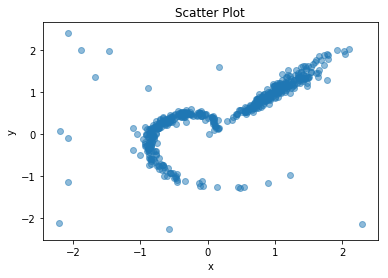
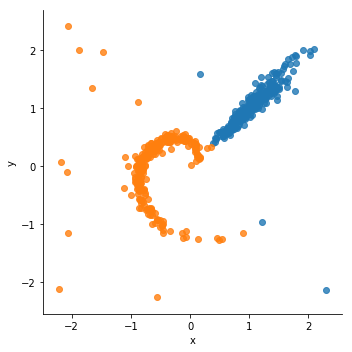
**Step 1: Read Data**

Read the csv data to python and generate the scatter plot. Rename two columns as x and y.



**Step 2: K-means Clustering**

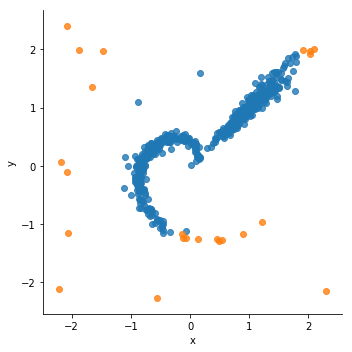
The scatter plot shows there are two distinct groups following a linear trend (Group 1) and a curved trend (Group 2). This pattern is a centroid-based clustering, so I used k-means methodology with parameter 2. Staring from the initial points, whenever a new data point is iterated, I perform a 2-step process of (1) clustering based on distance to group center and (2) updating of group center.



**Step 3: Identify outliers**

The distance to respective group center is used as a measure of outlier. I used a standard IQR (Interquartile Range Rule) to decide whether a data point is an outlier. Denote as 25% percentile, and as 75 percentile. The upper bound is and the lower bound is .

Applying this rule, the outliers are shown in plot below in orange.



However, this method has two drawbacks:

1. The distance is always greater than 0, and outliers are data points with large distance. The two-sided IQR is not quite suitable for such situation.
2. From the plot, it’s clear that some points are misclassified as outliers. Especially for the spiral shaped Group 2, some points at the far end belong to the trend but are far from the center.

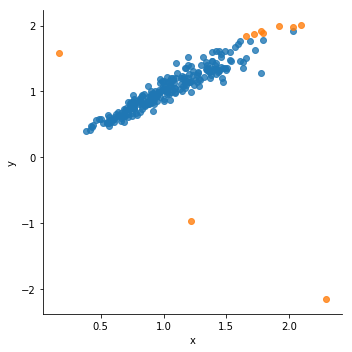
So I decided to look at each group individually.

**Step 4: Group 1**

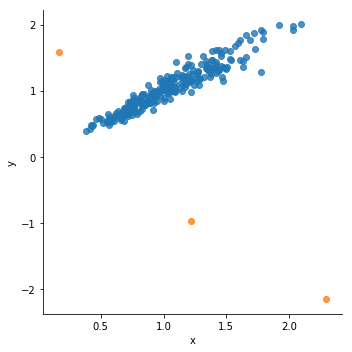
The first group is the blue points on the k-means plot. Note that x and y has linear relationship, I used Cook’s distance, which is commonly used in linear regression to find outliers or influencers. Cook’s distance is calculated as



The general rule of thumb is 4/(n-k-1), which is 0.016 in our data. Using this threshold, the outliers are identified in the plot below.



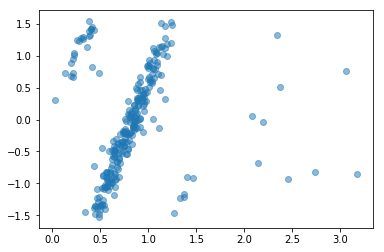
Note that some data points belong to the group are also classified as outliers, though their cook distance is much smaller than the three obvious outliers. So I increased the threshold a bit to 0.05, and obtained the following outlier plot. Now those points are no longer above the threshold.



**Step 5: Group 2**

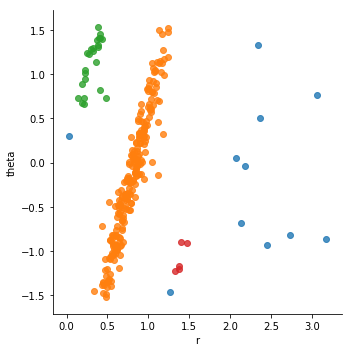
For Group 2, x and y have a spiral trend from [0, 0] to outside. At first, I was considering some kind of transformation like PCA or kernel function. I searched online and found there’s a mathematical representation and in a Cartesian coordinate system a spiral can be described by two parametric equations:

So, I transformed x and y to r and theta, and obtained the following scatter plot for r and theta:

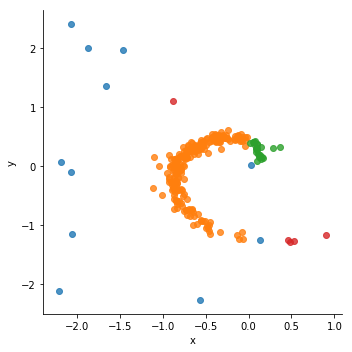


From this plot, it’s clearer than x and y plot and it shows several separate groups. In this case, a density-based clustering is better than centroid-based clustering since the shape is long and narrow. There are two parameters for density-based clustering, the radius and the minimum points of neighborhood. By looking at the red circled part, they seem to follow a pattern but only have a few points. I set the minimum points to 5 and adjust and used visualization to determine the final parameter choice. When is 0.3, the clustering result is shown as below.

There are three groups in green, orange and red. The blue ones are outliers. It looks reasonably well.



Plot in x and y space. There’s one red point far from the rest, but they still belong to the same group. Also, two blue points are close to the main stream, but they are outliers.



**Step 6: Combine two groups**

The final step is just to combine together Group 1 and Group 2 results and output the outliers. 14 outliers are identified.

