Ramen Exploration

December 4, 2020

[1]:
```
import matplotlib.pyplot as plt
```

[297]:
```
import pandas as pd
import janitor
```

[369]:
```
# Read in the dataset
df = pd.read_csv("ramen-ratings.csv")
```

[192]:
```
df.head()
```

[192]:

	Review #	Brand	Variety	Style	Country	Stars	Top Ten
0	2580	New Touch	T's Restaurant Tantanmen	Cup	Japan	3.75	NaN
1	2579	Just Way	Noodles Spicy Hot Sesame Spicy Hot Sesame Guan…	Pack	Taiwan	1	NaN
2	2578	Nissin	Cup Noodles Chicken Vegetable	Cup	USA	2.25	NaN
3	2577	Wei Lih	GGE Ramen Snack Tomato Flavor	Pack	Taiwan	2.75	NaN
4	2576	Ching's Secret	Singapore Curry	Pack	India	3.75	NaN

[193]:
```
df.columns
```

[193]:
```
Index(['Review #', 'Brand', 'Variety', 'Style', 'Country', 'Stars', 'Top Ten'],
dtype='object')
```

[374]:
```
df = df[df['Stars'] != 'Unrated']
df['Stars'] = df['Stars'].astype(float)
```

[375]:
```
df.head()
```

[375]:
```
   Review #           Brand  \
0      2580       New Touch
1      2579        Just Way
2      2578          Nissin
3      2577         Wei Lih
4      2576  Ching's Secret

                                                Variety Style Country  Stars  \
0                           T's Restaurant Tantanmen    Cup   Japan   3.75
1  Noodles Spicy Hot Sesame Spicy Hot Sesame Guan…  Pack  Taiwan   1.00
2                       Cup Noodles Chicken Vegetable   Cup     USA   2.25
3                       GGE Ramen Snack Tomato Flavor  Pack  Taiwan   2.75
4                                    Singapore Curry  Pack   India   3.75

  Top Ten
0     NaN
1     NaN
2     NaN
3     NaN
4     NaN
```

[377]:
```
average_ramen = df['Stars'].mean()
average_ramen
```

[377]:
```
3.6546759798214974
```

[197]:
```
good = df[df['Stars'] > average_ramen]
bad = df[df['Stars'] <= average_ramen]
```

[198]:
```
good.head()
```

[198]:
```
   Review #           Brand                                Variety Style  \
0      2580       New Touch           T's Restaurant Tantanmen    Cup
4      2576  Ching's Secret                     Singapore Curry  Pack
5      2575   Samyang Foods                Kimchi song Song Ramen  Pack
6      2574         Acecook  Spice Deli Tantan Men With Cilantro   Cup
7      2573     Ikeda Shoku                Nabeyaki Kitsune Udon  Tray

       Country  Stars Top Ten
0        Japan   3.75     NaN
4        India   3.75     NaN
5  South Korea   4.75     NaN
6        Japan   4.00     NaN
7        Japan   3.75     NaN
```

[199]:
```
bad.head()
```

[199]:
```
   Review #       Brand                                                Variety  \
1      2579    Just Way  Noodles Spicy Hot Sesame Spicy Hot Sesame Guan…
2      2578      Nissin                        Cup Noodles Chicken Vegetable
3      2577     Wei Lih                        GGE Ramen Snack Tomato Flavor
8      2572  Ripe'n'Dry                             Hokkaido Soy Sauce Ramen
9      2571        KOKA                The Original Spicy Stir-Fried Noodles

  Style    Country  Stars Top Ten
1  Pack     Taiwan   1.00     NaN
2   Cup        USA   2.25     NaN
3  Pack     Taiwan   2.75     NaN
8  Pack      Japan   0.25     NaN
9  Pack  Singapore   2.50     NaN
```

1 What Styles of Ramen are Good?

[200]:
```
df['Style'].unique()
```

[200]:
```
array(['Cup', 'Pack', 'Tray', 'Bowl', 'Box', 'Can', 'Bar', nan],
      dtype=object)
```

[201]:
```
ramen_by_style = df.groupby(['Style']).mean()
```

[202]:
```
ramen_by_style_good = ramen_by_style[ramen_by_style > average_ramen]
ramen_by_style_good['Stars'].dropna()
```

[202]:
```
Style
Bar     5.000000
Bowl    3.670686
Box     4.291667
Pack    3.700458
Name: Stars, dtype: float64
```

[203]:
```
ramen_by_style_bad = ramen_by_style[ramen_by_style < average_ramen]
ramen_by_style_bad['Stars'].dropna()
```

