Module 5 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', 'sklearn.linear_model import LinearRegression' libraries)

1.1) 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 (this will be referred to as model 1 throughout this question).

(Note: save each of these objects as different variable names as it will make question 1.7 easier for you) **(10 points)**

1.2) What is the Root Mean Squared Error (RMSE) of this model? (Hint: import 'from sklearn.metrics import mean_squared_error' and numpy's sqrt function to help solve for this) **(10 points)**

1.3) Now repeat exercises 1.1 and 1.2, but first split the data into train (80%) and test (20%) datasets and find the RMSE of the test set (this will be referred to as model 2 throughout this question) **(10 points)**

(Hint: import 'from sklearn.model_selection import train_test_split' to help solve this)

1.4) Now find the RMSE of the train set. **(5 points)**

1.5) What do you notice about the difference between the RMSE on the entire dataset (model 1), the RMSE on the 20% test/holdout set (model 2), and the RMSE on the 80% train set (model 2)? Why do you think this is? **(10 points)**

1.6) Build another regression model to predict the total assets of an airline from the customers served by the airline using a 75%/25% train-test dataset split. Evaluate the RMSE of this model as well (this will be referred to as model 3 throughout this question).
(Note: your predictor variables must a DataFrame, not a Series to use sklearn's linear model) **(10 points)**

1.7) What are the coefficients of the 3 models? (look up in the sklearn documentation on how to find this) **(10 points)**

1.8) What do you notice about these coefficients? Research what linear regression coefficients mean if you are not sure. **(5 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:
Golfer: name of the player
   - Average Drive distance
   - Fairway Percentage
   - Greens in regulation: in percentage
   - Average putts per round
   - Sand attempts per round
   - Sand saves: in percentage
   - Total Winnings per round
   - Log: Calculated as (Total Win/Round)
   - Total Rounds
   - Id: Unique ID representing each player **(10 points)**

2.1) Use agglomerative clustering on this dataset to find out which players have similar performance in the same season. To do this, perform the following:

   - First, remove the columns 'Id' and 'Golfer' from the dataset
   - Normalize the data using 'from sklearn.preprocessing import StandardScalar' and the method 'fit_transform()'
   - Save this result into a dataframe
   - Next, use 'import scipy.cluster.hierarchy as shc' and 'import matplotlib.pyplot as plt' to visualize the a dendrogram of this data
   - Use the 'sch.linkage()' method with the linkage as ward and the metric as Euclidean to create the clusters
   - Then use the 'sch.denogram()' method and 'plt.show()' to visualize the denogram
   - Once we've plotted this denogram, we see that a good number of clusters is 4.
   - Use 'from sklearn.cluster import AgglomerativeClustering' and implement a model that has 4 clusters, linkage as ward, and the metric as Euclidean
   - Print the cluster labels for this model on our normalized dataset **(9 points)**

2.2) What is the difference between agglomerative clustering and divisive clustering? **(1 point)**