
           Wednesday     31.906180
        Subscriber Friday      11.722610
           Monday      12.132253
           Saturday     13.184543
           Sunday      13.231205
           Thursday     11.479284
           Tuesday      11.516129
           Wednesday     12.072781
        Name: duration, dtype: float64
```

NYC.

```
In [36]: data_file3.groupby(['user_type', 'day_of_week'])['duration'].mean()
```

```
Out[36]: user_type  day_of_week
        Customer  Friday      32.781385
           Monday      33.257251
           Saturday     33.215728
           Sunday      32.848197
           Thursday     31.388124
           Tuesday      33.443487
```

	Wednesday	33.798065
Subscriber	Friday	14.361409
	Monday	13.203220
	Saturday	14.828011
	Sunday	14.883411
	Thursday	13.287923
	Tuesday	12.904069
	Wednesday	13.170957

Name: duration, dtype: float64

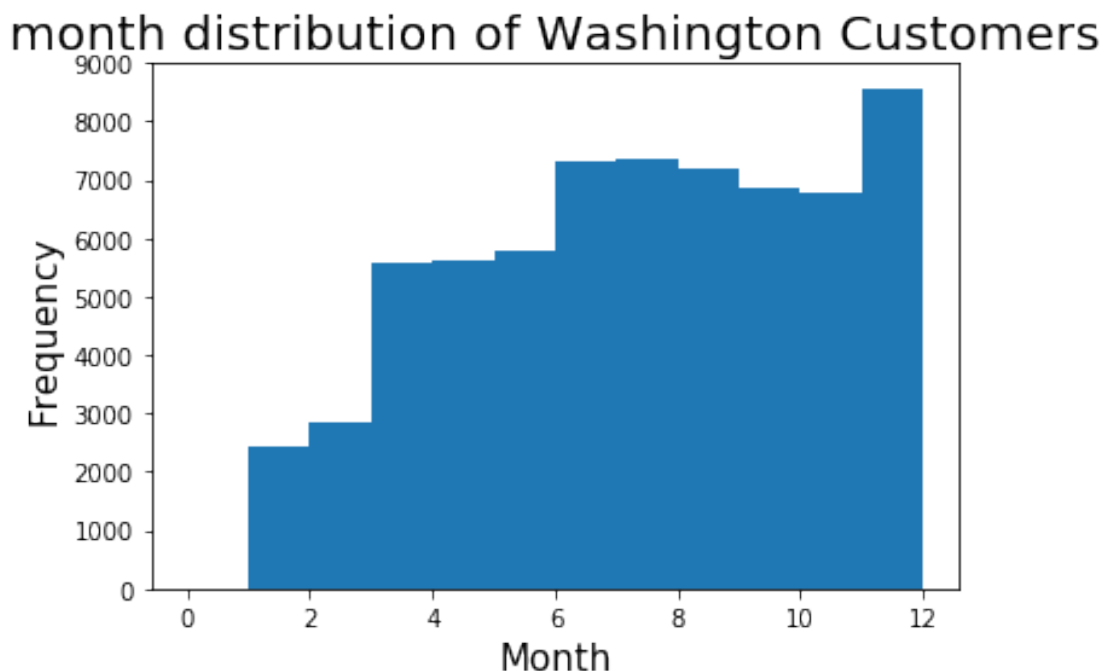
Answer(2):- In **Wahington** highest mean duration for Customers is **44.59(Saturday)** and for Subscribers **13.09(Saturady)**. In **Chicago** highest mean duration for Customers is **31.91(Wednesday)** and for Subscribers **13.23(Sunday)**. In **NYC** highest mean duration for Customers is **33.79(Wednesday)** and for Subscribers **14.88(Sunday)**. So, I can conclude that in all three cities maximum mean duration is almost high in **Weekend** means people spend more time in weekend

3.Which city has highest average duration?

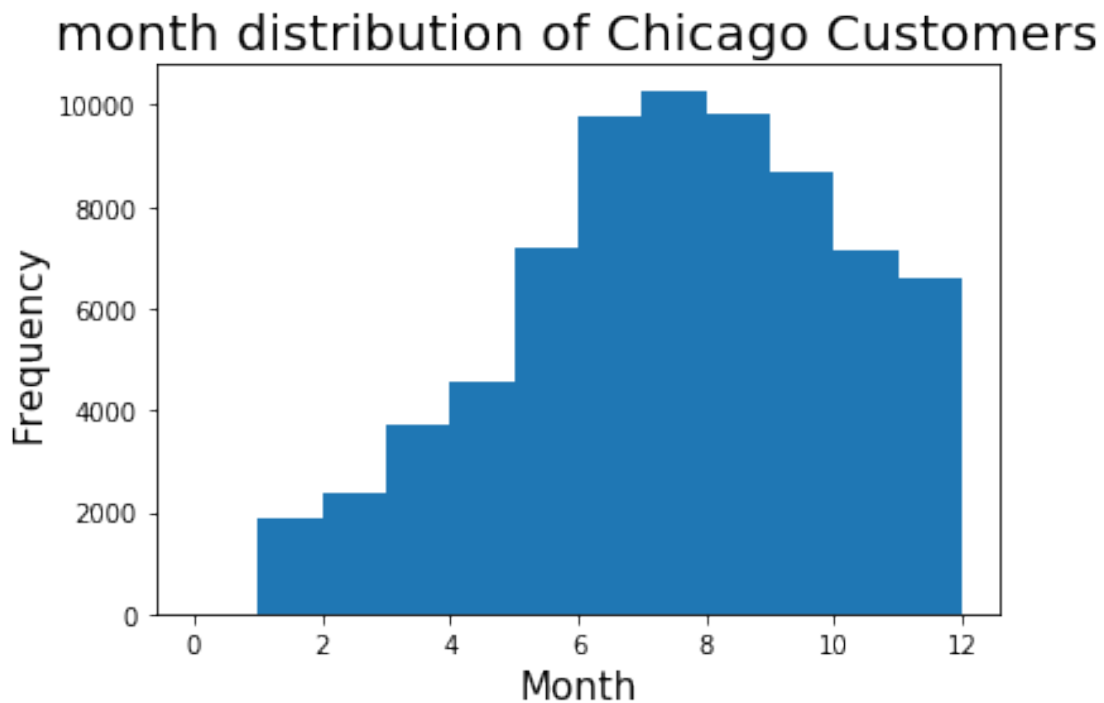
Answer:- Washington city has highest average of total duration.

4.In which month maximum trips made in each city?

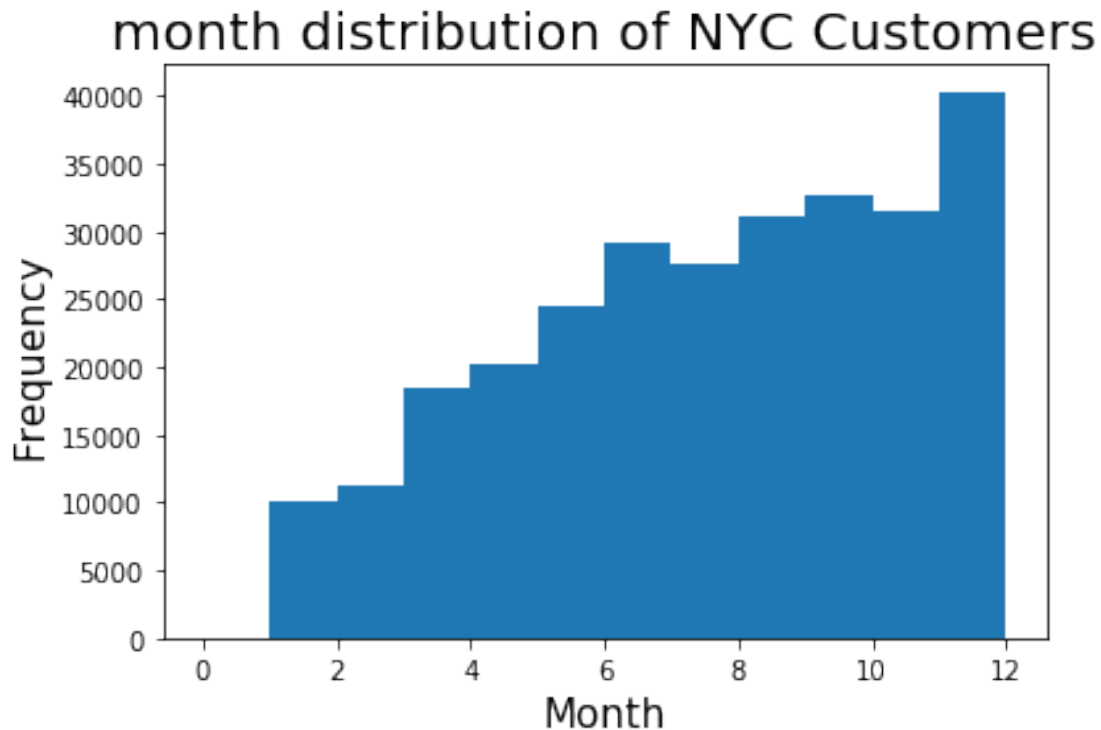
```
In [37]: d = data_file1['month']
bin_edges = [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]
plt.hist(d, bin_edges)
plt.title('month distribution of Washington Customers', fontsize = 20)
plt.xlabel('Month', fontsize = 15)
plt.ylabel('Frequency', fontsize = 15)
plt.show()
```



