**Data Collection:** For the sake of simplicity, I separate the dataset that I downloaded from Kaggle into 2 different categories: Lyft XL and UberX.

The dataset from Kaggle has 10 Features:

+**Distance**

+Cab\_Type

+Time\_Stamp

+Destination

+Source

+**Price**

+Surge Multiplier

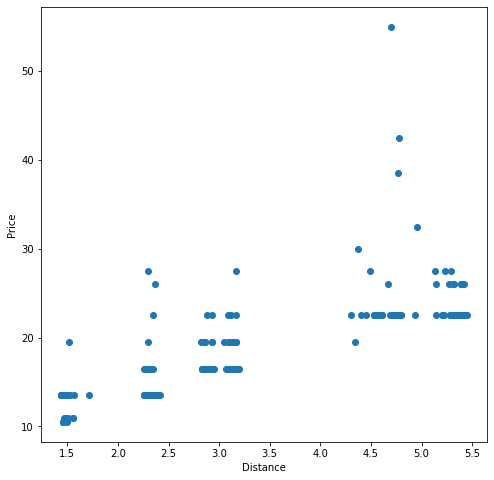
+ID  
+Product ID

+Company Name

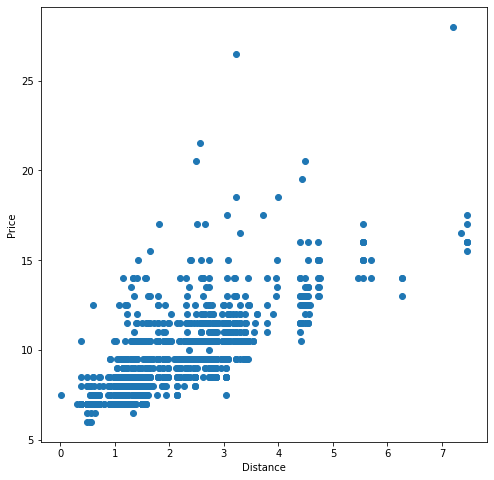
While other features affect very **little** to our Price Calculation, **Distance** plays an important role in determining our **Price**. Therefore, we reduce our dataset to 2 features: **Distance and Price**

The main focus of my analysis is the correlation between distance and price and the 2 companies. I take 300 data points for each model.

**LYFT XL**



**UBER X**

****

**Methods:** We used SciKit-Learn built-in classes for both unsupervised and supervised methods.

For an unsupervised method, we used DBScan for separating dataset into different clusters and identifying data points that don’t fit the model.

In order to clean the data and identify data points that don’t fit the model, we can filter out data points with ‘label’ equals ‘-1’, or in other words, data points that are recognized as “noises”.

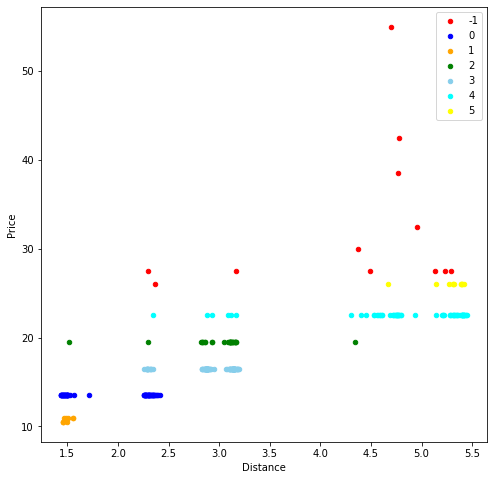
For the argument:

+ I set Epsilon as 1.5 using domain knowledge

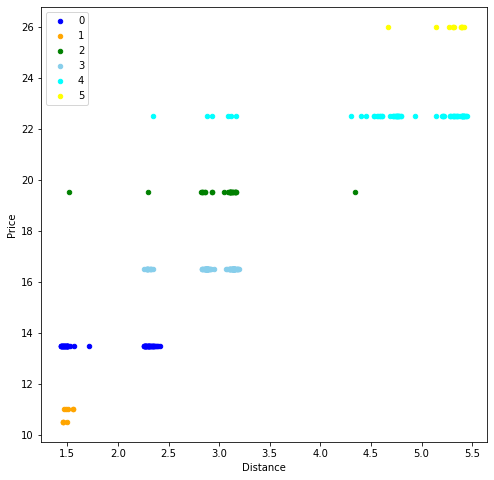
+ I set Minimum Points as 6 using domain knowledge

**Results and Discussion - (Metric):**

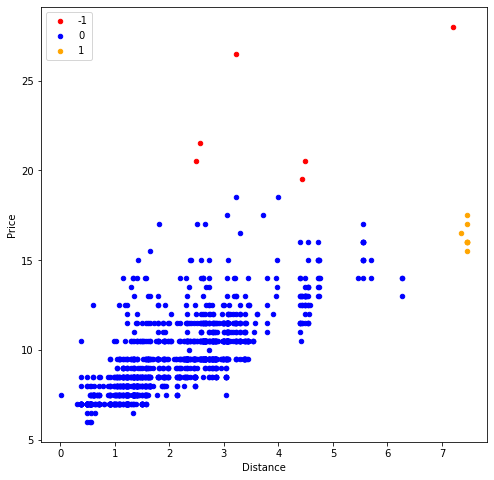
Lyft XL Dataset after Clustered:

****

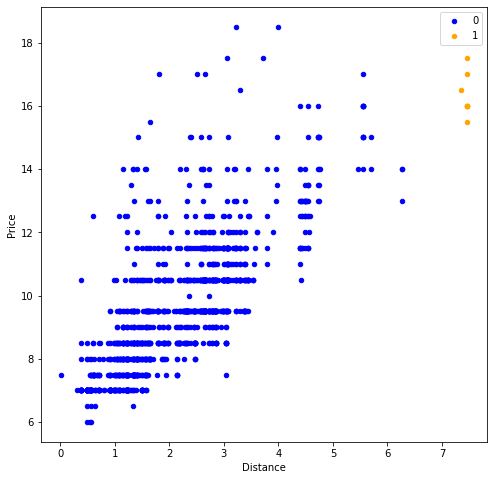
Lyft XL Dataset after Cleaned:



Uber X Dataset after Clustered:



Uber X Data after Cleaned:



**SCORE RESULTS**

Silhouette score Lyft XL **Before** cleaning data 0.8418462430662733

Silhouette score Lyft XL **After** cleaning data 0.8745290606295495

Silhouette score Uber X **Before** cleaning data 0.6591394025999412

Silhouette score Uber X **After** cleaning data 0.6605194682771542

DBScore score Lyft XL **Before** cleaning data 0.5355765875852511

DBScore score Lyft XL **After** cleaning data 0.18702255021113948

DBScore score Uber X **Before** cleaning data 0.4750692086398424

DBScore score Uber X **After** cleaning data 0.2789933018221577

**Further Improvement**

* Our model can further improved my normalized our dataset (Since 1.5 Eps for **Distance** has different impact then 1.5 Eps in **Price**)
* Using Elbow Method to determine the optimal value of Epsilon.