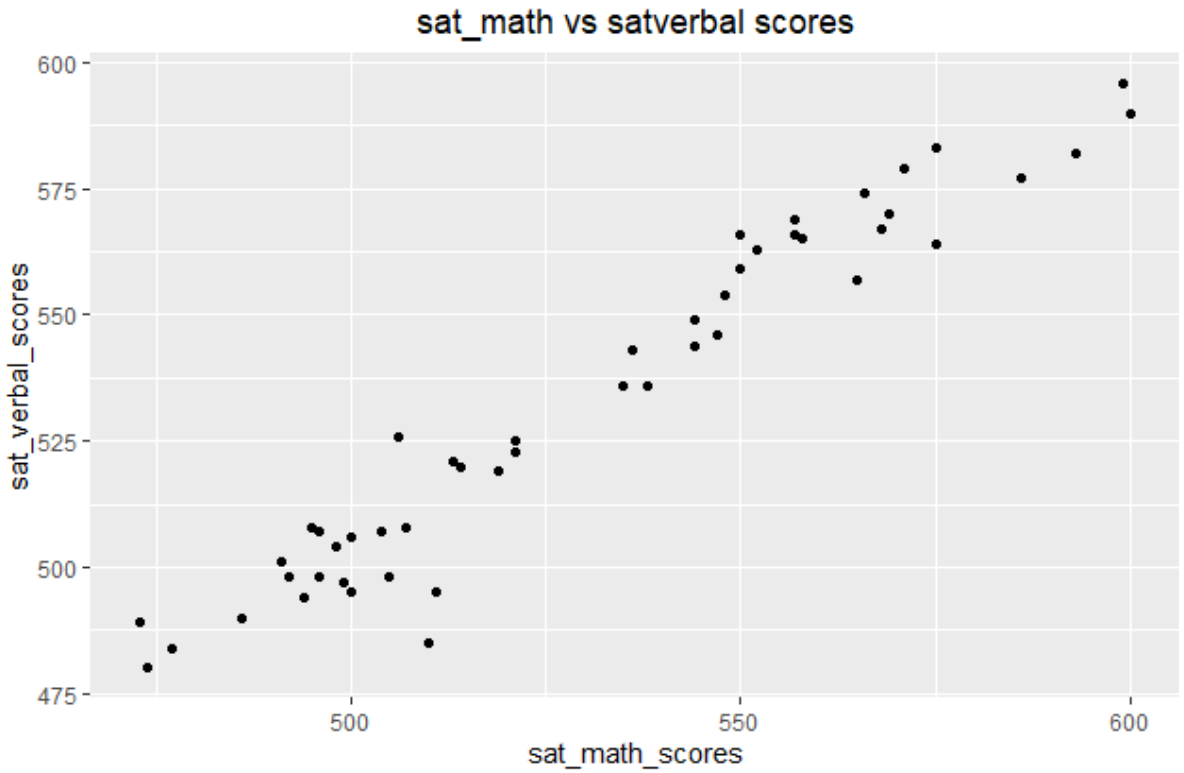
Assignment 12 (50 pts): Cluster Analysis, k-means

- **create a scatterplot of satmath vs satverbal scores; visually describe if there appear to be any data clusters (25 pts)**

1. Cleaned the dataset StateSAT.csv by checking the missing values and creating dummy numerical values for the categorical values in python and exported the data frame into clean.csv file.
2. Removed the states data column in the excel file as clustering can't happen with categorical variables ,even though mapping of states column is done, we get 51 unique numerical values and clustering the relationship between satmath and satverbal scores based on states column can give us 51 different data points (clusters) , which don't make any sense .
3. The data cleaning process(python code) is in the attached ML_Assignment_Cleaning.ipynb file
4. The Clustering Analysis and Visualization of clusters is done in R in the attached ML_Assigment.R file

Scatter plot of satmath vs satverbal scores:

