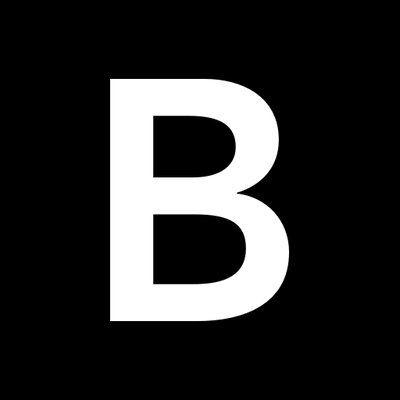
**loomberg Data Scientist Case Study**

****

**BACKGROUND**

Citibike is a bike sharing system in New York City that allows users to rent bicycles for short trips around the city. The system has been in operation since 2013 and has grown significantly in popularity over the years.   
  
As a data scientist, you will help Bloomberg employees pedal around town with a data driven approach! Help us analyze a sample of Citibike data from 2013 to 2016 and provide insights into usage patterns, detect potential outliers/anomalies, forecast future demand, and provide insights into bike stations based on the data.

**DATA**

The dataset provided includes information on Citibike trips taken between 2013 and 2016. The data includes information such as the start and end time of the trip, the starting and ending location of the trip, the duration of the trip, and other relevant information.

**TASKS**

Please choose 2 of the following focuses for your assignment. (Feel free to tackle them all for an extra challenge!)

1. Detection of potential outliers and anomalies for Citibike usage:

Using exploratory data analysis EDA and time series techniques, you should identify any potential outliers or anomalies in the Citibike data. This could involve identifying unusually high or low global usage patterns, individual trips that are significantly longer or shorter than average, or other patterns that stand out from the norm. Once potential outliers or anomalies have been identified, you should investigate the cause of these patterns and determine whether they are genuine or errors in the data.

2. Citibike usage forecast:

Using time series analysis and forecasting techniques, you should develop a model to forecast Citibike usage in the future. The model should take into account factors such as seasonality, trends, and other relevant variables that may impact Citibike usage. The model should be able to provide accurate predictions of future usage patterns based on your chosen metric of usage (e.g. trips taken per day, the number of bikes rented per day, etc. Your Choice!)

3. Insights into bike stations based on the data:

Using the Citibike data, you should provide insights into the usage patterns of different bike stations around the city.

**Anything else?**

We are happy to review additional insights you find interesting outside of above scope.

**DELIVERABLES**

Please generate the report within a **Jupyter Notebook** and upload it to your **GitHub**. Please share the GitHub repository link with us.  
  
  
The case study should include a report summarizing your findings and recommendations, as well as any code, visualizations, or other materials that were used in your analysis. The report should be structured in a way that is easy to understand for both technical and non-technical audiences and should clearly communicate the key insights and recommendations that were identified during the analysis.

Thank you and Good Luck!

## 

## 

## 

## **TABLE COLUMN DESCRIPTIONS**

##### 

##### **citibike-trips-sample.csv**

Tripduration - Trip Duration (in seconds)

Starttime - Start Time, in NYC local time.

Stoptime - Stop Time, in NYC local time.

Start\_station\_id - Start Station ID

Start\_station\_name - Start Station Name

Start\_station\_latitude - Start Station Latitude

Start\_station\_longitude - Start Station Longitude

End\_station\_id - End Station ID

End\_station\_name - End Station Name

End\_station\_latitude - End Station Latitude

End\_station\_longitude - End Station Longitude

Bikeid - Bike ID

Usertype - User Type (Customer = 24-hour pass or 7-day pass user, Subscriber = Annual Member)

Birth\_year - Year of Birth

Gender - Gender (unknown, male, female)

Customer\_plan - The name of the plan that determines the rate charged for the trip

##### **citibike-stations.csv**

Station\_id\_int - Unique identifier of a station.

Name - Public name of the station.

Short\_name - Short name or other type of identifier, as used by the data publisher.

Latitude - The latitude of the station. The field value must be a valid WGS 84 latitude in decimal degrees format.

Longitude - The longitude of a station. The field value must be a valid WGS 84 latitude in decimal degrees format.

Region\_id - ID of the region where the station is located.

Rental\_methods - Array of enumerables containing the payment methods accepted at this station.

Capacity - ANumber of total docking points installed at this station, both available and unavailable.

Eightd\_has\_key\_dispenser

Num\_bikes\_available - Number of bikes available for rental.

Num\_bikes\_disabled - Number of disabled bikes at the station.

Num\_docks\_available - Number of docks accepting bike returns.

Num\_docks\_disabled - Number of empty but disabled dock points at the station.

Is\_installed - Is the station currently on the street?

Is\_renting - Is the station currently renting bikes?

Is\_returning - Is the station accepting bike returns?

Eightd\_has\_available\_keys