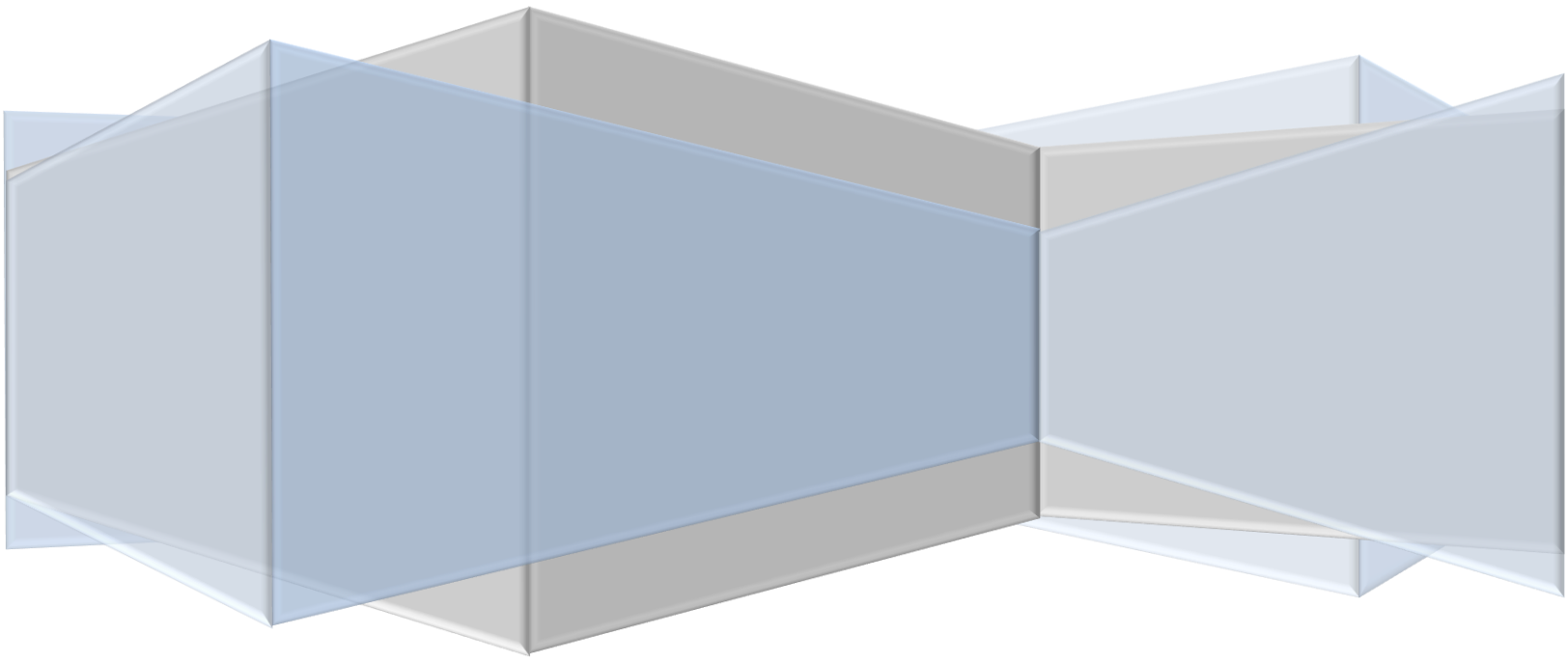
Rutgers Data Science Boot Camp

Matplotlib

Homework Assignment 05

Mark Visco



Assignment

Using the dependencies and data files (csv files) below build and display a Bubble Plot showing the relationship between four key variables:

- Average Fare ($) Per City
- Total Number of Rides Per City
- Total Number of Drivers Per City
- City Type (Urban, Suburban, Rural)

Additionally, three Pie Charts will also be created showing the following:

- % of Total Fares by City Type
- % of Total Riders by City Type
- % of Total Drivers by City Type

Based on the above data, provide a written description of three observable trends.

Methodology

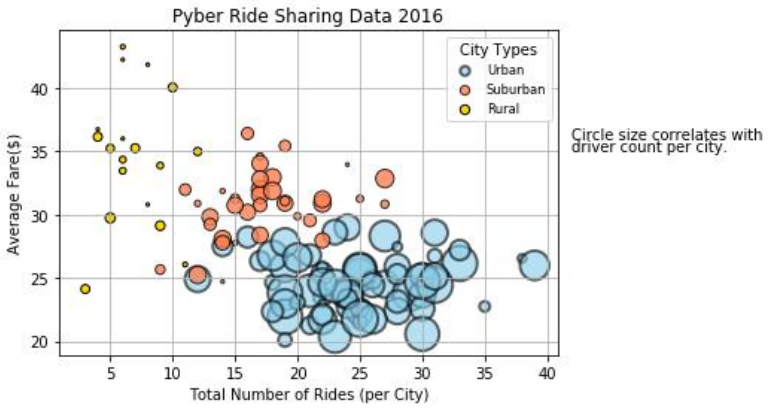
The first order of business was reading in the two csv files and determining a common data point (city name) so the two files could then be combined into one master file.

