##Question 4

##a)

Example 1: Suppose a doctor want to find out whether or not a patient has diabetes (this is the response). The doctor collects several information from the patient such as blood glucose, weight, height, and etc. (these variables are the predictors). Then the doctor decides whether or not the patient has diabetes based on the information he collected. This is an example of classification. The goal is prediction.

Example 2: Suppose a medical institute wants to investigate which variables indicate diabetes (whether or not a patient has diabetes is the response). The institute collects data from a sample of random patients of both diabetes and non-diabetes. Several variables of each patient are collected such as blood glucose**,** weight, height, gender and etc. By statistical analysis, the institute decide which variables are necessary or significant in order to predict if a patient has diabetes (these variables are the predictors). This is an example of classification. The goal is inference.

Example 3: Suppose we are given Capital Bikeshare data from the second quarter of 2016. Each row of the data set is a bike that is out for service. Columns in the data set are variables such as duration of the trip, bike identification number, date, and etc. (these variables are the predictors). We would like to predict whether or not a bike is in repair shop instead of used by customers based on the given variables. So the response is whether or not a bike is in repair shop. This is an example of classification. The goal is prediction.

##b)

Example 1: Suppose we want to predict a student’s GRE score based on his or her GPA and hours spent for preparing GRE test that we know. We also know there is a function that takes his or her GPA and hours spent for studying GRE as input and output GRE score. We plug in this function and calculate the student score. This is an example of regression. The response is a student’s GRE score. The predictors are his or her GPA and hours spent for preparing GRE test. The goal is prediction.

Example 2: Suppose we want to generate a function to predict a student’s GRE score based on two variables-his or her GPA and hours spent for preparing GRE test. We collect a sample of random students who have taken GRE test along with their GPA and hours spent for preparing GRE test. Then, we fit a model with multiple linear regression to see if the variables are significant. This is an example of regression. The response is a student’s GRE score. The predictors are his or her GPA and hours spent for preparing GRE test. The goal is inference.

Example 3: Suppose we want to find out whether or not hours of studying per week can predict a student’s GPA at Georgetown University. We collected a sample of random students along with their hours of studying per week and GPA at Georgetown University. Then, we fit a model with linear regression to see if hours of studying per week is a significant predictor of GPA. This is an example of regression. The response is a student’s GPA at Georgetown University. The predictors are his or her hours spent for studying per week. The goal is inference.

##c)

Example 1: Suppose we want to cluster tweets on twitter based on their sentiments of six so-called “basic” emotions (i.e. anger, disgust, happiness, sadness, surprise, and fear). Each tweet is assigned a score for each of the emotion which is the percentage of words appear in the emotional category. Then, we conduct a K-means cluster analysis to determine if the tweets fell into emotion patterns that could be used for grouping purposes.

Example 2: Suppose we want to group students at Georgetown University based on their hobbies, major, years of study. In this case, cluster analyses such as K-nearest neighbors might be useful.

Example 3: Suppose we want to group species based on species’ characteristics such as bipedalism or reptile, amphibian or not, and etc. In this case, cluster analyses such as K-nearest neighbors might be useful.