CS 584 Machine Learning

Pradyot Mayank(A20405826)
Spring 2019 Midterm Test

Question 11 (25 points)

You can use the Chicago's 311 Service Request to report street potholes. After a request has been received, the Department of Transportation will first assess the severity of the pothole, and then schedule road crew to fill up the pothole. After the pothole is filled, the service request will be closed.

You are provided with this CSV file **ChicagoCompletedPotHole.csv** for analyzing the city's efforts to fill up street potholes. The data contains 17,912 observations. Each observation represents a completed request which was created between December 1, 2017 and March 31, 2018 and was completed between December 4, 2017 and September 12, 2018. The data has the following seven variables:

Name	Level	Description
1) CASE_SEQUENCE	Nominal	A unique index for identifying an observation
2) WARD	Nominal	Chicago's ward number from 1 to 50
3) CREATION_MONTH	Nominal	Calendar month when the request was created
4) N_POTHOLES_FILLED_ON_BLOCK	Interval	Number of potholes filled on the city block
5) N_DAYS_FOR_COMPLETION	Interval	Number of days elapsed until completion
6) LATITUDE	Interval	Latitude of the city block
7) LONGITUDE	Interval	Longitude of the city block

You will first identify clusters in the data, and then use a classification tree to profile the clusters. Here are the specifications for performing the analyses.

K-Means Clustering

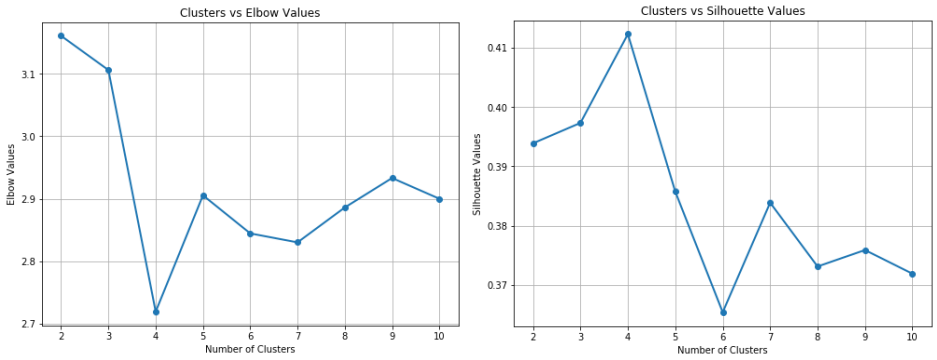
1. Use $\log_e$(N_POTHOLES_FILLED_ON_BLOCK), $\log_e$(1 + N_DAYS_FOR_COMPLETION), LATITUDE, and LONGITUDE (i.e., you need to perform the transformations before clustering)
2. The maximum number of clusters is 10 and the minimum number of clusters is 2
3. The random seed is 20190327
4. Use both the Elbow and the Silhouette methods to determine the number of clusters

Classification Tree

1. Use N_POTHOLES_FILLED_ON_BLOCK, N_DAYS_FOR_COMPLETION, LATITUDE, and LONGITUDE (without any transformations) as the predictors
2. The maximum number of branches is 2
3. The maximum depth is 2
4. The random seed is 20190327.
5. The grow criterion is the Gini's value

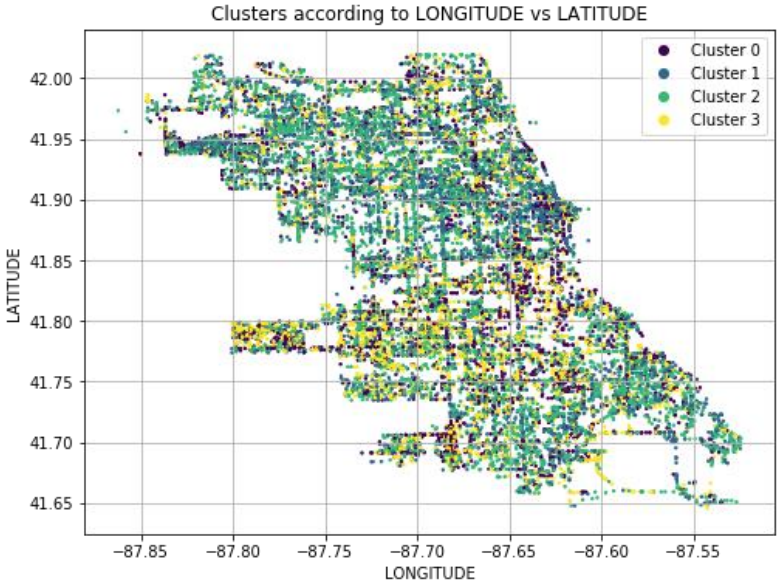
Please answer the following questions.