[203]:
```
Style
Can     3.500000
Cup     3.498500
Tray    3.545139
Name: Stars, dtype: float64
```

Looks like better-than-average ramen are: - Bar - Bowl - Box - Pack

And worse-than-average ramen are: - Canned - In a cup - In a tray

2 Where is Good Ramen From?

[204]:
```
# first of all, where is ramen from?
df['Country'].unique()
```

[204]:
```
array(['Japan', 'Taiwan', 'USA', 'India', 'South Korea', 'Singapore',
       'Thailand', 'Hong Kong', 'Vietnam', 'Ghana', 'Malaysia',
       'Indonesia', 'China', 'Nigeria', 'Germany', 'Hungary', 'Mexico',
       'Fiji', 'Australia', 'Pakistan', 'Bangladesh', 'Canada', 'Nepal',
       'Brazil', 'UK', 'Myanmar', 'Netherlands', 'United States',
       'Cambodia', 'Finland', 'Sarawak', 'Philippines', 'Sweden',
       'Colombia', 'Estonia', 'Holland', 'Poland', 'Dubai'], dtype=object)
```

[205]:
```
ramen_by_country = df.groupby(['Country']).mean()
#ramen_by_country.sort_values('Stars', ascending=False).head()

ramen_by_country
```

[205]:

Country	Review #	Stars
Australia	2005.000000	3.138636
Bangladesh	1953.857143	3.714286
Brazil	2093.600000	4.350000
Cambodia	1822.400000	4.200000
Canada	1361.512195	2.243902
China	1327.189349	3.421893
Colombia	1265.500000	3.291667
Dubai	753.000000	3.583333
Estonia	1099.500000	3.500000
Fiji	1644.500000	3.875000
Finland	1797.666667	3.583333
Germany	2055.518519	3.638889
Ghana	2545.000000	3.500000
Holland	769.250000	3.562500
Hong Kong	1693.423358	3.801825
Hungary	1929.777778	3.611111
India	1522.225806	3.395161
Indonesia	991.484127	4.067460
Japan	1375.928977	3.981605
Malaysia	1623.838710	4.154194
Mexico	2275.160000	3.730000
Myanmar	1879.000000	3.946429
Nepal	1223.071429	3.553571
Netherlands	1866.000000	2.483333
Nigeria	2502.000000	1.500000
Pakistan	1672.666667	3.000000
Philippines	613.893617	3.329787

```
Poland          972.000000  3.625000
Sarawak        1707.333333  4.333333
Singapore      1702.917431  4.126147
South Korea    1003.863192  3.790554
Sweden         1519.333333  3.250000
Taiwan         1296.339286  3.665402
Thailand       1141.120419  3.384817
UK             1217.797101  2.997101
USA            1086.730650  3.457043
United States  2007.000000  3.750000
Vietnam         858.870370  3.187963
```

[207]:
```
ramen_by_country_good = ramen_by_country[ramen_by_country['Stars'] >␣
↪average_ramen]
ramen_by_country_good['Stars'].sort_values().dropna()
```

[207]:
```
Country
Taiwan           3.665402
Bangladesh       3.714286
Mexico           3.730000
United States    3.750000
South Korea      3.790554
Hong Kong        3.801825
Fiji             3.875000
Myanmar          3.946429
Japan            3.981605
Indonesia        4.067460
Singapore        4.126147
Malaysia         4.154194
Cambodia         4.200000
Sarawak          4.333333
Brazil           4.350000
Name: Stars, dtype: float64
```

[257]:
```
ramen_by_country_bad = ramen_by_country[ramen_by_country < average_ramen]
ramen_by_country_bad['Stars'].sort_values().dropna()
```

[257]:
```
Country
Nigeria        1.500000
Canada         2.243902
Netherlands    2.483333
UK             2.997101
Pakistan       3.000000
Australia      3.138636
Vietnam        3.187963
Sweden         3.250000
Colombia       3.291667
```

```
Philippines    3.329787
Thailand       3.384817
India          3.395161
China          3.421893
USA            3.457043
Ghana          3.500000
Estonia        3.500000
Nepal          3.553571
Holland        3.562500
Dubai          3.583333
Finland        3.583333
Hungary        3.611111
Poland         3.625000
Germany        3.638889
Name: Stars, dtype: float64
```

3 Can we Predict whether Ramen will be good?

[213]:
```
good_ramen_style_list = ramen_by_style_good['Stars'].dropna().index.unique()
good_ramen_style_list_stars = ramen_by_style_good['Stars'].dropna().unique()

bad_ramen_style_list = ramen_by_style_bad['Stars'].dropna().index.unique()
bad_ramen_style_list_stars = ramen_by_style_bad['Stars'].dropna().unique()
```

