**K-means Clustering in Tableau Visualizing Custom Sales territory based on the Analysis**

Top of Form

**K-means algorithm is used for Clustering in Tableau 10.0. Let us understand the Definition of “Clustering” and some details about the algorithm “K-means” in its simplest form so that we clearly understand what we are trying to achieve.**

**Clustering** is the partitioning of a data set into subsets (clusters), so that the data in each subset share some common trait.

**K-means:** For a given number of clusters say “K” the algorithm partitions the data into “K” clusters. Each Cluster has a center (centroid) that is the mean value of all the points in that cluster.

If you are in interested in understanding the inner mechanism of the details of the K-means algorithm that Tableau uses, you can go through the user guide. In its simplest form you can understand that Tableau uses the Calinski-Harabasz criterion to assess cluster quality. The greater the value of this ratio, the better the cluster. So If a user does not specify the number of clusters, Tableau picks the number of clusters corresponding to the first local maximum of the Calinski-Harabasz index. By Default, k-means will be run for up to 25 clusters if the first local maximum of the index is not reached for a smaller value of k. The users can set value to a maximum of 50 for the number of cluster.

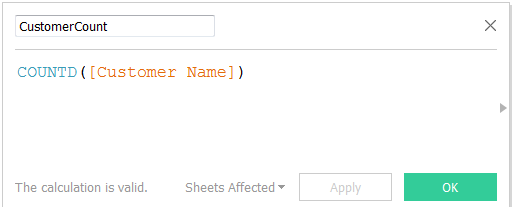
**Hands On Example:**

Here we will try to create a use case with sample superstore as we all know this data very well and we do not want to waste time understanding a new data set. That being said the example that we go though is easily implementable on any other data set as the concept remains same.

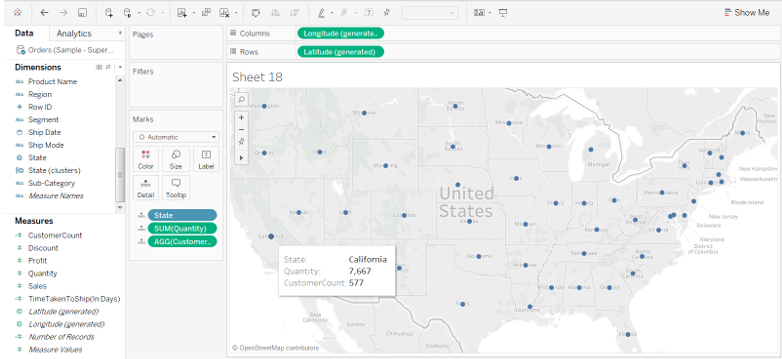
With sample super store data set, we want to do some quick resource planning. We basically want to figure out how many Sales Person do we need to place for the Sales Territory that we define. So in order to define Sales Territory we need to find out the cluster of states that share some common trait and based on which we can plan our resources.

In this example the feature that define the common trait would be the “Total Number of Customer” and the “Total Quantity Being Sold"

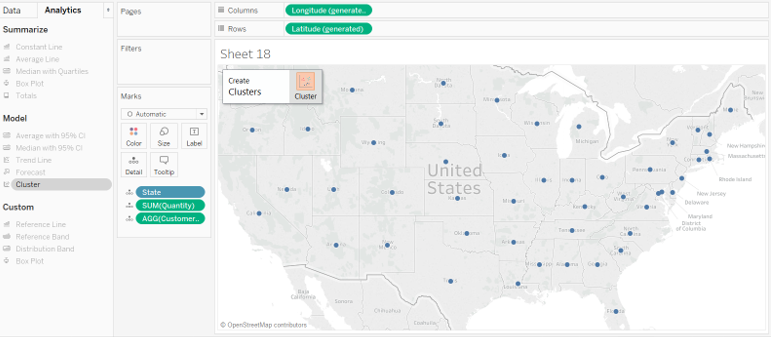
We already have the measure “Quantity”. In order to know the total number of distinct Customer, we create a calculated field “CustomerCount”



