MIS2502: Data Analytics
Assignment: Clustering Using R

For this assignment, you'll be working with the Jeans.csv file and the Clustering.r script. This file has data from 689 stores that sell four different types of jeans: leisure, fashion, stretch, and original. The marketing division of the company wants to identify groups of stores that sell a similar mix of their product so that they can roll out promotions specific to those stores.

The data file contains the following fields:

Variable Name	Variable Description
StoreID	Store identification number
Fashion	The number of pairs of "fashion" style jeans sold last month
Leisure	The number of pairs of "leisure" style jeans sold last month
Stretch	The number of pairs of "stretch" style jeans sold last month
Original	The number of pairs of "original" style jeans sold last month
TotalSold	The total number of jeans sold last month

You'll need to modify the script with the following information to perform the analysis:

- Set the input filename to the store's dataset.
- Set the number of clusters to create (NUM_CLUSTER) to 5.
- Set the variable list (VAR_LIST) to use the Fashion, Leisure, Stretch, and Original variables.

Answer the following questions (complete the worksheet at the end of this document):

1) Which cluster is the largest (write the number of the cluster)?
2) How many stores are in the largest cluster?
3) Describe the sales of cluster 1 for each type of jeans (compared to the overall average across all stores)? (write one or two sentences)
4) Describe the sales of cluster 5 for each type of jeans (compared to the overall average across all stores)? (write one or two sentences)
5) In which of the five clusters of stores do original jeans sell the best?

6) What is the range of withinss errors for the five clusters?

_____(lowest) to _____(highest)

7) What is the average betweenss error for all five clusters?

Now rerun the script, this time with 20 clusters. Then answer the following questions:

8) Describe the sales of cluster 15 for each type of jeans (compared to the overall average across all stores)? (write one or two sentences)

9) Describe the sales of cluster 20 for each type of jeans (compared to the overall average across all stores)? (write one or two sentences)

10) What is the range of withinss errors for the 20 clusters?

_____(lowest) to _____(highest)

11) What is the average betweenss error for all 20 clusters?

12) Which scenario (5 clusters or 20 clusters) produces clusters with better cohesion?

13) Which scenario (5 clusters or 20 clusters) produces clusters with better separation?

14) Besides cohesion and separation, what other advantage does the 5 cluster scenario have over the 20 cluster scenario? (write one or two sentences)

What to submit:

Send a single email to your instructor with the following attachments:

- The completed, working R script that produced the analysis for the 20 cluster scenario.
- The output file "ClusteringOutput.txt" and "ClusteringPlots.pdf" for the 20 cluster scenario.
- The completed worksheet provided on the last page.

Answer Sheet for Assignment: Clustering Using R

Name _____

Fill in the worksheet below with the answers to the questions on pages 1 and 2 of the assignment:

Question	Answer
1	
2	
3	
4	
5	
6	Lowest: _____ Highest: _____
7	
8	
9	
10	Lowest: _____ Highest: _____
11	
12	
13	
14	