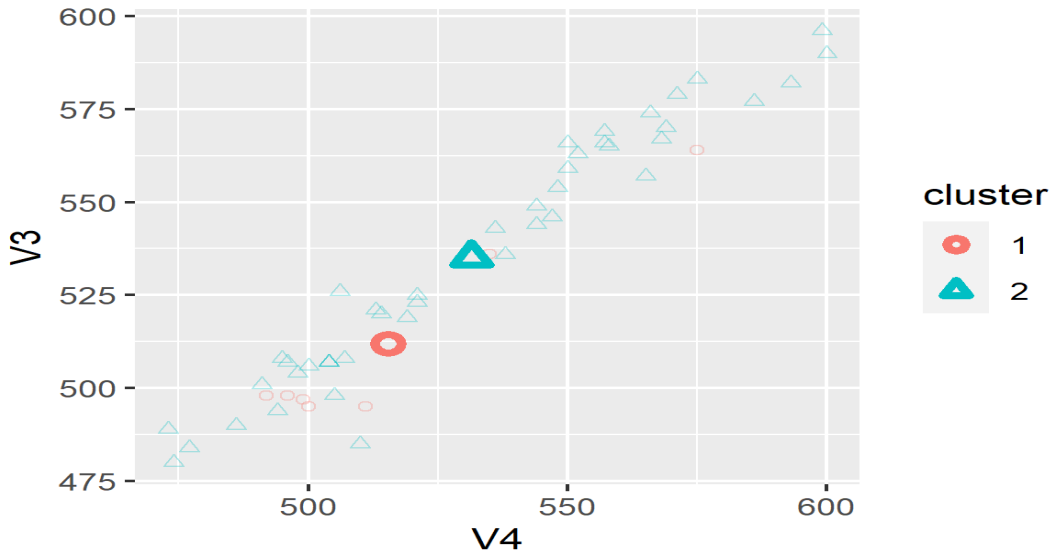


When a large data set is decomposed into smaller groups of related data elements, Clusters are formed. It is observed that there are 6-8 small clusters when sat_math and

sat_verbal scores are plotted. sat_math and sat_verbal scores are linearly related numerical variables. The linear relationship between them is almost strong and positive which can be inferred from the above visualization. This also says when the sat_math scores increase, the sat_verbal scores also increases.

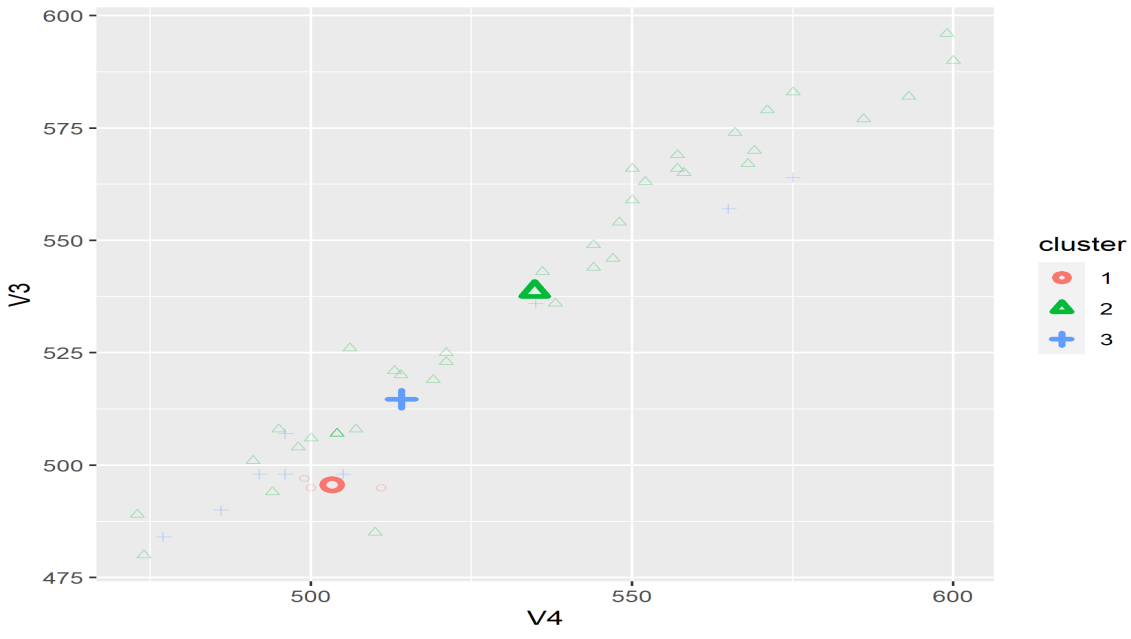
- **run the cluster code with different values of k (centers):**
  - **describe the clusters that appear to be found; how do they correspond to other data items in the dataset? How would you interpret the results? (25 pts)**

