**Bill Yerkes**

**CS5542 Big Data Apps and Analytics**

**In Class Programming –7**

**8th October 2020**

**Submit ICP Feedback in Class. :** [**Lnik to Feed back Form**](https://docs.google.com/forms/d/e/1FAIpQLSesllFh5_STnj7RbHyQainRG_2EIKw1csp8ObP5FWjpVnGVOg/viewform)

**K-means clustering:**

**Use a different data and use the model provided in ICP7 to perform clustering. You must try 5 different number of clusters (for example n\_clusters= 5 or n\_clusters=6,7,8,or 9 etc) based on elbow curve and for each cluster visualize the clustering results and report your findings in detail.**

ICP Requirements:

1. Successfully executing the code and trying 5 different number of clusters based on elbow curve and visualizing those clusters using python plotting libraries (75 points)
2. Using a new and good dataset (5 points)
3. overall code quality (10 points)
4. Pdf Report quality, video explanation (10 points)

Submission Guidelines:

Same as previous ICPs.

ICP Report:

**What I learned in the ICP:**

I learned about data clustering. I was able to put into practice what we went over in the lecture on Tuesday. I practice with different data sets, and settled on a data set of latitude and longitude information about craiglist car sales. I also saw that when the data is connected it is hard to see how the different clusters will form, but after they are formed you can see how the groups are related and that there is separation in the data that is not initially visible.

**Description of what task I was performing:**

Process new data set, normalize the data, and find the “elbow” to determine the optimal number of clusters.

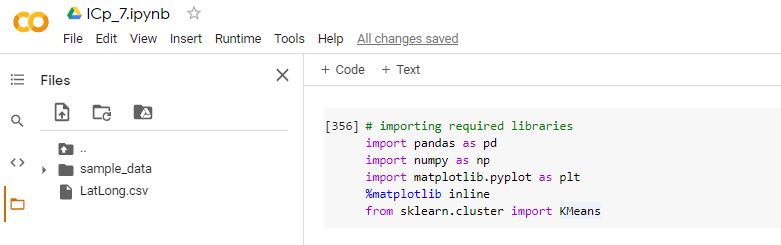
**Challenges I faced:**

Finding a data set to work with. Cleaning the data and removing rows which caused the different steps to fail.