[214]:
```
# maybe, might be tricky
df = pd.read_csv("ramen-ratings.csv")

# Let's modify our dataframe to have the mean values we got
good_ratings = dict(zip(good_ramen_style_list, good_ramen_style_list_stars))
bad_ratings = dict(zip(bad_ramen_style_list, bad_ramen_style_list_stars))

ratings = {**bad_ratings, **good_ratings}

ratings
```

[214]:
```
{'Can': 3.5,
 'Cup': 3.4984999999999995,
 'Tray': 3.545138888888889,
 'Bar': 5.0,
 'Bowl': 3.6706860706860707,
 'Box': 4.291666666666667,
 'Pack': 3.7004581151832467}
```

[215]:
```
df['Style_value'] = df['Style'].replace(ratings)
df['Style_label'] = df['Style'].replace(good_ramen_style_list, value=1)
df['Style_label'].replace(bad_ramen_style_list, value=0, inplace=True)
```

```
df.head()
```

[215]:

	Review #	Brand	Variety	Style	Country	Stars	Top Ten	Style_value	Style_label
0	2580	New Touch	T's Restaurant Tantanmen	Cup	Japan	3.75	NaN	3.498500	0.0
1	2579	Just Way	Noodles Spicy Hot Sesame Spicy Hot Sesame Guan…	Pack	Taiwan	1	NaN	3.700458	1.0
2	2578	Nissin	Cup Noodles Chicken Vegetable	Cup	USA	2.25	NaN	3.498500	0.0
3	2577	Wei Lih	GGE Ramen Snack Tomato Flavor	Pack	Taiwan	2.75	NaN	3.700458	1.0
4	2576	Ching's Secret	Singapore Curry	Pack	India	3.75	NaN	3.700458	1.0

[262]:
```
good_countries = dict(zip(ramen_by_country_good.index,␣
↪ramen_by_country_good['Stars']))

try:
    del ramen_by_country_bad['Review #']
except KeyError:
    print("Not there")

try:
    ramen_by_country_bad = ramen_by_country_bad['Stars'].dropna()
except KeyError:
    print("already droppped")

bad_countries = dict(zip(ramen_by_country_bad.index, ramen_by_country_bad.
↪values))
country_ratings = {**good_countries, **bad_countries}

country_ratings
```

```
Not there
already droppped
```

[262]:
```
{'Bangladesh': 3.7142857142857144,
 'Brazil': 4.35,
 'Cambodia': 4.2,
 'Fiji': 3.875,
 'Hong Kong': 3.8018248175182485,
 'Indonesia': 4.067460317460317,
 'Japan': 3.981605113636364,
 'Malaysia': 4.154193548387097,
 'Mexico': 3.73,
 'Myanmar': 3.9464285714285716,
 'Sarawak': 4.333333333333333,
 'Singapore': 4.126146788990826,
 'South Korea': 3.7905537459283383,
 'Taiwan': 3.665401785714286,
 'United States': 3.75,
 'Australia': 3.1386363636363637,
 'Canada': 2.2439024390243905,
 'China': 3.4218934911242602,
 'Colombia': 3.2916666666666665,
 'Dubai': 3.5833333333333335,
 'Estonia': 3.5,
 'Finland': 3.5833333333333335,
 'Germany': 3.638888888888889,
 'Ghana': 3.5,
 'Holland': 3.5625,
 'Hungary': 3.611111111111111,
 'India': 3.3951612903225805,
 'Nepal': 3.5535714285714284,
 'Netherlands': 2.4833333333333334,
 'Nigeria': 1.5,
 'Pakistan': 3.0,
 'Philippines': 3.3297872340425534,
 'Poland': 3.625,
 'Sweden': 3.25,
 'Thailand': 3.3848167539267022,
 'UK': 2.9971014492753625,
 'USA': 3.457043343653251,
 'Vietnam': 3.187962962962963}
```

[275]:
```
df['Country_value'] = df['Country'].replace(country_ratings)
df['Country_label'] = df['Country'].replace(list(good_countries.keys()),␣
 ↪value=1)
df['Country_label'].replace(list(bad_countries.keys()), value=0, inplace=True)

df.head()
```