k = 2



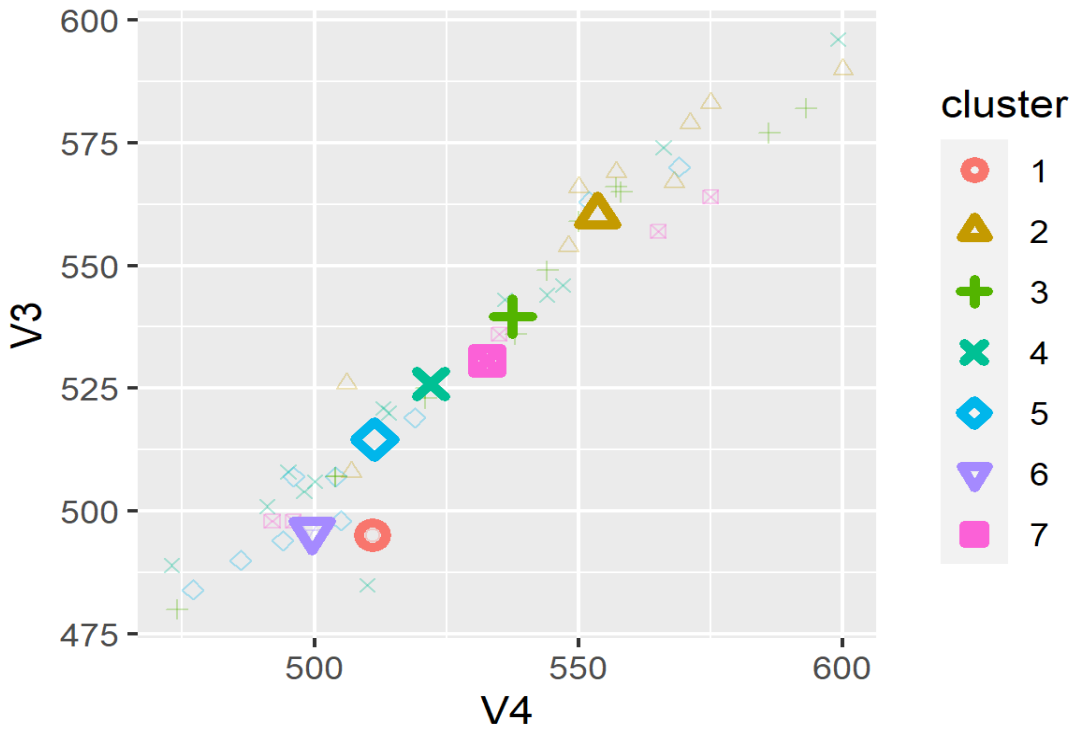
With Centres value 2, two clusters are formed in a scatterplot of sat_math and sat_verbal scores.