```
In [96]: d = data_file2['month']
bin_edges = [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]
plt.hist(d, bin_edges)
plt.title('month distribution of Chicago Customers', fontsize = 20)
plt.xlabel('Month', fontsize = 15)
plt.ylabel('Frequency', fontsize = 15)
plt.show()
```



```
In [38]: d = data_file3['month']
bin_edges = [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]
plt.hist(d, bin_edges)
plt.title('month distribution of NYC Customers', fontsize = 20)
plt.xlabel('Month', fontsize = 15)
plt.ylabel('Frequency', fontsize = 15)
plt.show()
```



Answer(4):- In Washington city, maximum trips made between **November-December(11-12)** months. In Chicago city, maximum trips made between **July-August(7-8)** months. In NYC city, maximum trips made between **July-August(7-8)** months.

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 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 want to explore **Libraries systems** of two different universities. In this exploration I would like to investigate that how examination influence book issues and which subject book issue most. I will also investigate that if students do not return books within time then how much fine they will have according to delay. Which university library has maximum books of machine learning.

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 Note-

book 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 [39]: from subprocess import call
         call(['python', '-m', 'nbconvert', 'Bike_Share_Analysis.ipynb'])
```

```
Out[39]: 0
```

Resoures:- 1.<https://www.youtube.com/watch?v=eirjyP2qcQ>.