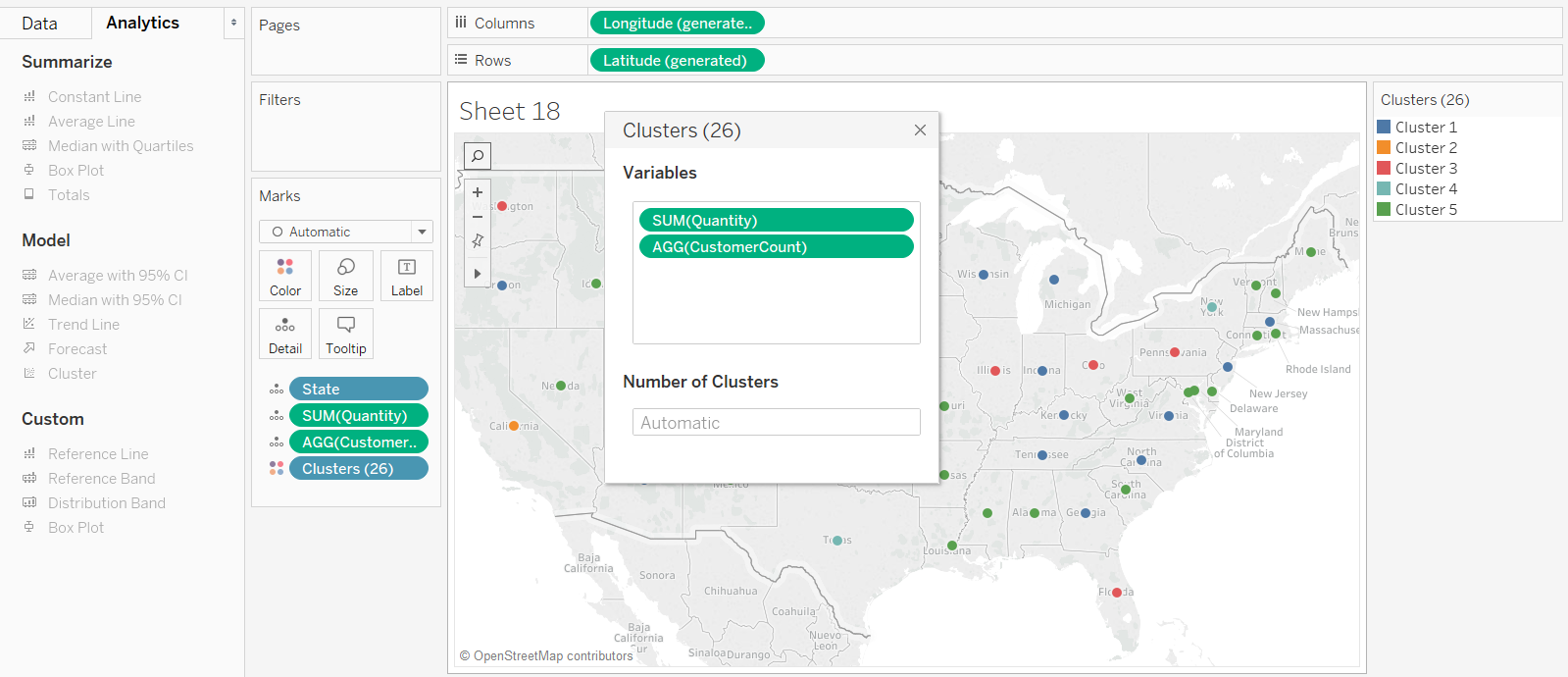
Now we drag the “States” on the Map and drag “Customer Count” and “Quantity” to the “details” marks



Next go to the Analytics (tab) next to “Data” and drag and Drop “Clusters” to the visualization.