k=3



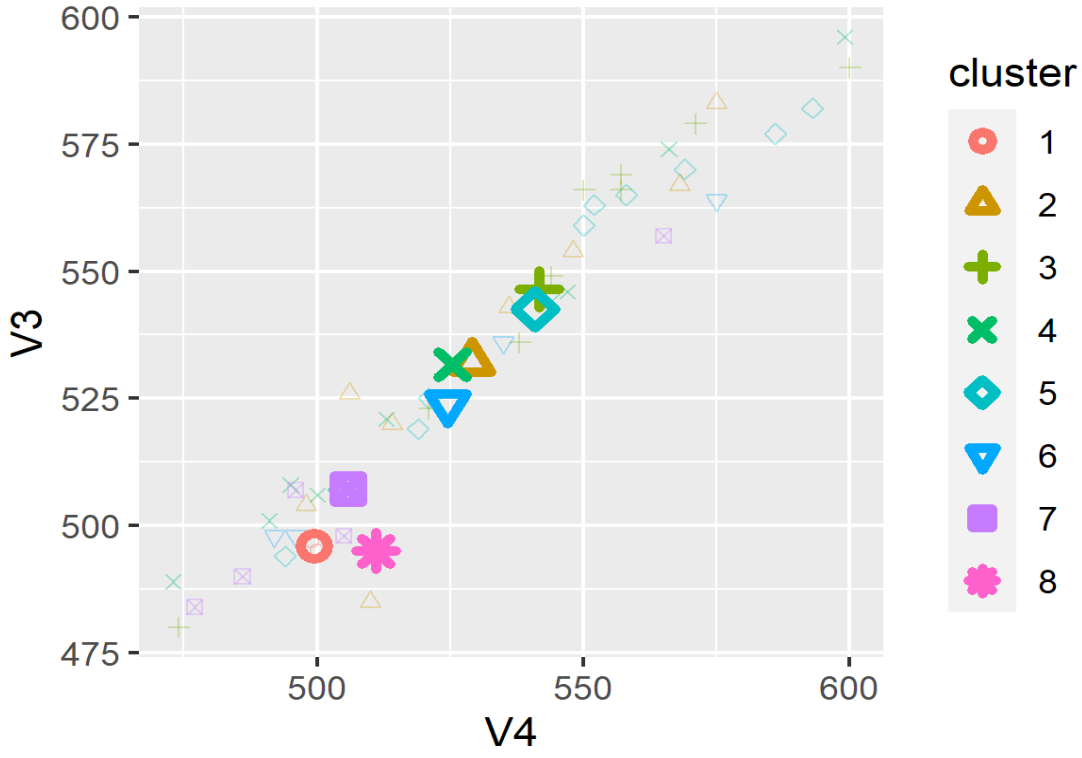
With Centres value 3, three clusters are formed in a scatterplot of sat_math and sat_verbal scores.

K=7



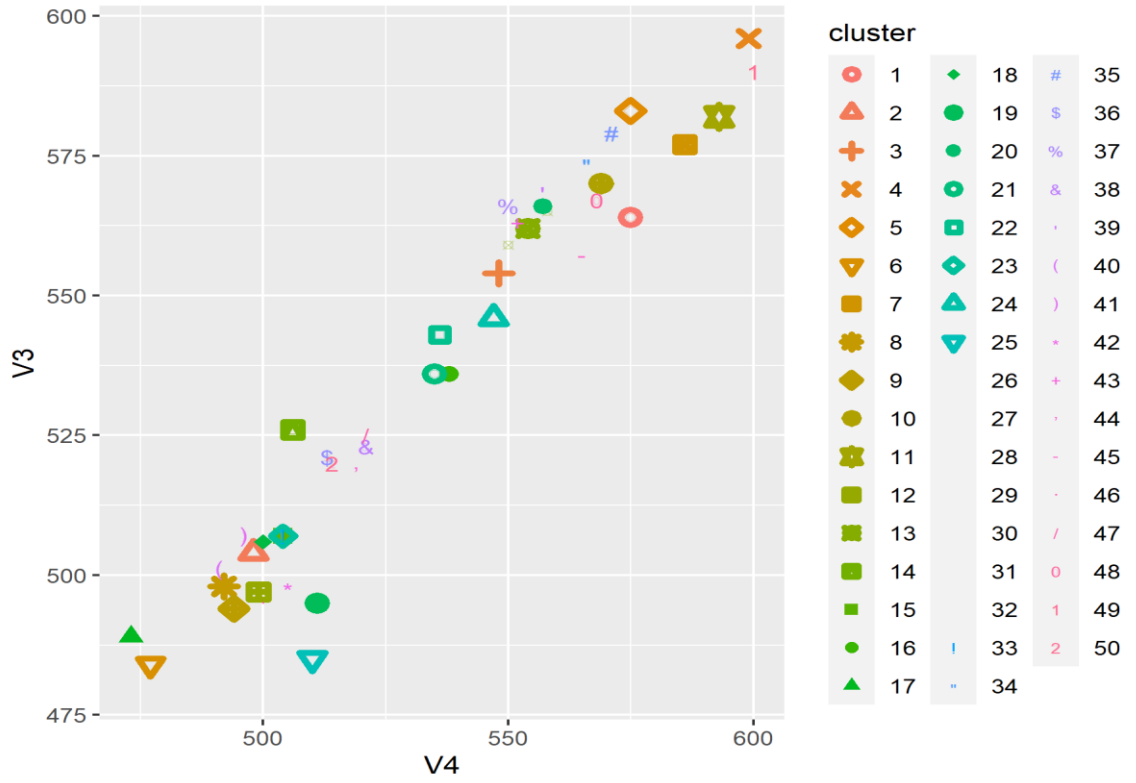
With Centres value 7, seven clusters are formed in a scatterplot of sat_math and sat_verbal scores.

