Introduction

Every Tuesday after my piano lesson I go to my local Asian market. I've been religiously visiting them on a weekly basis for about a year now. I noticed they have issues with consistent inventory. This local Asian market is well loved and highly regarded in the community but looking on social media platforms, one of the biggest complaints that customers have is the unpredictability on the availability of their desired products. It seems the market has random imports of random things at random weeks. This can mean overstock of products that no one will buy and understock of staple goods.

This current inventory of trial and error of random import of products at random times seems costly as I'm sure the shipping costs are expensive. It's important to acknowledge the Covid situation and the geopolitical tensions in the Pacific as most of the products are shipped from that region. Of course grace has to be given to the supply chain issues. However, I still think a more thoughtful approach to inventory is best! Focusing on just one of the many products that the Asian market sells,ramen, I would like to propose a data driven inventory.

In order to provide meaningful and insightful data to the Asian market I am searching for a reliable and centralized source. To my luck this is available on Kaggle (https://www.kaggle.com/datasets/residentmario/ramen-ratings?datasetId=9366&sortBy=voteCount). The source actually originates from The Ramen Rater (https://www.theramenrater.com/) website.

Dataset

The dataset is a csv file. There is one review per row. There are 2580 rows of reviews excluding the header and seven columns which will be expanded to ten (more on that in the Data Cleaning section). The original columns are

- Review #- the ramen lover's review ID
- Brand- the ramen brand
- Variety- the ramen name
- Style- whether the ramen is packaged in a cup, bow, or tray
- Country- of the ramen
- Stars- perceived quality of the ramen on a scale of 5 (1 low, 5 high)
- Top Ten Brand- if the ramen ever made it to the top ten ranking and in what year

Data Cleaning

Preliminary scan

I typically spend around five minutes looking over a dataset to get a sense of what and how contents are recorded. This means looking at data types (integer, float, strings, characters, datetimes, etc). Then I evaluate the uniformity of the data to anticipate how much data manipulation (standardizing and modifying) is required. I also look for red flags such as multiple blank rows and columns or faulty data. The preliminary scan stage of the data cleaning process is a good practice as you can either save or waste time.

This dataset is not very big so some of the data cleaning process took place in excel. Also the data is quite organized so data cleaning will be minimal in terms of standardizing and modifying data. The original columns of the ramen-ratings.csv are: Review #, Brand, Variety, Style, Country, Stars, Top Ten. Using the “Top Ten” column I create three more columns to parse out the data values. The detailed data cleaning is done in RStudio Cloud using R.

Data modification: adding three customized columns

I sort the "Top Ten" column alphabetically making all the blanks rows fall to the bottom. The rows with data are coded as "Y" for yes, the blank rows are coded as "N" for no in the new column "Top Ten Y/N”. I create two more new columns to store the parsed data from the "Top Ten".

- The year is extracted and stored in the "Top Ten Year" column
    - =LEFT(J2,4) which grabs the first 4 characters
- The rank number was extracted and stored in the "Top Ten Rank" column
    - =RIGHT(J2,LEN(J2)-SEARCH("#",J2)) which grabs all characters after the hashtag sign

Then the entire table is copy pasted as value to lock in the values.

The rest of the data cleaning, analysis, and visualizing are done in R (https://github.com/ckraft-bot/RamenRatings/blob/main/Ramen.r)