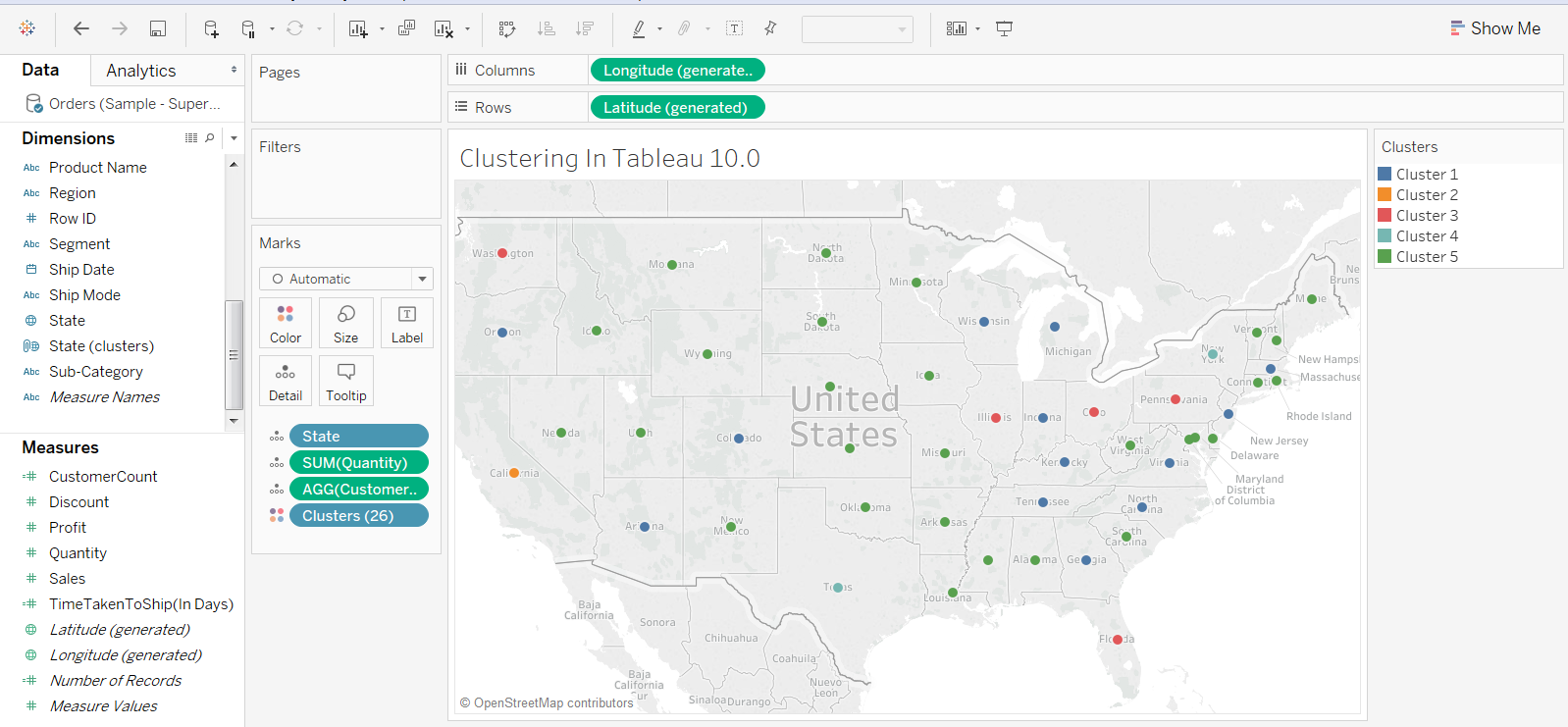


This will automatically pop up the below dialogue box with the variable “Sum(Quantity)” and “Agg(CustomerCount)” in the Variables [ *remember we had put these two fields in details]* and these variables are being used by Tableau to compute clusters. We can add additional variables to it, if we want more of them to be used for computing cluster.

