University of San Diego®
SHILEY-MARCOS
SCHOOL OF ENGINEERING

Module 6 Assignment Questions

Note that the answers to each of these questions should be the direct result of running appropriate Python or R code and not involve any manual processing of dataset files. Answers without either the code or results will not receive any grade.

1. For the next exercise, you are going to use the “airline_costs.csv” dataset.
   The dataset has the following attributes:
   - i. Airline name
   - ii. Length of flight in miles
   - iii. Speed of plane in miles per hour
   - iv. Daily flight time per plane in hours
   - v. Customers served in 1000s
   - vi. Total operating cost in cents per revenue ton-mile
   - vii. Revenue in tons per aircraft mile
   - viii. Ton-mile load factor
   - ix. Available capacity
   - x. Total assets in $100,000s
   - xi. Investments and special funds in $100,000s
   - xii. Adjusted assets in $100,000s

   (Implement this exercise in Python language; import ‘pandas’, ‘statsmodels.api’ libraries) Use a linear regression model to predict the number of customers each airline serves from its length of flight and daily flight time per plane. Next, build another regression model to predict the total assets of an airline from the customers served by the airline. Do you have any insight about the data from the last two regression models? **(20 points)**

2. For this clustering exercise, you are going to use the data on women professional golfers’ performance on the LPGA, 2008 tour (“lpga2008.csv” dataset). The dataset has the following attributes:
   - i. Golfer: name of the player
   - ii. Average Drive distance
   - iii. Fairway Percentage
   - iv. Greens in regulation: in percentage
   - v. Average putts per round
   - vi. Sand attempts per round
   - vii. Sand saves: in percentage
   - viii. Total Winnings per round
   - ix. Log: Calculated as (Total Win/Round)
   - x. Total Rounds
   - xi. Id: Unique ID representing each player

(Implement this exercise in R language; import ‘cluster’ library)
Use agglomerative clustering and divisive clustering on this dataset to find out which players have similar performance in the same season. Visualize the clusters using dendrograms for both types of clustering models. **(20 points)**