K =8



With Centres value 8, eight clusters are formed in a scatterplot of sat_math and sat_verbal scores.

K =50



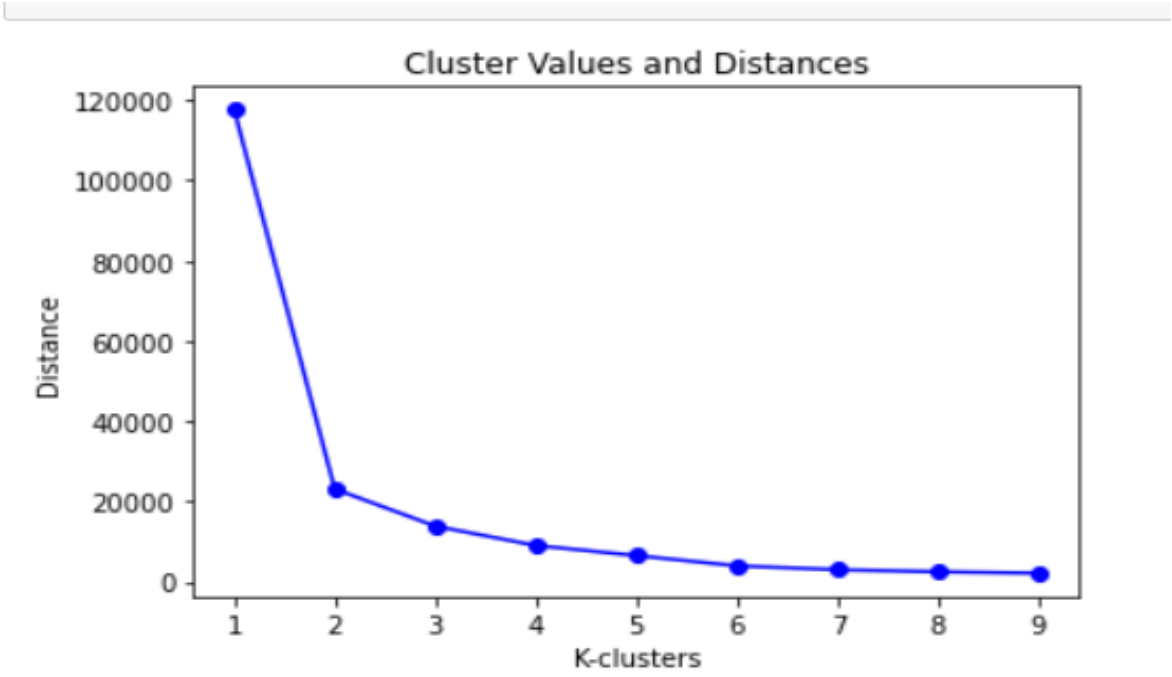
When Centres value is approximately equal to the number of observations or instances, then each data point represents a cluster and that doesn't make sense.

Finding K value:

Depending on the datasets, different k values will be appropriate for grouping. So, way to find a k value is to create clusters and then analyse the sum of squared distances.

