# [Week 4 Assignment](https://bbhosted.cuny.edu/webapps/assignment/uploadAssignment?content_id=_39523712_1&course_id=_1705333_1&group_id=&mode=view)

Centrality measures can be used to predict (positive or negative) outcomes for a node.

Your task in this week’s assignment is to identify an interesting set of network data that is available on the web (either through web scraping or web APIs) that could be used for analyzing and comparing centrality measures across nodes.  As an additional constraint, there should be at least one categorical variable available for each node (such as “Male” or “Female”; “Republican”, “Democrat,” or “Undecided”, etc.)

# Data

The data I’ve collected from web is

Chicago Divvy Bicycle Sharing Data

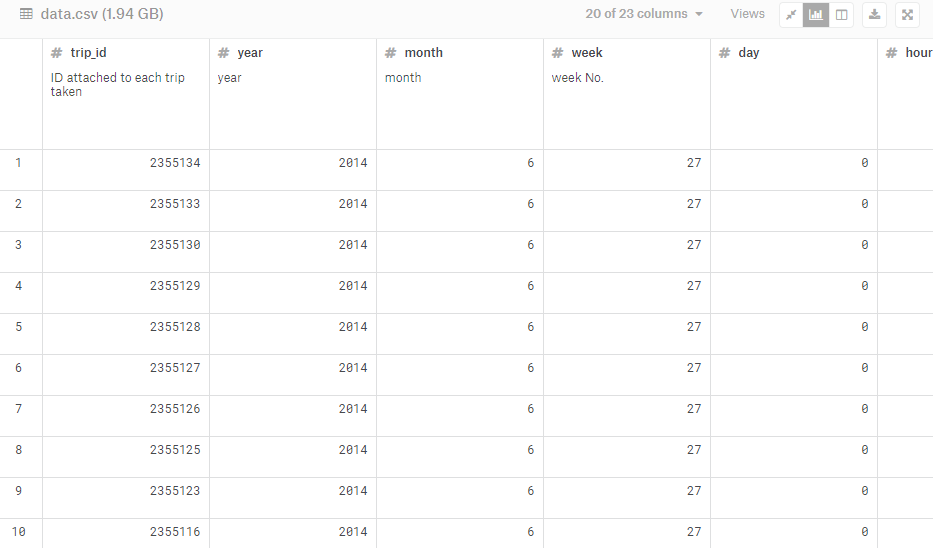
<https://www.kaggle.com/devisangeetha/divvy-bike-share-eda-network-analysis/data>

The data set has several attributes. From this data we can make a network of trips the riders did using from\_station\_name and to\_station\_name variable.

This data set satisfies our constraint of at least on categorical variable. Here our varible of interset is gender.

Using this variable has Customer and Subscriber category. We can compare between two categories of user using our centrality measure.

**Data Preview:**

****

In addition to identifying your data source, you should create a high level plan that describes how you would load the data for analysis, and describe a hypothetical outcome that could be predicted from comparing degree centrality across categorical groups.

# Loading the data

I’ll use python pandas libraries pd.read\_csv() function to load the data. Then I’ll subset the necessary 4 variables only gender, to\_station\_name,from\_station\_name. Due to large size of the data I’ve used excel to subset the data. Then loaded into python.

After this I’ll create a network from this data using networkx python library.

We’ll use degree of a node as a measure of centrality for our analysis.

From our data the hypothetical outcome :

Null Hypothesis: There is no significant difference between the degree measure of rides between two types of users

Alternative Hypothesis: There is significant difference between the degree measure of rides between two types of users

For this we’ll first find the degree of from\_station\_name, then we’ll do a t test to test the hypothesis.