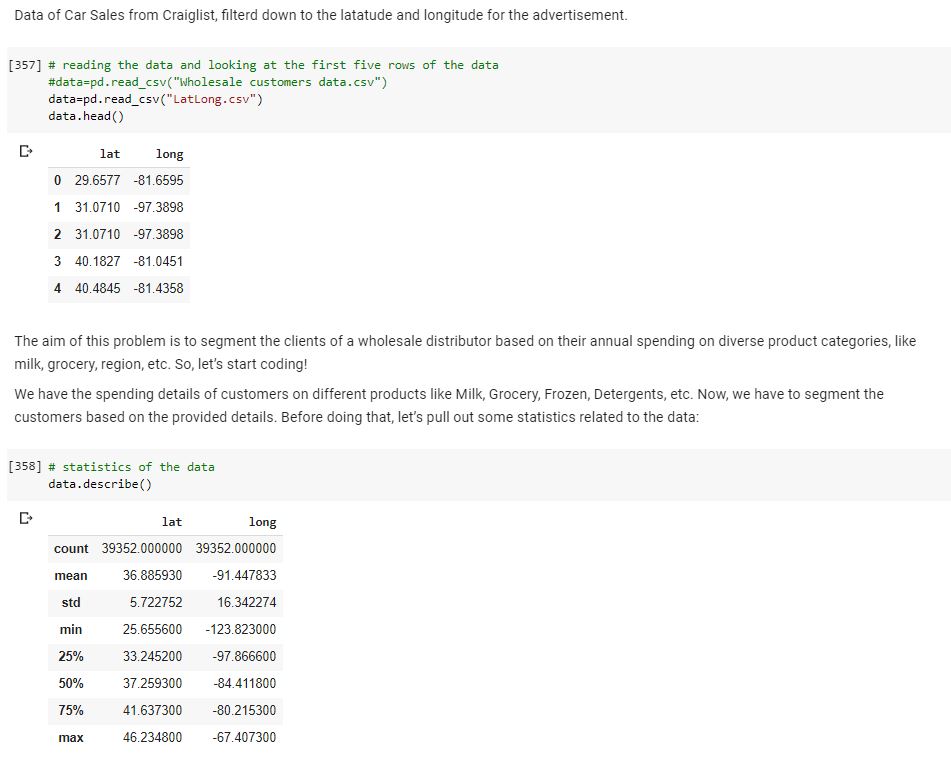
**Screen Shots**

**GitHub:**

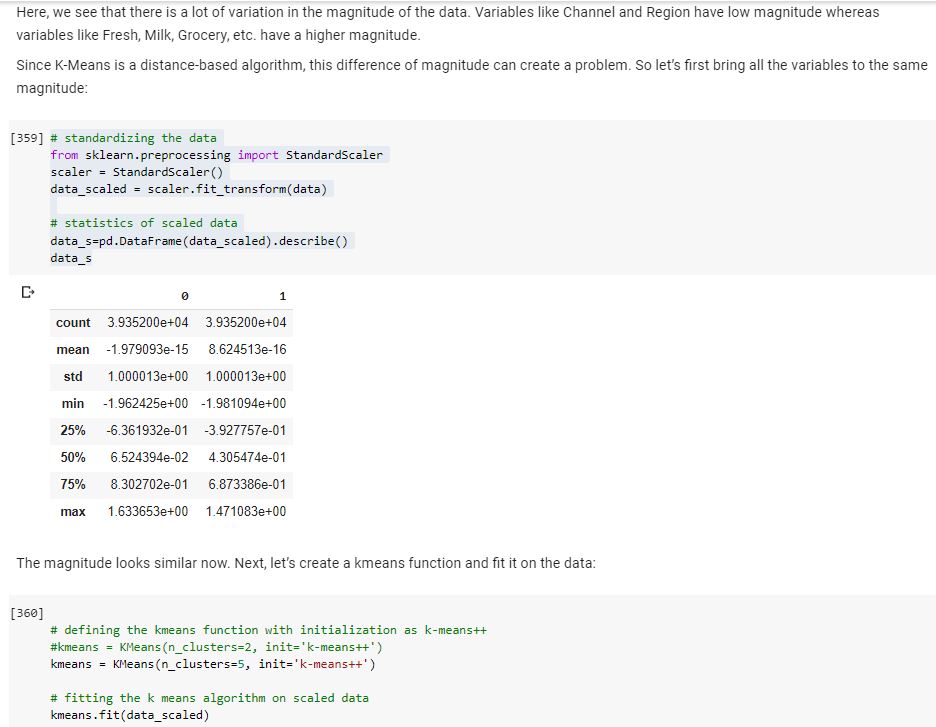
### **Initialize and Install Libraries**