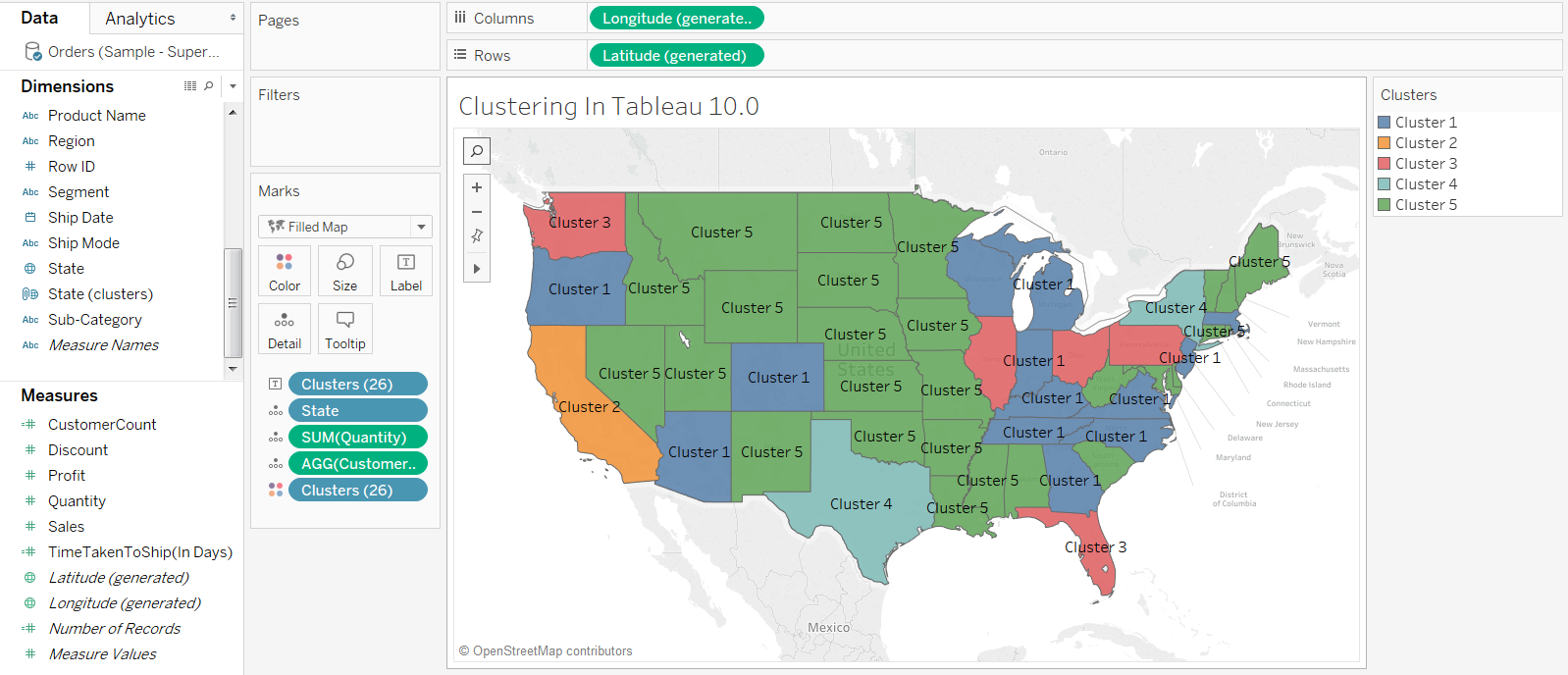
Also in the background Tableau creates 5 clusters on the “color” shelf and mark the different states with different color accordingly. Please note since user has not specified the number of clusters, Tableau picks the number of clusters as -5 corresponding to the first local maximum of the Calinski-Harabasz index.

If we want to identify more number of clusters in our data, we can provide a value based on our requirement and less than 50. We will leave it to 5 for now for our further analysis.Also, in this example I have used “Quantity” and “Customer Count” as variable to compute the cluster. There is no guarantee that these are the ideal fields to be selected. Clustering is an iterative process of Analytics I.e. Experimentation leads to Discovery leading to more experimentation.