a) (5 points) How many clusters have you determined? Please provide the Elbow and the Silhouette charts and state your arguments. The charts must be properly labeled.
There are 4 clusters according to Elbow and Silhouette charts. Since there is a sharp decrease of elbow value when number of clusters are 4 and maximum value of Silhouette is attained when the number of clusters are 4.
So, we can say optimum number of clusters determined are 4.



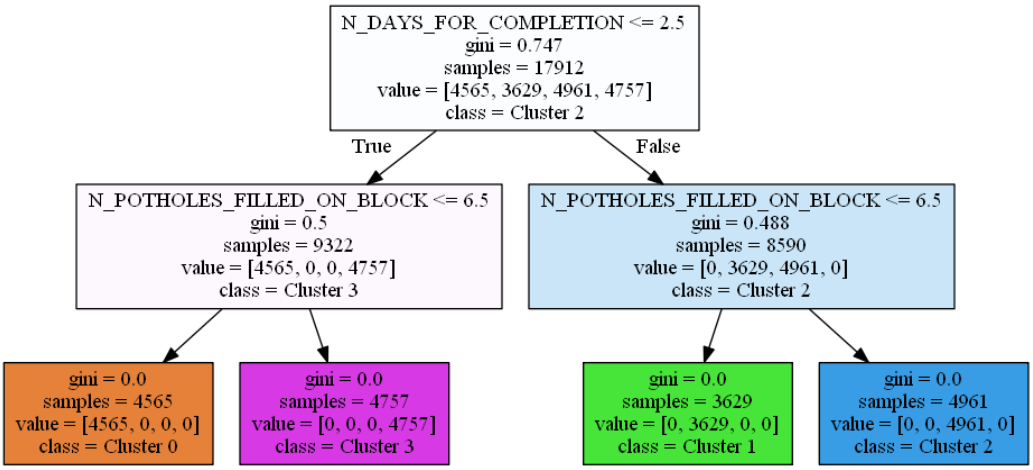
b) (5 points) Generate a scatterplot of LATITUDE (y-axis) versus LONGITUDE (x-axis) using the Cluster ID as the color response variable. You may need to adjust the marker size and set the aspect ratio to one in order to make the scatterplot more readable.

Scatterplot of LATITUDE (y-axis) versus LONGITUDE (x-axis) using the Cluster ID as the color response variable



c) (5 points) How many leaves in your classification tree? Please provide the properly labeled tree diagram.

There are 4 leaves in the classification tree.
Classification tree:



d) (5 points) What is the Root Average Squared Error of your classification tree? Please give your answer up to four decimal places.
Root Average Squared Error of the classification tree is zero as the gini index of each leave is 0. This means all the given samples are assigned to each one of the 4 clusters.

e) (5 points) Based on your classification tree, please describe the profiles of the clusters which are at least 99% correctly classified by the classification tree.
All the profiles of the clusters are correctly classified by the decision tree as each leaf contains a sample belonged to a particular cluster.

Cluster ID	Samples	N_DAYS_FOR_COMPLETION	N_POTHOLES_FILLED_ON_BLOCK
Cluster 0	4565	<=2.5	<=6.5
Cluster 1	3629	>2.5	<=6.5
Cluster 2	4961	>2.5	>6.5
Cluster 3	4757	<=2.5	>6.5