Eric Penrose, Regan Maher, Reneise White  
10/23/2020  
ETL Project  
Written report

**Ramen Noodle Dataset**

**Project Outline**: Provide a Ramen Noodle analysis ready database. The data focused on ratings, countries, and words that will provide insight into whether these factors have impacts on the ratings.

**Final Project File Detail:  
*- Ramen\_Ratings\_1.csv* –** Kaggle dataset number 1  
- ***Ramen\_Ratings\_2.csv*** – Kaggle dataset number 2  
- ***Ramen\_Soup.py*** – Project script that includes scraper, transformer, cleaning, and loader.  
- ***Full\_Ramen.csv*** – Final dataset output from the Ramen\_Soup.py program.  
- ***RamenDB*** (database)– Mongo database.  
- ***RamenDB*** (collection) – Mongo collection that the dataset is stored.  
-***Written Report\_etl****.****docx*** – Final written report for project.

**Data Sources:**

- Kaggle - Ramen Noodle review CSV from theramenrater.com  
- Scraping - theramenrater.com for additional reviews and missing elements from the CSV.

**Cleansing and Transformation**:

- ***Country Clean Up*** - Some of the country data was misspelled and cities were identified as countries. Those items needed to be corrected in the data set prior to adding a continent column.

- ***Continent Column*** - Utilized the country column to developed a continent column.

- ***Review Word Count*** - Determine the length of the review (Blurb) column and did a len function to determine that total.

- ***Word Frequency*** - Initially placed the review (Blurb) column into a word cloud generator to determine the most frequently used words.

- **Word Standardizing** - Had to normalize some of the words as they were standard throughout all reviews (i.e., Finished, UPC, JAN, EAN)

- **Key Word Selection** - Once the word list was identified, programed to count the number of occurrences in each review for each word.

- **Bar Code Split** - The bar code information was part of the review (Blurb) and was typically found at the end of each review.

- **Bar Code Column Creation** - Created another column and did another split on the review to place the bar code info into the newly created column.

**Database Structure:** Mongo was used to store this dataset into a database.

**Future Questions for Analysis:** This dataset was built to help answer the following questions.

1. Which continents manufacture better tasting ramen based on reviews?

2. Which countries manufacture better tasting ramen based on reviews?

3. Are there particular words that are associated with a positive or a negative rating?

4. Does the length (word count) of a review rating equate to a positive or negative rating?

5. With the provided bar code information, how does price impact the ratings?