[275]:
```
   Review #           Brand  \
0      2580       New Touch
1      2579        Just Way
2      2578          Nissin
3      2577         Wei Lih
4      2576  Ching's Secret

                                                Variety Style Country Stars  \
0                          T's Restaurant Tantanmen    Cup   Japan  3.75
1  Noodles Spicy Hot Sesame Spicy Hot Sesame Guan…  Pack  Taiwan     1
2                       Cup Noodles Chicken Vegetable   Cup     USA  2.25
3                       GGE Ramen Snack Tomato Flavor  Pack  Taiwan  2.75
4                                    Singapore Curry  Pack   India  3.75

  Top Ten  Style_value  Style_label  Country_value  Country_label
0     NaN     3.498500          0.0       3.981605              1
1     NaN     3.700458          1.0       3.665402              1
2     NaN     3.498500          0.0       3.457043              0
3     NaN     3.700458          1.0       3.665402              1
4     NaN     3.700458          1.0       3.395161              0
```

[274]:
```
df.to_csv('new_ramen_dataset.csv')
```

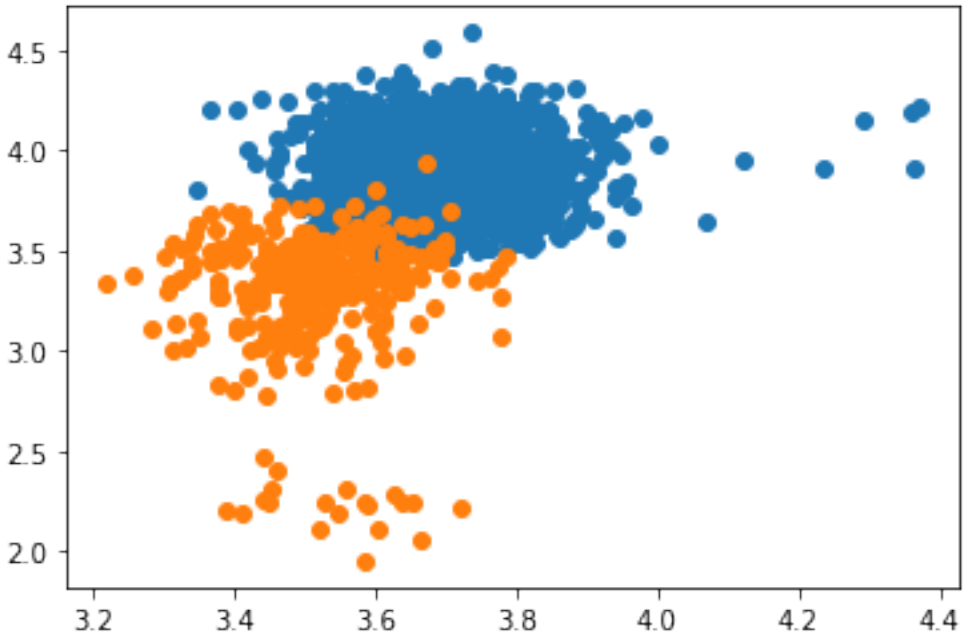
[308]:
```
# Adding some jitter to make it separate off the boundaries of the reviews
# This should help the k-means since it gives more significane to to the␣
 ↪distance metric used

df = janitor.functions.jitter(
    df=df,
    column_name='Style_value',
    dest_column_name='Style_value_jitter',
    scale=0.1,
    clip=None,
    random_state=None,
)

df = janitor.functions.jitter(
    df=df,
    column_name='Country_value',
    dest_column_name='Country_value_jitter',
    scale=0.1,
    clip=None,
    random_state=None,
)
```

[343]:
```
better_ramen_style = df[df['Style_value'] > average_ramen]
better_ramen_country = df[df['Country_value'] > average_ramen]
```

```
better_ramen = better_ramen_style.merge(better_ramen_country)


worse_ramen_style = df[df['Style_value'] <= average_ramen]
worse_ramen_country = df[df['Country_value'] <= average_ramen]
worse_ramen = worse_ramen_style.merge(worse_ramen_country)

#better_ramen['Style_value'].mean(), better_ramen['Country_value'].mean(),␣
↪worse_ramen['Style_value'].mean(), worse_ramen['Country_value'].mean()
(len(df) - (len(better_ramen) + len(worse_ramen)))
```

[343]: 1067

[342]:
```
plt.scatter(better_ramen['Style_value_jitter'],␣
↪better_ramen['Country_value_jitter'])
plt.scatter(worse_ramen['Style_value_jitter'],␣
↪worse_ramen['Country_value_jitter'])

plt.show()
```