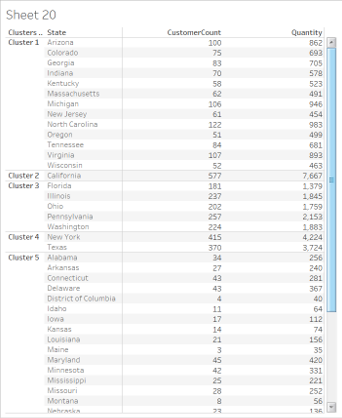
So we get the below visualization on point Map for the 5 Clusters that Tableau Identified based on the Quantity Ordered and the Number of Customers variable.



We can change it to “Filled Map” as my personal favorite and put Clusters on Label to identify the states belonging to different cluster.

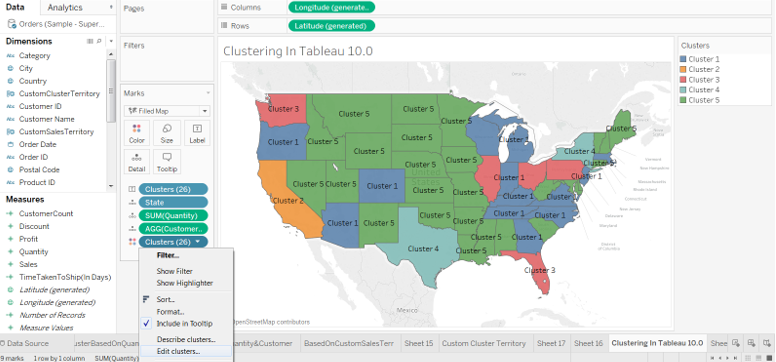


To do further analysis on the Cluster that tableau provided, you can generate a cross tab of the data and finalize if it suits your need or you would like to make some changes.