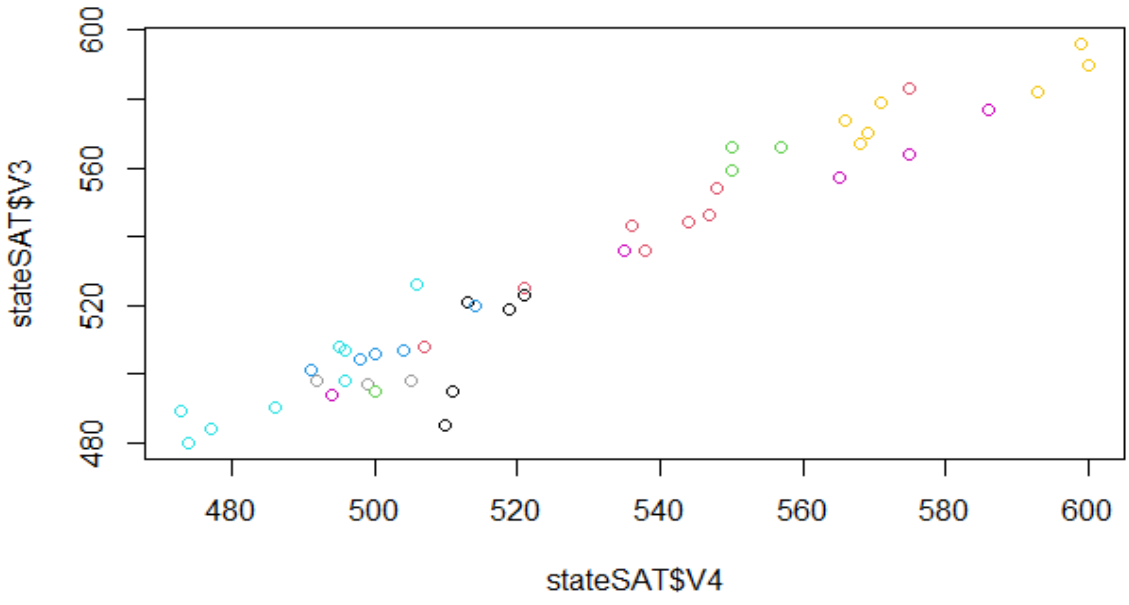
A common approach known as elbow method is also used to find the values of k

This can be seen through below visualization. Here at k = 2, we can get observe optimal bias and variance i.e., neither under fitting nor over fitting.

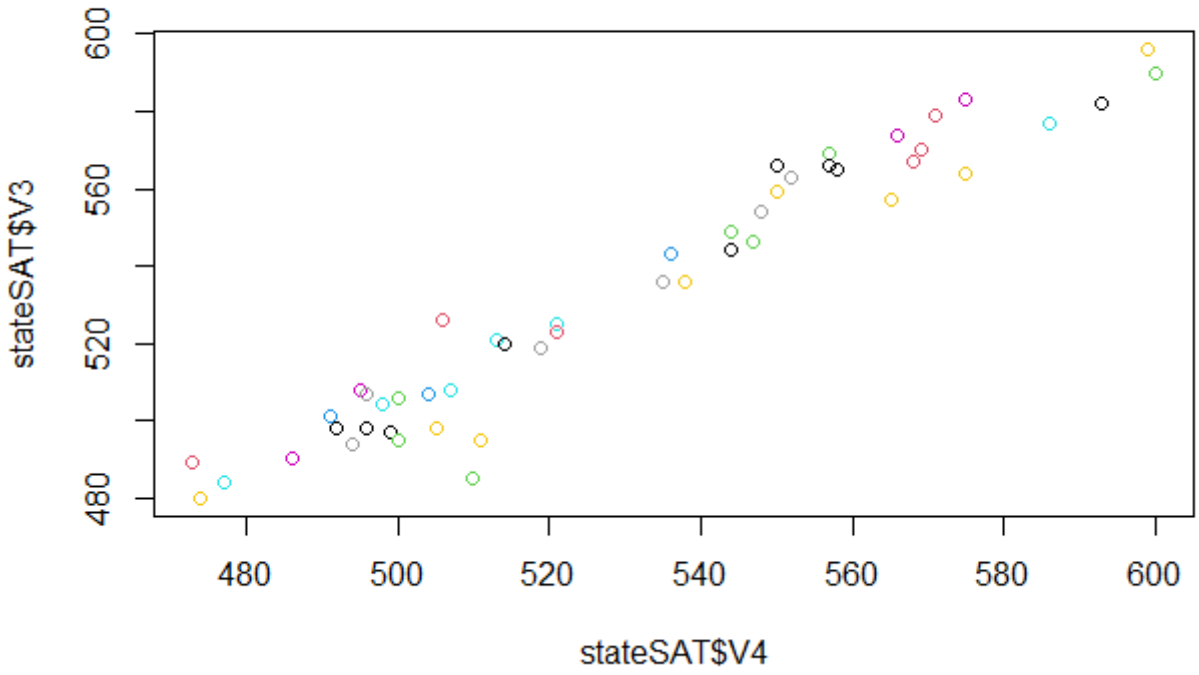


Influence of Region on cluster formation:

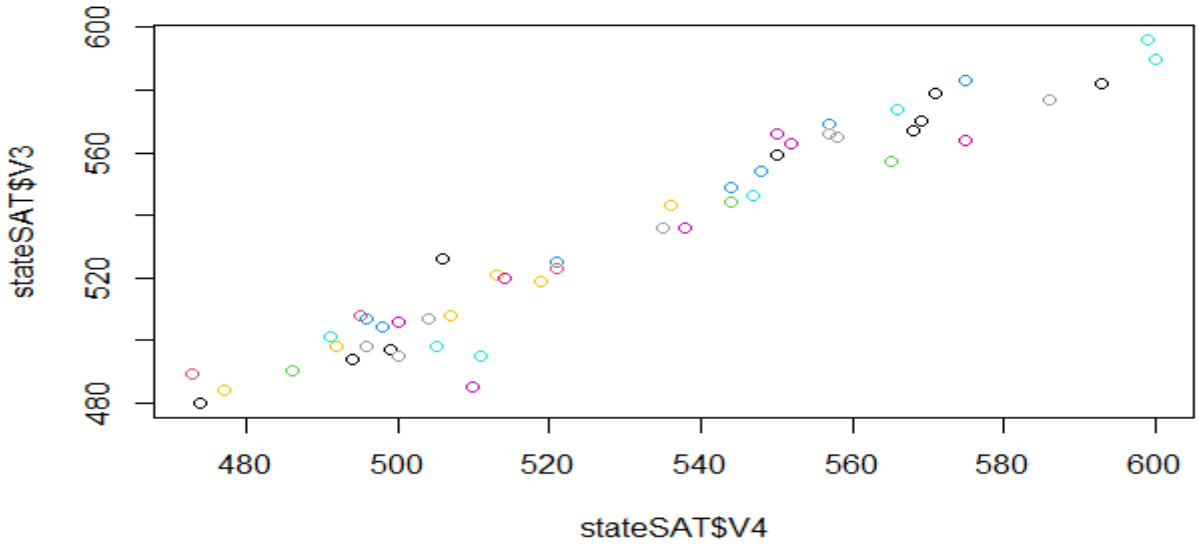
Cluster formation between sat_math and sat_verbal scores based on the V1 region gives us 8 clusters. This value of K i.e., 8 is appropriate for this dataset. But this cluster model may not work well with other unseen data when k is 8. That is, Other countries may have different number of regions. So k value 2 is optimal.



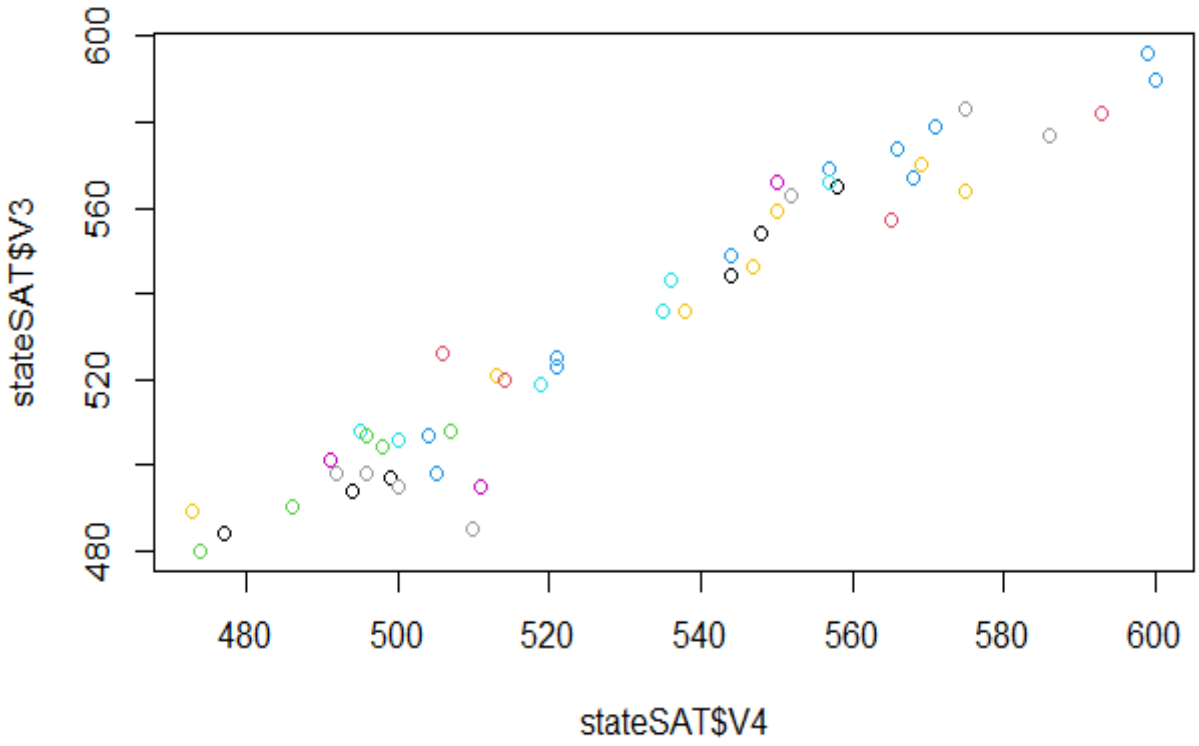
Influence of Percent taking on cluster formation: The influence of this data column is not as comparable with data column Region. With Region data column we can clearly see the differentiation of clusters compared to Percent taking