### **Load Data, Describe Data**

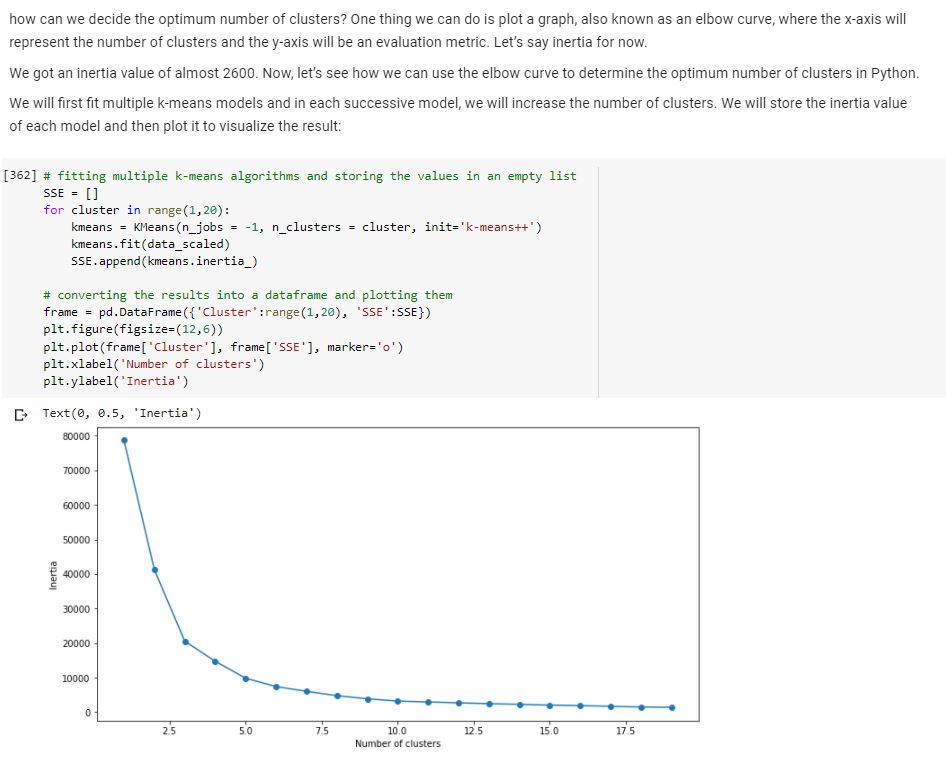


**Standardize the Data**

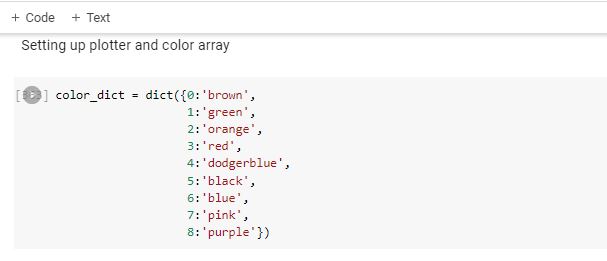


**Inertia on the Fitted Data**

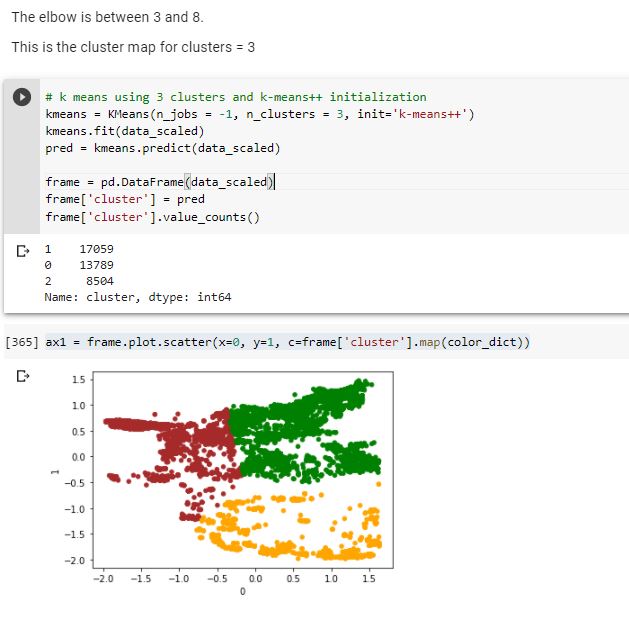
### **Elbow Curve**



**Color Array**

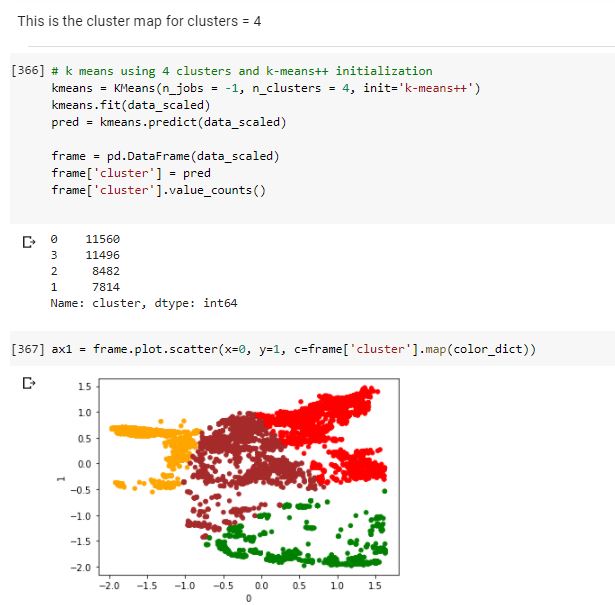
**Cluster Map, Clusters = 3**

Initial clustering with two groups on the top and one group (green) on the bottom



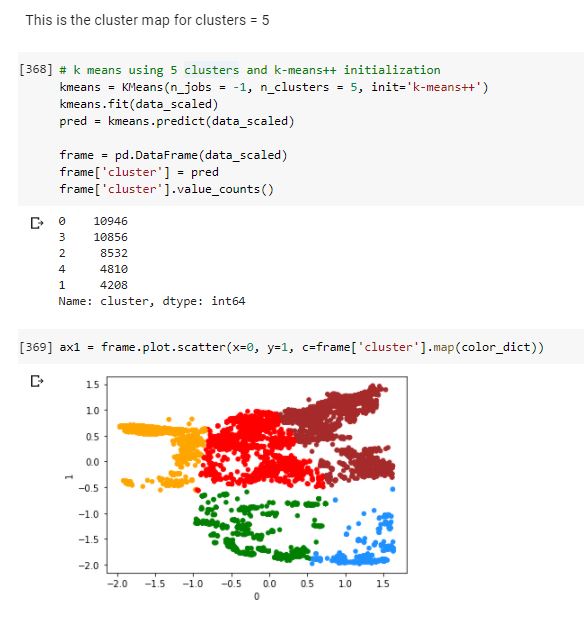
**Cluster Map, Clusters = 4**

Moving from 3 to 4 clusters, the two clusters on the top are now divided into 3 with a small part of the bottom cluster added to the center cluster.



