Store Lab – Part I

This lab is about understanding the mechanisms driving sales for Basic Sun Preparations, a skincare product.

**Data description**

Monthly sales data for a skincare product.

The dataset contains sales data for multiple stores across the nation, and includes information about the region where the store is located, store characteristics, product shelf space use, and store sales performance.

**Instructions:**

• Read the lab all the way through

• Complete the lab

• Clearly label all .R scripts and submit relevant scripts, word documents and slides

Section 0: Store Lab dataset

Open **Store data.xlsx** and inspect the dataset. Review the data dictionary, located in the **dictionary** sheet of the workbook and make sure you understand the meaning of each column. Please bring any questions to class.

*\* Copy All the files in a folder and set that folder as your working Directory in RStudio.*

Section 1: Explore the dataset and practice R, basic operations

**Use the "Store data.xlsx" file for these exercises.**

1. Read the dictionary to understand the variables in the dataset.
2. Import the data into RStudio as a data frame, use "File" 🡪 "Import Dataset" 🡪 "From Excel..."
3. Explore the data to learn the following information:
   1. # of store formats included in the data
   2. # of regions
   3. # of years of sales
4. Calculate the total tax-included sales (SALES\_TISP) of Store #19 for all time
5. Create a new copy of your data frame, named "Store\_Copy" and use this data frame for the rest of the problems
6. Use View(Store\_Copy) to visually inspect the data
7. **Challenge:** Find the number of stores which don't have 12 months of sales information
   1. Hint: Check that there are no duplicates of STORE\_ID and MONTH, one approach:

sum(!duplicated(Store\_data[,c("STORE\_ID", "MONTH")]))

* 1. Hint: Try using table() with two column arguments to make a two-way table of STORE\_ID and MONTH. Use STORE\_ID as the first argument
  2. Hint: Try converting your table to a data frame using as.data.frame.matrix()
  3. Hint: Try summing the columns or rows to see which stores have fewer than 12 months of data using rowSums(), you can append a "total" column to your existing data frame
  4. Hint: Try subsetting the data frame to find the stores where the total column is less than 12

**PAUSE HERE, CHECK ANSWERS WITH CLASS, BRING FORWARD ANY QUESTIONS**

Section 2: Manipulate the dataset

**Use the Store\_Copy data frame for these exercises.**

1. Add new columns with the following definitions

Note: log10() computes the base 10 logarithm in R

* + - * **Avg\_Price**: SALES\_TISP/SALES\_UNITS
      * **Space\_Yield**: SALES\_TISP/SPACE
      * **NDSA\_Yield**: SALES\_TISP/NDSA
      * **LOG\_TISP**: Log10(SALES\_TISP)
      * **LOG\_TESP**: Log10(SALES\_TESP)
      * **LOG\_Space**: Log10(SPACE)
      * **LOG\_NDSA**: Log10(NDSA)
      * **LOG\_Price**: Log10(Avg\_Price)
      * **LOG\_Space\_Yield**: Log10(Space\_Yield)
      * **LOG\_NDSA\_Yield**: Log10(NDSA\_Yield)
      * **Date**: create a data column with the format: YEAR-MONTH-01 (a date column using YEAR as the year, MONTH as the month and 01 as the day of the month)

Hint: Store\_Copy$Date <- as.Date(with(Store\_Copy, paste(YEAR, MONTH, 1,sep="-")), "%Y-%m-%d")

**Use the Store\_Copy data frame for these exercises.**

1. Create a new column called **RegionsText** – and transform the Region data as follows
   * 1 🡪 "South"
   * 2 🡪 "Pacific"
   * 3 🡪 "Mid-Atlantic"
2. For SALES\_TISP and NDSA, calculate the following descriptive statistics by **Format** – and fill in the table below (you will need a separate table for NDSA):
   * Total
   * Average
   * Max
   * Min
   * Cutoff value for highest 1% (hint, use quantile())
   * Cutoff value for lowest 1% (hint, use quantile())
   * How many values are in the highest 1%?
   * How many values are in the lowest 1%?

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Convenience** | **Chemist** | **Destination** |
| Total |  |  |  |
| Average |  |  |  |
| Max |  |  |  |
| Min |  |  |  |
| Cutoff value for highest 1% |  |  |  |
| Cutoff value for lowest 1% |  |  |  |
| Count how many values are in the highest 1% |  |  |  |
| Count how many values are in the lowest 1% |  |  |  |

**PAUSE HERE, CHECK ANSWERS WITH CLASS, BRING FORWARD ANY QUESTIONS**

1. Remove the highest 1% and lowest 1% of the data identified from above for SALES\_TISP and store in a new data frame called Store\_Data\_No\_Outliers
2. Using Store\_Data\_No\_Outliers calculate the total sales for each format by month, and report your results in the following table

|  |  |  |  |
| --- | --- | --- | --- |
| **Month** | **Convenience** | **Chemist** | **Destination** |
| Jan |  |  |  |
| Feb |  |  |  |
| Mar |  |  |  |
| Apr |  |  |  |
| May |  |  |  |
| Jun |  |  |  |
| Jul |  |  |  |
| Aug |  |  |  |
| Sep |  |  |  |
| Oct |  |  |  |
| Nov |  |  |  |
| Dec |  |  |  |

**Save your work for use in a later lab: save("StoreLabPt1.RData")**