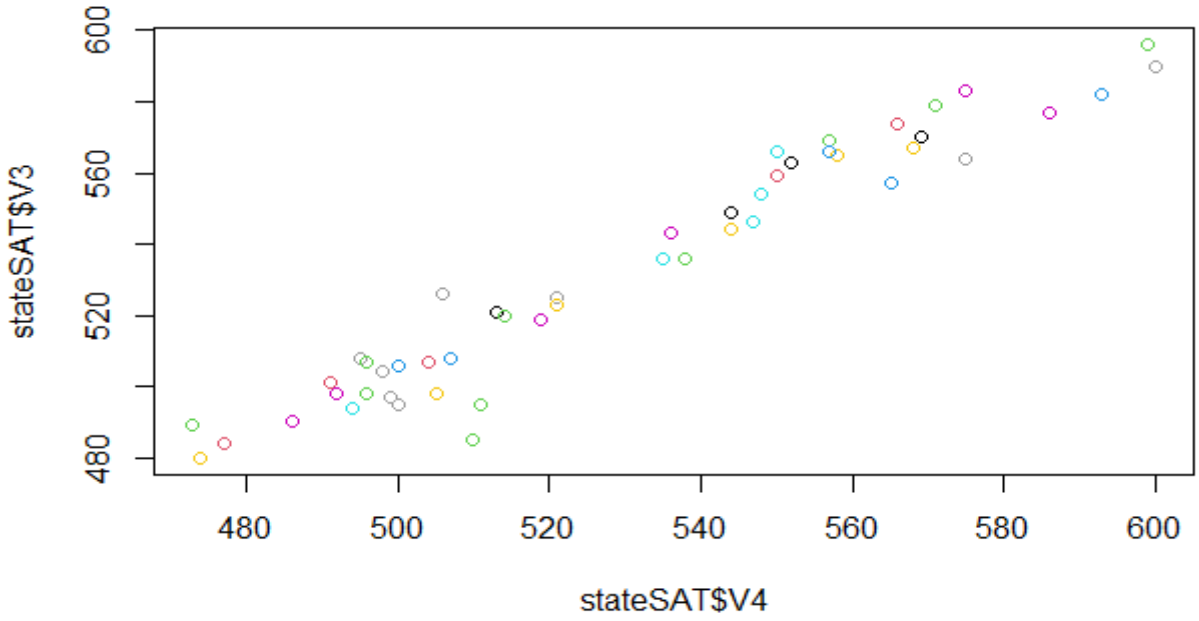
Influence of Percentnohs on cluster formation: The influence of this data column is not as comparable with data column Region. With Region data column we can clearly see the differentiation of clusters compared to Percentnohs



Influence of teacher pay on cluster formation: The influence of this data column is not as comparable with data column Region. With Region data column we can clearly see the differentiation of clusters compared to teacher pay.



Influence of population on cluster formation: The influence of this data column is not as comparable with data column Region. With Region data column we can clearly see the differentiation of clusters compared to population.



Finally, we can interpret that k =2 is optimum for this data set from the elbow method approach as explained in the pdf.