[337]:
```
test_country = 'Taiwan'
test_style = 'Bowl'

test_country_2 = 'Korea'
test_style_2 = 'Pack'
```

```
def generate_ramen_point(country: str, style: str):
    try:
        style_value = ratings[style]
        country_value = country_ratings[country]
        print(style_value, country_value)
    except KeyError as e:
        print(f"No data on input {str(e)}")


generate_ramen_point(test_country, test_style)
generate_ramen_point(test_country_2, test_style_2)
```

```
3.6706860706860707 3.665401785714286
No data on input 'Korea'
```

[360]:
```
better_ramen['label'] = 'GOOD'
```

[361]:
```
worse_ramen['label'] = 'BAD'
```

[362]:
```
worse_ramen
```

[362]:

	Review #	Brand	Variety	Style	Country
0	2578	Nissin	Cup Noodles Chicken Vegetable	Cup	USA
1	2541	Nissin	Cup Noodles Very Veggie Spicy Chicken Flavor R…	Cup	USA
2	2538	Nissin	Cup Noodles Very Veggie Beef Flavor Ramen Nood…	Cup	USA
3	2535	Nissin	Cup Noodles Very Veggie Chicken Flavor Ramen N…	Cup	USA
4	2513	Pringles	Nissin Top Ramen Chicken Flavor Potato Crisps	Can	USA
..	…	…	…	…	…
293	157	Nissin	Sotanghon	Cup	Philippines
294	156	Maruchan	Instant Lunch Cajun Style With chili Piquin Sh…	Cup	USA
295	154	Nissin	Creamy Chicken	Cup	USA
296	153	Sapporo Ichiban	Shrimp Flavor	Cup	USA
297	142	Nissin	Chow Mein Teriyaki Beef	Tray	USA

```
    Stars Top Ten  Style_value  Style_label  Country_value  Country_label  \
0    2.25     NaN     3.498500          0.0       3.457043              0
1       5     NaN     3.498500          0.0       3.457043              0
2       5     NaN     3.498500          0.0       3.457043              0
3       5     NaN     3.498500          0.0       3.457043              0
4     3.5     NaN     3.500000          0.0       3.457043              0
..    …     …          …           …             …            …
293     2     NaN     3.498500          0.0       3.329787              0
294     3     NaN     3.498500          0.0       3.457043              0
295  1.75     NaN     3.498500          0.0       3.457043              0
296  1.75     NaN     3.498500          0.0       3.457043              0
297   4.5     NaN     3.545139          0.0       3.457043              0

     Style_value_jitter  Country_value_jitter label
0              3.487514              3.540047   BAD
1              3.699937              3.557153   BAD
2              3.472149              3.395145   BAD
3              3.458815              3.605592   BAD
4              3.463731              3.340618   BAD
..                  …                 …   …
293            3.436894              3.287358   BAD
294            3.545600              3.469000   BAD
295            3.348925              3.626128   BAD
296            3.319839              3.351370   BAD
297            3.454717              3.652747   BAD

[298 rows x 14 columns]
```

[363]:
```
final_df = better_ramen.append(worse_ramen)
```

[364]:
```
final_df.head()
```

[364]:
```
   Review #          Brand                                              Variety  \
0      2579       Just Way  Noodles Spicy Hot Sesame Spicy Hot Sesame Guan…
1      2577        Wei Lih                        GGE Ramen Snack Tomato Flavor
2      2575  Samyang Foods                               Kimchi song Song Ramen
3      2572     Ripe'n'Dry                             Hokkaido Soy Sauce Ramen
4      2571           KOKA                The Original Spicy Stir-Fried Noodles

  Style      Country Stars Top Ten  Style_value  Style_label  Country_value  \
0  Pack       Taiwan     1     NaN     3.700458          1.0       3.665402
1  Pack       Taiwan  2.75     NaN     3.700458          1.0       3.665402
2  Pack  South Korea  4.75     NaN     3.700458          1.0       3.790554
3  Pack        Japan  0.25     NaN     3.700458          1.0       3.981605
4  Pack    Singapore   2.5     NaN     3.700458          1.0       4.126147
```

	Country_label	Style_value_jitter	Country_value_jitter	label
0	1	3.841945	3.697278	GOOD
1	1	3.622133	3.767675	GOOD
2	1	3.627783	3.774688	GOOD
3	1	3.819662	3.996573	GOOD
4	1	3.930415	4.020176	GOOD

[365]:
```
final_df['label'].unique()
```

[365]:
```
array(['GOOD', 'BAD'], dtype=object)
```

[366]:
```
final_df.to_csv('final_ramen.csv')
```

[ ]:
```

```