Submission Instructions

- Every student must upload his or her own homework and all source files (R scripts, Excel spreadsheets, Tableau workbooks, etc.) in one **zip** folder titled MP3_FirstName_LastName.zip.
- Please do not submit .rar files.
- If you are partnering up with another student, then only 1 submission is required. Do not forget to write the names of the students you worked with (2 other students in the course maximum).
- **The report should be a Word file, double-spaced in Times New Roman 12pt with 1-inch margins, with a cover page, an executive summary, table of contents, list of figures, list of tables, main body of report: introduction (what is the problem about), answers to questions below, conclusions, references, appendix, all pages numbered.**

Mini-Proj3

A company in Chile markets sensors to retail stores so that the stores can better keep track of the traffic in the stores, clients who made purchases, number of salespeople in the stores, etc.

The data MiniProj3.csv has daily data for 30 stores in Chile over a year. First column is code of store, second column is name of store which is also city where the store is, then date of sensor data as given by year, month, day, name of day (in Spanish), week of the year (a number between 1 and 52), traffic data (0 if sensor was malfunctioning or not installed yet), number of customer, sales, average ticket (=average amount on receipt), number of salespeople who worked that day, sales per salesperson, number of tickets (receipts) for salesperson, average number of products per ticket (receipt), average price in Chilean pesos, average number of salespeople working at the same time. You do not have to use all the data for your project. 0 in the data means missing data (sensor didn't detect anything, or was not installed yet) or the store was closed.

You can use R or Excel as software for the plots and summary tables (PivotTable in Excel). Hierarchical clustering and k-means clustering must be done in R, obviously.

Deliverables

1- **Question 1:** Cluster the stores in 3-10 groups based on traffic volume and pattern during the year.

   a. **(*Data exploration – 2pts*)** Plot the traffic over time for at least 5 stores. Try to pick 5 stores that don't all show the same traffic pattern over time.
   b. **(*Data analysis – 2pts*)** Create a summary table with the average traffic per day of the week for each store. Each store should occupy one row. (Note: some stores are closed on Sundays, others are open only half a day.)
   c. **(*Hierarchical clustering – 2pts*)** Do hierarchical clustering with the data in (b) only. Justify an appropriate number of clusters and provide cluster composition (which stores in which clusters).
   d. **(*K-means clustering – 2pts*)** Do k-means clustering with the same number of clusters as in (c). Provide cluster composition. Has it changed compared to (c)? If so, why?
   e. **(*Refined data analysis – 2pts*)**. Augment your table from (b) using more dimensions of your choice. Possible ways to proceed would be to look up those cities' population or geographic location (near the sea, near a neighboring country, mining town, small town…), average traffic over the year, existence of peaks in traffic at key dates during the year (Christmas, Mother's Day, etc.), ratio of maximum to minimum traffic over the year, or a pattern of your choice. You may also want to remove some stores from further analysis if they had no data or too little data.
   f. **(*Hierarchical clustering revisited – 2pts*)** Repeat (c) with your new summary table from (e). What is the number of clusters you recommend now? The composition of clusters? Can you describe the clusters in words to a business manager?
   g. **(*K-means clustering revised – 2pts*)** Repeat (d) with your new summary table from (e). What is the number of clusters you recommend now? The composition of clusters? Can you describe the clusters in words to a business manager?
   h. **(*Conclusions – 2pts*)** Provide your recommended number of clusters and their composition. Justify your answer. For each cluster, plot for each day of the year (January 1 to December 31) the average traffic as well as the minimum and maximum traffic. (The idea is that we would use the cluster's average traffic to predict future traffic. Since we only have barely one year of data at each store, we can't predict future traffic at each store using only the past historical data at this store.)

2- **Question 2:** Explore the relationship, if any, between salespeople (or salespeople working concurrently) and customers (or sales). (I am not after a specific in-depth answer – this could keep us busy for an entire semester – but am more interested in seeing how you approach the problem.)
   a. (***Data analysis – 2pts***) In each of the clusters you found in Question 1 (h), identify a store that seems to make the best use of its salespeople (delivers best results, measured by number of customers or sales, for the smallest number of salespeople). Justify your answer. For each of those stores, plot how sales vary with the clusters
   b. (***More clustering – 2pts***) Cluster stores based on average number of salespeople only. You can use hierarchical clustering or k-means clustering. Justify the number of clusters and provide cluster composition. For each cluster, which store has the best sales? (as measured by average sales, or average sales outside Christmas, where demand is so high that it doesn't seem the stores need salespeople to sell their goods, or a metric of your choice.) Which store has the worst sales?
   c. (***More clustering – 2pts***) Repeat (b) using average number of salespeople working concurrently. Does your clustering change?
   d. (***Conclusions – 1pt***) Based on above, which store would you say is underperforming (maybe is overstaffed, or should retrain its salespeople to make them more productive)?

3- All software files (R scripts, Excel spreadsheets, Tableau workbooks, etc.)

Grading Scheme

- 16 points for Question 1
- 7 points for Question 2
- 1 point for the overall quality of the written report and strength of final recommendation.
  - Is the report clear and well-written?
  - Are the guidelines above followed?
  - Is the problem explained?
  - Are the results well documented?
  - Is this a report that I can show to a company?