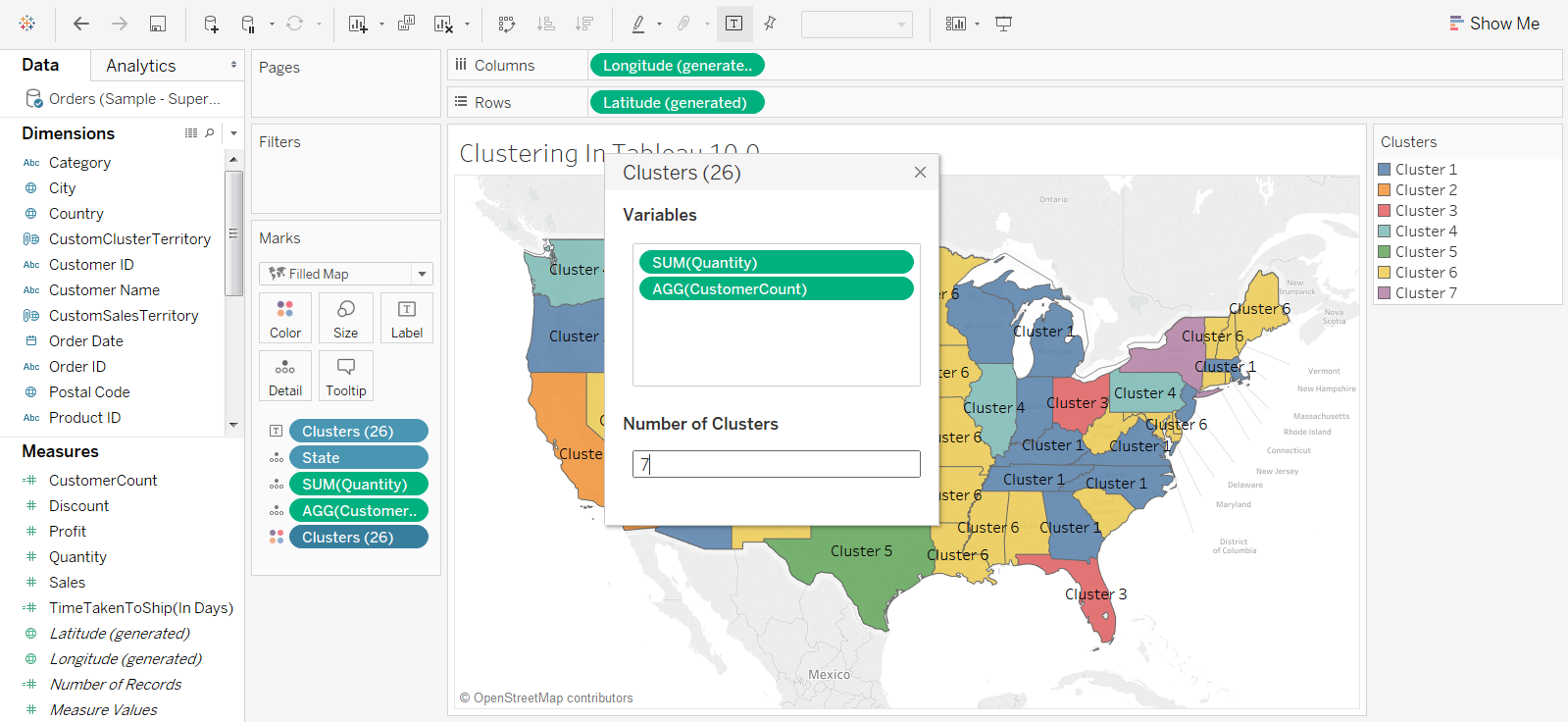


So as we can see California is kept singly in “Cluster2”  ,while “New York” and “Texas” are placed in “Cluster4” . Based on our analysis if we decided , it would be better to have 7 Clusters then we can go ahead and edit the “Cluster”

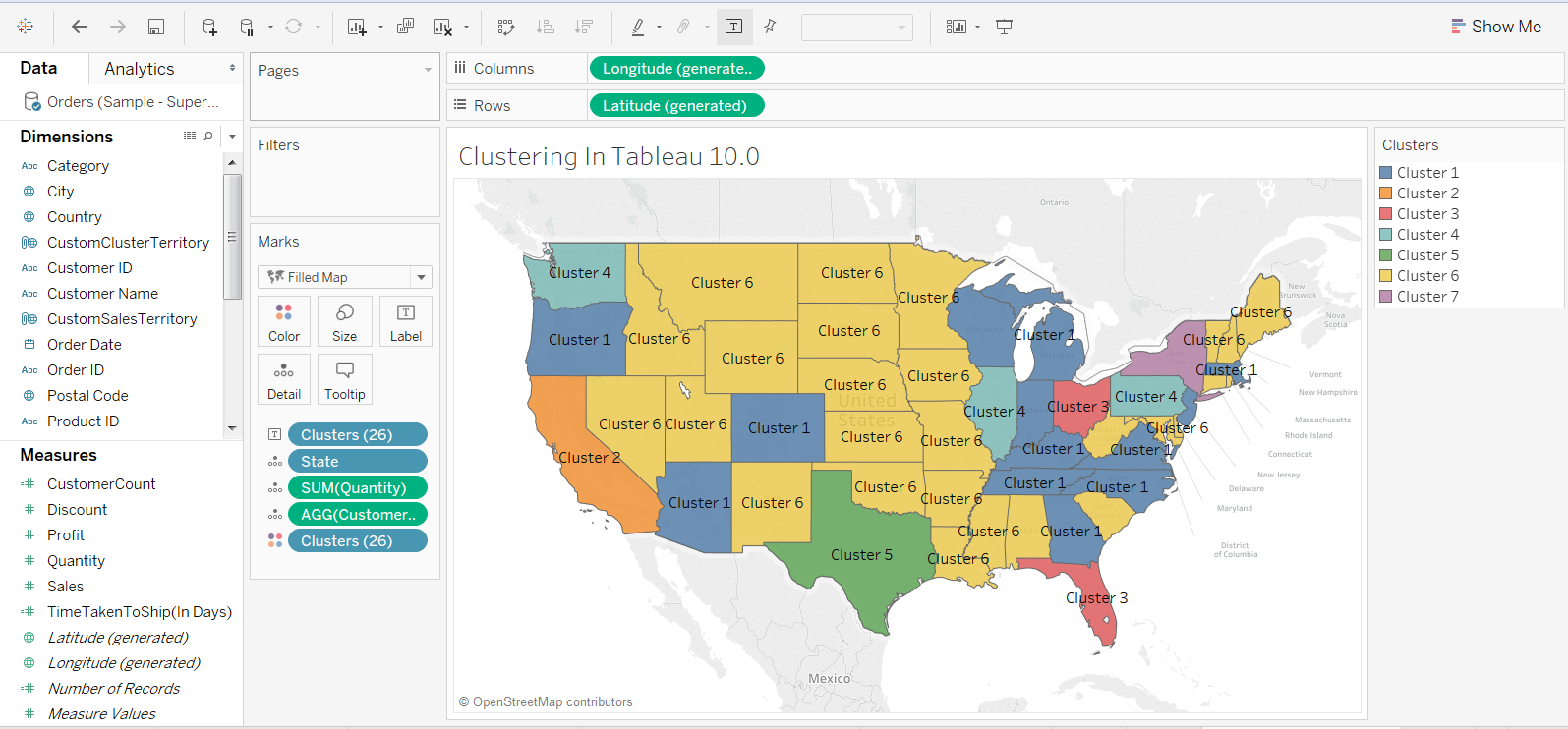
Click on the Clusters on the “Color Marks” shelf and select “edit clusters”