Out[1]:

	city	date	fare	ride_id	driver_count	type
0	Lake Jonathanshire	2018-01-14 10:14:22	13.83	5739410935873	5	Urban
1	Lake Jonathanshire	2018-04-07 20:51:11	31.25	4441251834598	5	Urban
2	Lake Jonathanshire	2018-03-09 23:45:55	19.89	2389495660448	5	Urban
3	Lake Jonathanshire	2018-04-07 18:09:21	24.28	7796805191168	5	Urban
4	Lake Jonathanshire	2018-01-02 14:14:50	13.89	424254840012	5	Urban

Plotting the Data

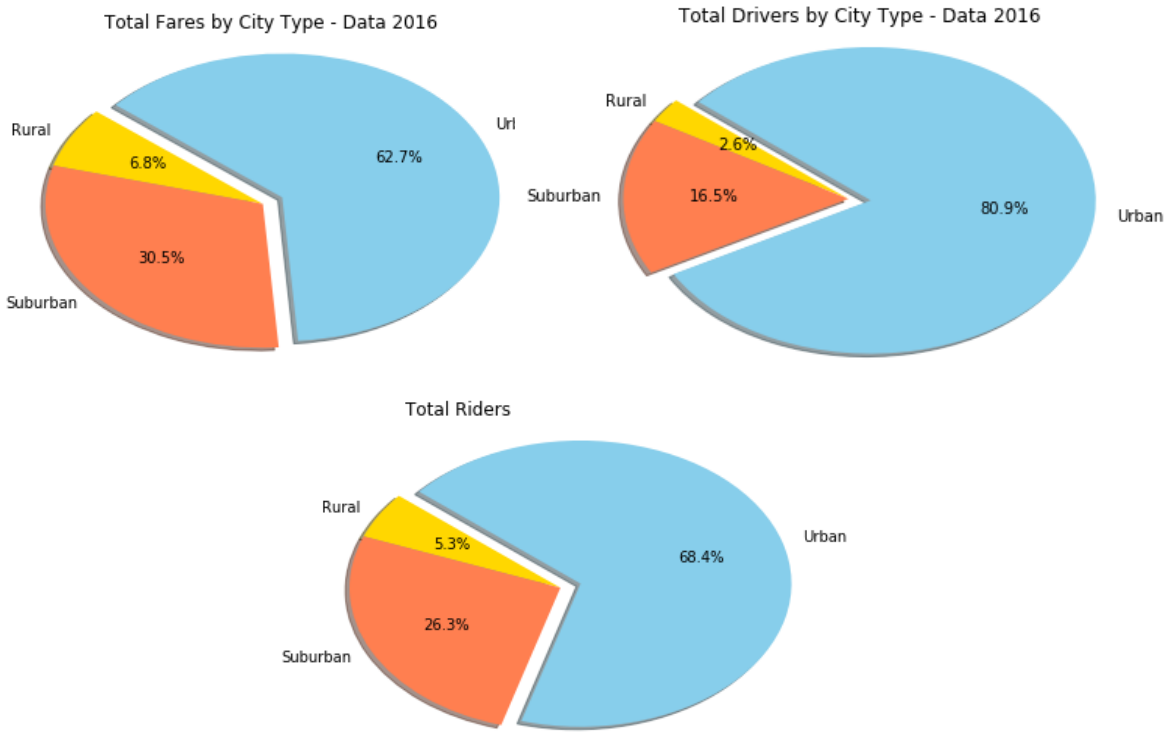
In order to create the bubble plot, three subset files (Urban, Suburban and Rural) needed to be created and extracted from the master file. For each subset, a scatter diagram was created then all three scatter diagrams were incorporated into one graphical representation.



This Bubble Plot is showing a lot of information in one graph. Each bubble represents the number of drivers per city. The size of the bubble correlates to the number of drivers per city. The color of the bubble differentiates between the type of city and the X axis of the graph represents the average fare per city and the Y axis indicates the number of trips per city.

In addition to the Bubble Plot, three pie charts were also created showing by city type (Urban, Suburban and Rural) a breakdown of the average fare, drivers and rides (or trips).

To generate the graphs below, three new and different subsets needed to be created. This time, instead of subsets by city type, we needed subsets by average fare, drivers and riders/trips. Once these were generated, the following pie charts were created.



Observations

At first glance your attention is drawn to the sections of the graph in blue as it easily dominates each graph. Upon further investigation and in my opinion, nothing remarkable is being depicted by these graphs. Urban areas by nature are more densely populated than rural areas and depending upon the size of the suburb, it could go either way. Therefore, it isn't any surprise the Urban pie sections are the largest.

Initially, the breakdown of the fares on the Bubble Plot seemed to show the opposite of what I would have expected, since urban areas are typically more costly then rural areas. I think I just contradicted myself! Upon further consideration, riders are most likely travelling shorter distances than in rural areas and that could be the reason for the lower average fare. However, if you factor in the element of time, then we will be looking at a different set of graphs and ones which I think will present a more meaningful picture. It wasn't until I started to analyze the results that the element of time became a consideration. Incorporating trip duration (mileage and time) into the data would prove extremely useful. Then you would be able to determine a cost per mile or cost per time period (i.e. ¼ hr., ½ hr., 1 hr.) which I think is a more realistic comparison.

For now, the graphs confirm urban areas have more riders, more drivers and will account for a larger portion of the total fares. However, introducing a time element might present a whole new dynamic to this assignment.