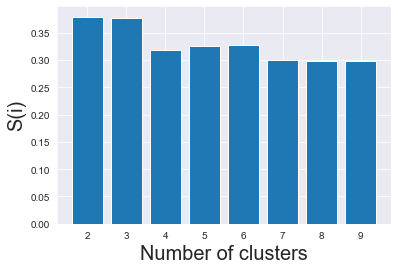
List of questions and recommandations

Cleaning, Treatment

* Basicly youd id data wrangling, and have you done something else? Could you explain quickly what data wrangling have solved in our case?

Q1

b)

* why did you use the k mean to solve it?
* Because the data wasn't a labeled data, a labeled data is data that has a column to target, in our problem, if the data was labeled, we'd have a column in the data that tells if a student is a potential donor or not, therefor we can target that column to build a supervised ML model to predict any other unlabeled data to see whether that unlabeled student is a donor or not, also in our problem, we already have all the data we may need, we already have the data about the student that we may target, there're no future student that we may need to label.
* You have used 2 or 3 clusters?
* I used 2 clusters as if you took a look on the following silhouette scores graph: 

You'll find that cluster 2 is nearly the same as cluster 3, so it won't make any difference if we choose any of them, but our problem requires determining whether a student is a potential donor or not (just two options required), so that I choose to use 2 clusters instead of 3.

Q2

a)

b)

* Can you interpret me the graph of the top 5 city?
* That graph shows the most cities with students that have the highest donor score that I created from the columns (Affinity Score, Capacity Score and Bequest Gift Score) after normalizing them and multiplying them to each other's, that produced a column that I can use to see much a student is really a potential donor, then I grouped the data based on cities to get the most cities that has the top donor scores on average, that lead me to figure out that (Prince William, Lac-Caymant, Reserve Mines, Albert Bridge and Lakeville-Westmorland) are the top 5 cities with highest donor scores that we should target respectively.
* We should only target city based on lifetime giving soly and not use multiple variables you think?
* Well, I rethought about that and yes you're right, Because of the potential outliers in the column that I was estimating by, it's better to use multiple columns instead.
* When you have removed the outliner basicly it has removed the highest gifts right? That’s why per exemple edmenton has 5 milions in excel and 4 mil in the graph right?
* When I removed the outliers, I removed it from a subset that I created especially for the graph, so that won't spoil the data, and after considering your recommendation in the previous question, I'll remove the removing outliers step.