Enter Number of Clusters as “7” and you can see 7 Clusters getting created in the background. Click the “X” button remove the Clusters screen.



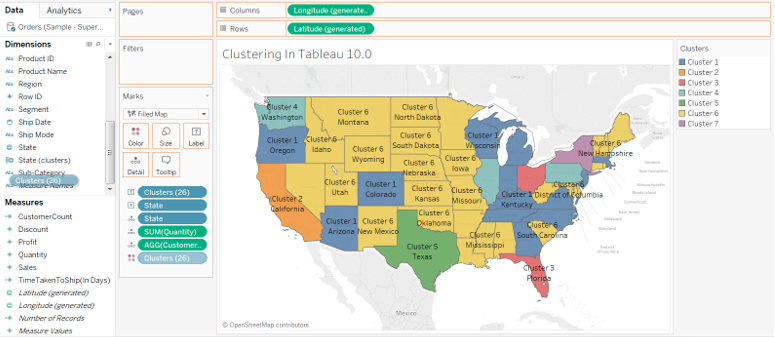
And you can see 7 clusters now

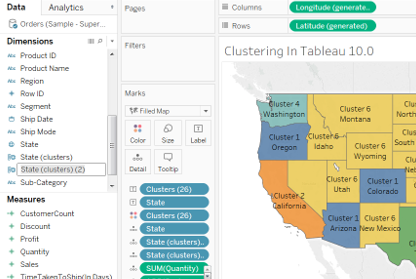


We can do further analysis to finalize the cluster structure. Once this gets finalize we can create these 7 Clusters as “Custom Territory on Map” (Another new feature Introduced in tableau 10.0).

