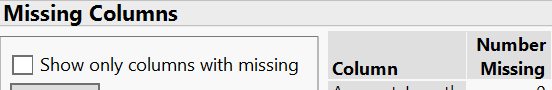
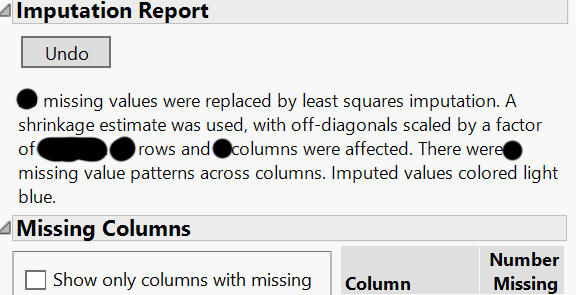
BDA 620

1. Use hierarchical clustering on the “Colleges and Universities” data set. Use “School” as a label. Use Average, Centroid, and Ward methods and run each with the data standardized.
   1. How many clusters does each algorithm produce? Copy the dendrogram for each method with color and mark clusters.
   2. Copy the cluster means tables for each method and explain how each cluster is grouped in each method?
   3. Use k-means clustering with k=2,3,4,5,6,7, and 8. What is the optimal number of clusters among k=2,3,4,5,6,7, and 8? Consider the optimal *k*, create parallel coordinate plots, copy it on your answer file and explain why it is optimal among the other alternatives.
   4. Based on the clustering results with optimal *k*, do you have any outliers in your data set? If yes, which cluster/s could be outlier/s and why?
2. Use the “Telephone\_Calls” data set.
   1. Find the number of outliers with “Tail Quantile” = 0.1 and Q=3 (Note that Q is a multiplier used to determine outliers for the chosen variable). How many outliers are there in the data set? Copy the table that shows the number of outliers for each variable.
   2. Find the number of outliers with “Tail Quantile” = 0.3 and Q=3. How many outliers are there in the data set? Copy the table that shows the number of outliers for each variable.
   3. After detecting the outliers in part b, remove the outliers from the data set. How many observations do you have after removing the outliers?
   4. After removing the outliers in part c, detect the number of missing values. Copy the window that demonstrates the number of missing values in each variable. Part of the table is shown below.



* 1. Impute the missing values using Multivariate Normal Imputation. Copy the “Imputation Report” that occurred at the same window with “Missing Columns” as shown in the sample picture.



* 1. Apply hierarchical clustering to the clean data set. Copy the dendrogram and report the number of recommended clusters in the final data set.
  2. Use k-means clustering to the clean data set with k=15 through 25. What is the optimal number of clusters using the optimal CCC value?