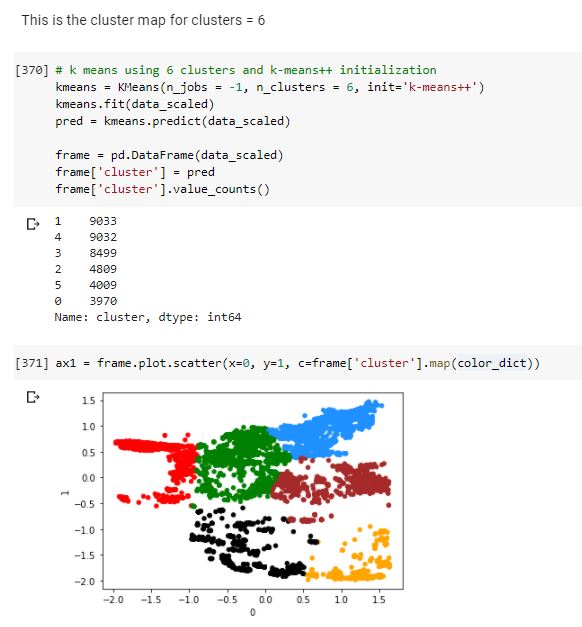
**Cluster Map, Clusters = 5**

Moving from 4 to 5 clusters, the bottom cluster is now divided in two, with the void in the middle of it as the main dividing area.



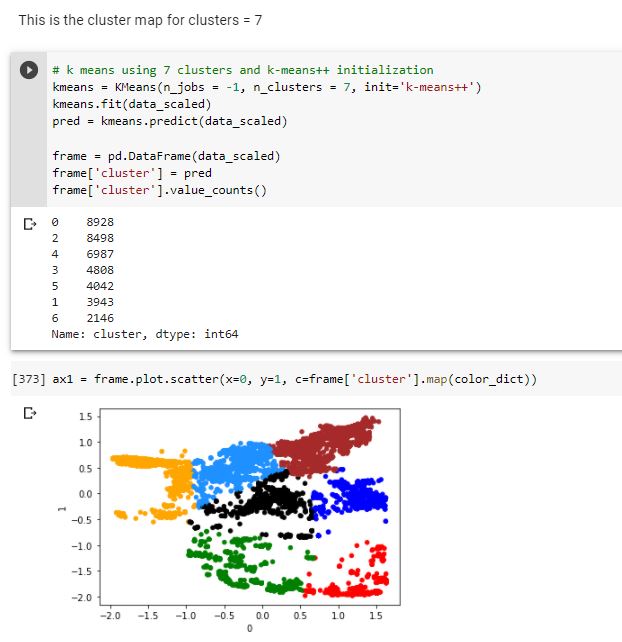
**Cluster Map, Clusters = 6**

Moving from 5 to 6 clusters, the top right cluster is now divided in two, with the two 'fingers' of the cluster seperating



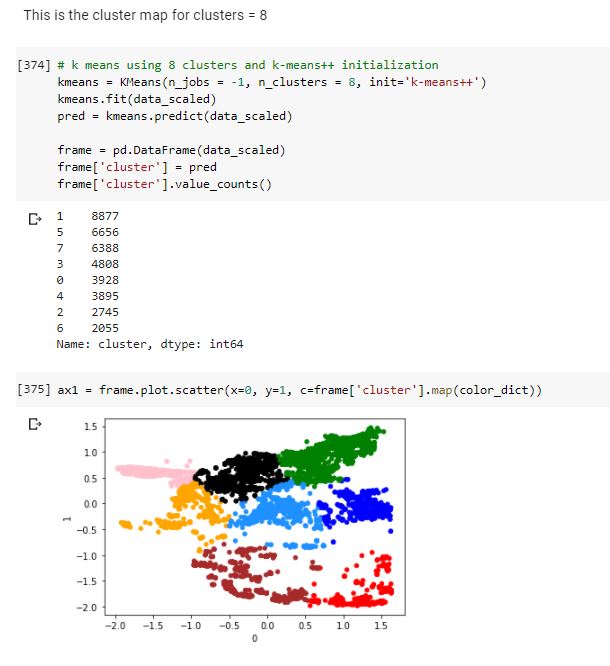
**Cluster Map, Clusters = 7**

Moving from 6 to 7 clusters, the upper center cluster is now divided in two, with the void in the middle of it as the main dividing area.



**Cluster Map, Clusters = 8**

Moving from 7 to 8 clusters, the top left cluster is now divided in two, with the two 'fingers' of the cluster separating.



[**Video Link**](https://www.youtube.com/watch?v=-korGxyKtPw)

**Any in site about the data or the ICP in general**

The ICP was quite enjoyable. Helped put context around the lecture from Tuesday.