Gary Chen

IST-652

Homework 1: Structured Data

Date: 8/8/2021

**Homework 1 Write up**

Data and its source:

We'll be analyzing the Ramen-Rating data set.

This Data set was found on Kaggle.

<https://www.kaggle.com/residentmario/ramen-ratings>

The original file has 2580 rows and 7 columns.

Gary Chen- HW 1.ipynb – Jupyter notebook file

* Please read the code with Jupyter notebook. I used the “display” function to print out dataframes. I am not sure if other programs have the “display” function. I used “display” instead of “print” because it formatted the output of dataframes better.

**Data Cleaning, Exploration, and Transformation:**

I imported the ‘ramen-ratings.csv’ file and imported the pandas’ library to convert the csv into a dataframe. I checked the shape to see how many rows and columns I was working with. There are 2580 rows and 7 columns. I checked the column names and got back the first 5 rows of data with the head function. This gave me insight on what each column can mean. Below is a chart of how I interpret the meaning of each column. ‘Review #’ can be the order number but for the purposes of this exercise, I interpreted it as ‘Number of reviews on the specific Variety/Brand of Ramen’. I needed another numeric variable show the thinking behind calculations and exploring datasets.

|  |  |
| --- | --- |
| **Columns** | **Description** |
| ‘Review #’ | Number of reviews on the specific Variety/Brand of Ramen |
| 'Brand' | Company or Brand that Manufactures the Ramen |
| 'Variety' | The Specific Description and flavor of the ramen |
| 'Style' | If the ramen comes in a Cup, Pack, Bowl, Tray, Box, Can or Bar |
| 'Country' | Country that the ramen is from |
| ‘Stars' | The average Rating of the ramen out of 5 stars |
| 'Top Ten' | If the Ramen was awarded a top 10 status in a given year |

I pulled the descriptive statistics from each of my numeric value columns which are ‘Review #’ and ‘Stars'. I saved them in their own variables and used the “describe” function. I found that the ‘Stars’ column was an object type and wasn’t returning mean, min, or max values. I used the “unique” function to find all distinct values from ‘Stars’. I found a string value of ‘Unrated’. I replaced all ‘Unrated’ values with 0.001. I wanted a float number so I can do calculations with the ‘Stars’ column. 0.001 is a small enough number where it shouldn’t skew my calculations and is unique in identifying ‘Unrated’ values. Then I converted the ‘Stars’ column to numeric type with the “to\_numeric” function. This allowed me to find the full descriptive statistics and do calculations with the ‘Stars’ column. I noticed that there were a lot of NULL values for the ‘Top Ten’ column. I checked for NULL values with the “isnull” function. I used the “loc” function to transform all NULL values from the ‘Top Ten’ column into ‘no’. Then I explore all values that do not equal to ‘no’. I found a couple of ‘\n’ values in the updated ‘Top Ten’ column. I didn’t think ‘\n’ denotes as a top 10 award, so I also replaced ‘\n’ values with ‘no’.

In the initial data exploration, I used the “value\_counts” function to get an idea of the overall counts and values for each column. The “describe” function gave me insight that there was an average of 1290 number of reviews for each ramen variety. The lowest number of reviews was 1 and the max was 2580 reviews. I noticed that the number of reviews is a good indicator on the popularity on a specific ramen variety. There were mostly 4.0, 5.0, and 3.75 ‘Stars’ for rating. I wanted to see which country had the most 5.0 ‘Stars’ rated ramens. I created a new dataframe (‘Ramen\_5Stars’) with only 5.0 stars rated ramen. Then I got the “value\_counts” by country which showed Japan, Malaysia, and Taiwan had the most 5.0 star rated ramen. This was good insight, but I couldn’t solely judge the best ramen or the top ramen cultured countries purely on ‘Stars’. It had to be a combination of all available data in my dataset to find answers to my questions. For ‘Brands’, I noticed that Nissin had by far the most variety. The countries with the most variety Japan, USA, and South Korea.

After applying these steps and exploring the data set a little bit, I was able to map out a path in transforming and calculating additional columns to do further analysis and answer my questions. I calculated a ‘Ramen Score’ for each variety that factors in ‘Review #’, ‘Stars’, and ‘Top Ten’ awards. I will go further in depth on these calculations in the later parts of the write up.

**Questions:**

The two main questions that I wanted to ask are below. I also have sub questions which helped me to better answer the main questions.

Which country has the strongest Ramen culture in terms of popularity, rating, brand, and variety?

- Which country has the most 5.0 stars rated ramens?

- Which Country and Brand has the most ramen variety?

- Which countries had the most ramen variety with top 10 awards?

- For each country, what're the highest rated ramen?

What're the top 10 best rated ramens based on rating and popularity?

- Which style of ramen was most popular and best rated?

- Which country had the most top 1000 Ramen Scores?

To answer my questions, I couldn’t go by just the number of stars or top ten awards. I created my own way of scoring each ramen based on popularity, ratings, and awards. I measured popularity based on the number of reviews the ramen variety receives. Ratings is based on stars and awards is based on if they received a top ten award. ‘Ramen Score’ was created to rank ramen variety and the ramen culture from each country. I got a lot of insights from the overall dataset, but I needed additional insights on purely the best ramen variety. I created a subset (‘Top\_Ramen’) of the top 1000 ramen based on the highest 1000 ‘Ramen Score’.