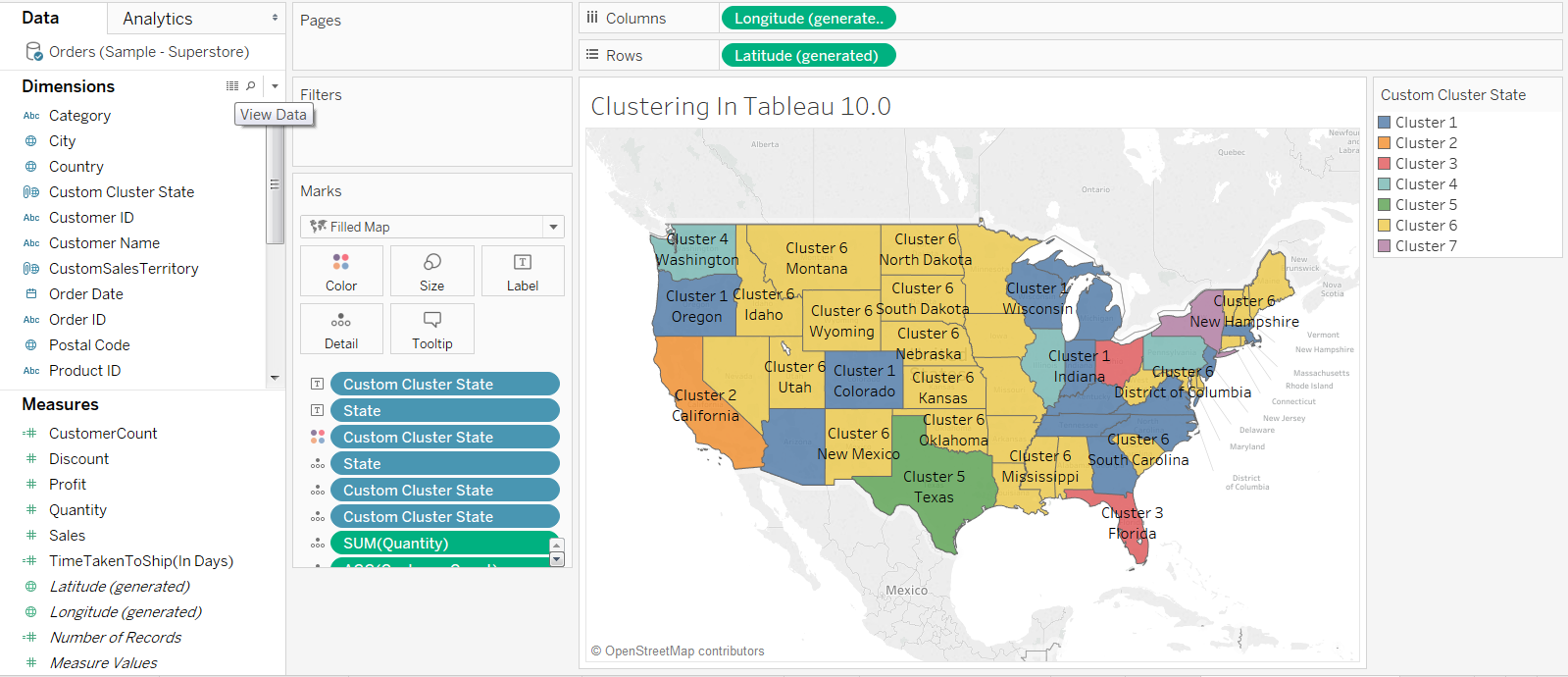
**Convert “Cluster” into “Custom Sales Territory”**

Drag and drop the Cluster from Marks Sheet to Dimensions

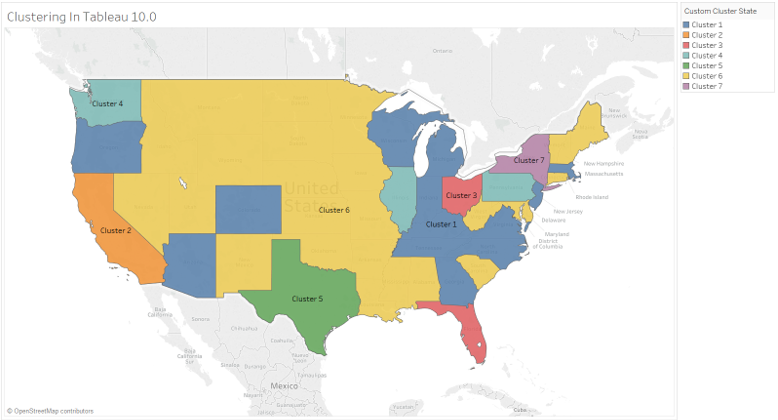




You can rename this as “Custom Cluster State” and Drag and Drop it on Marks shelf replacing the earlier cluster.



After that remove the “state” field from the Marks and you get the Custom Sales Territory define as per the K-Means Cluster that you identified.



Now you can use these custom sales territory for your resource planning and also have a quick look on the sales / profit that happened across these clusters to further strengthen your analysis.