**Program Description/ Data Transformation (Continued):**

Before calculating the ‘Ramen Score’ column, I created the ‘Top Ten – extra pts’ column. I wanted to award ramen varieties that have won a top 10 award with an additional of 500 points. I got that number from valuing the top 10 award with 100 unscaled reviews of 5 stars (100 \* 5 = 500). I imported the numpy library to use the “where” function to create the ‘Top Ten – extra pts’ column. The condition is if the ramen was top 10 awarded, the value will be 500 and if it wasn’t, the value will be 0. At first, the formula to calculate ‘Ramen Score’ was:

('Review #'/(3.65)) \* 'Stars' + 'Top Ten - extra pts'.

I needed to scale down the weight of the number of reviews since there was such a huge standard deviation and range. If the number of reviews was unweighted, then it would’ve skewed the scores for varieties that have a lot of reviews but very low star ratings. The scale of 3.65 was the divider because that was the average ‘Stars’ rating. After calculating the ‘Ramen Score’ with the 3.65 divider, it still rewarded number of reviews too highly. I decided to use 5.475 as the divider which was divvy from 3.65 \* 1.5. The final formula to calculate ‘Ramen Score’ is:

**ramen\_df['Ramen Score']** = ((ramen\_df['Review #'] / (3.65\*1.5)) \* ramen\_df['Stars'] + ramen\_df['Top Ten - extra pts']).

This yielded better and fairer results. This was shown in from the descriptive statistics on the ‘Ramen Score’ based on the 2 different formulas and from sorting the data frame.

As previously stated, I created a subset (‘Top\_Ramen’) of the top 1000 ramen based on the highest 1000 ‘Ramen Score’. Within the ‘Top Ramen’ dataframe, I did additional calculations and sorted the top ramen variety to get additional insights. In Order to get the highest ‘Ramen Score’ rated from each country, I did a “groupby” function by ‘Country’ to get the max ‘Ramen Score’. This output only gave me the max ‘Ramen Score’ by ‘Country’ but didn’t include which ramen variety received the score. I did a “for loop” to “append” the max ‘Ramen Scores’ to an empty list (‘filter\_score\_list’). Then used the “isin” function with the ‘filter\_score\_list’ to see which ‘Brand’, ‘Variety’, and ‘Style’ these ‘Ramen Scores’ belong to. This showed where each countries best ramen ranked on the global scale.

**Conclusion and Results:**

The results showed that the Nissin ‘Brand’ has ramen varieties in multiple countries. Nissin had 381 ramen variety and in the “head” exploration of the dataset, Nissin’s country is USA. That isn’t possible since the USA only has 323 varieties. This shows that a ‘Brand’ might not be indicative in measuring a country’s ramen culture. Japan, Malaysia, Taiwan, South Korea, and Singapore had the most 5.0 stars rated ramens. Even though USA had the 2nd most ramen varieties, it didn’t crack the top 5 in 5.0 stars or top 10 awarded ramens. The USA did rank number 5 in having the most ramen varieties in the ‘Top\_Ramen’ subset. USA did have 3 of the top 10 Ramen Scores but It doesn’t have the best quality relative to its massive quantity of varieties. The most popular ‘Style’ of quality ramen is by far pack ramen. In ‘Top\_10\_Ramen’ highest scored Ramen, 8 out of the top 10 ‘Ramen Scores’ was pack. The best Variety flavors in ‘Top\_10\_Ramen’ are Creamy, Shoyu, and Curry.

I ranked 1st Malaysia, 2nd Singapore, and 3rd Japan in having the best Ramen culture. All three finished in the top 5 of most 5.0 stars rated ramen and are in the top 3 of most top 10 awarded ramens. Singapore had the 3 of the top 10 ‘Ramen Scores’(‘Top\_10\_Ramen’) and Malaysia had 1 of the top 10 ‘Ramen Scores’ (‘Top\_10\_Ramen’). Japan and Malaysia are ranked 1 and 2 in having the most varieties in the “Top\_Ramen” subset. Considering the quality relative to the quantity of varieties from each of these countries, I ranked them in that order I stated before. While Japan did have the most overall varieties, variety with 5.0 stars, and most varieties in the ‘Top\_Ramen’ subset, Japan ranked 7 in the max ‘Ramen Score’ ranking (“Top\_Ramen\_filtered\_by\_country”). Singapore and Malaysia ranked higher in that subset (“Top\_Ramen\_filtered\_by\_country”). It was a close race, but Singapore and Malaysia are smaller countries, and the data suggest that they have a higher quality control over their varieties.

**Answers to sub questions:**

Which country has the strongest Ramen culture in terms of popularity, rating, brand, and variety?

- Which country has the most 5.0 stars rated ramens?

Japan 74

Malaysia 54

Taiwan 48

South Korea 44

Singapore 36

- Which Country and Brand has the most ramen variety?

Nissin 381

Nongshim 98

Maruchan 76

Mama 71

Paldo 66

Japan 352

USA 323

South Korea 309

Taiwan 224

Thailand 191

- Which countries had the most ramen variety with top 10 awards?

Singapore 7

Malaysia 6

Japan 6

South Korea 5

Indonesia 4

Thailand 3

Taiwan 2

China 1

Hong Kong 1

Myanmar 1

USA 1

- For each country, what're the highest rated ramen?

Please see in cell In[24] (Jupyter)

What're the top 10 best rated ramens based on ratings and popularity?

* Please see in cell In[20] (Jupyter)

- Which style of ramen was most popular and best rated?

Pack 576

Cup 222

Bowl 176

Tray 20

Box 4

Can 1

Bar 1

- Which country had to the most in top 1000 Ramen Scores?

Japan 159

Malaysia 108

Taiwan 87

Hong Kong 82

USA 79