IST 652 Final Project Report

Pet Food Market Analysis from Amazon

By Han Mo

Instead of using existing datasets as I did in mini projects, the data used here in final project is fetched from Amazon by applying the techniques of web scraping.

In the step of data preparation, too access data from amazon webpage, python requests package is required. There could be thousands of products for one single category. Therefore, to make the analysis more effective and manageable, I will fetch the first 20 pages of items. And the data attributes I get for analysis are 'product name', 'rating', 'rating count', 'price', 'unit_price', 'product_asin (Amazon Standard Identification Number, which is identical to every single product)', 'product_url', 'time_stamp'. In next step of data preparation, I scrape 4 categories of products, 'cat food', 'dog food', 'fish food' and 'bird food' (Treats are merged with food. See ipynb document), and data I fetched by that time was stored in CVS files. Before reading csv into dataframe, the package pandas, numpy and drive need to be imported and ready for next step. Then I can use pd.read_csv function to read the csv files. The top 5 lines of cat food is showed for demonstration purpose.

	product	rating	rating count	price	unit_price	product_asin	product url
0	Hill's Science Diet Dry Cat Food, Adult, Chick...	4.8 out of 5 stars	8,966	19.03	($4.76/lb)	B000084DWM	https://amazon.com/gp/slredirect/picassoRedire...
1	Blue Buffalo Sensitive Stomach Natural Adult D...	4.8 out of 5 stars	5,783	39.98	($2.67/lb)	B008EXERJK	https://amazon.com/gp/slredirect/picassoRedire...
2	Purina Beyond Grain Free, Natural, Chicken Adu...	4.7 out of 5 stars	4,533	45.99	($2.87/lb)	B07DWJYPRV	https://amazon.com/gp/slredirect/picassoRedire...
3	Meow Mix Original Choice Dry Cat Food	4.8 out of 5 stars	24,585	19.12	($1.52/lb)	B078229TVR	https://amazon.com/Meow-Mix-Original-Food-3-15...
4	Purina Friskies Canned Wet Cat Food 32 Count V...	4.7 out of 5 stars	22,843	20.78	($1.89/lb)	B00283MYSI	https://amazon.com/Purina-Friskies-Poultry-Adu...

There are some main operations I did for data preparation to be highlighted. Here I see some metrics like rating and unit price are very good measures of a product, but they are now constructed with a mix of numbers and string, which makes it very unfriendly for quantitive analysis. Therefore, in next step I are going to remove the unnecessary strings and split the numbers and units apart. In the meanwhile, I also want to replace the invalid information by 'N/A'. There is possibility that a product is going to appear on the web for multiple time because of advertising reason, so I remove duplicate on product_asin and set that as index. After all these

tidy-ups has been done, I concatenate all the sub-dataframe I have now and display the first several rows of prepared data for demonstration purpose again.

product_asin	product	rating	rating_count	price	unit_price	product_url	unit	currency
B000084DWM	Hill's Science Diet Dry Cat Food, Adult, Chick...	4.8	8966	19.03	4.76	https://amazon.com/gp/slredirect/picassoRedire...	lb	usd
B008EXERJK	Blue Buffalo Sensitive Stomach Natural Adult D...	4.8	5783	39.98	2.67	https://amazon.com/gp/slredirect/picassoRedire...	lb	usd
B07DWJYPRV	Purina Beyond Grain Free, Natural, Chicken Adu...	4.7	4533	45.99	2.87	https://amazon.com/gp/slredirect/picassoRedire...	lb	usd
B078229TVR	Meow Mix Original Choice Dry Cat Food	4.8	24585	19.12	1.52	https://amazon.com/Meow-Mix-Original-Food-3-15...	lb	usd
B00283MYSI	Purina Friskies Canned Wet Cat Food 32 Count V...	4.7	22843	20.78	1.89	https://amazon.com/Purina-Friskies-Poultry-Adu...	lb	usd
B001STX13U	Purina Fancy Feast Gravy Wet Cat Food Variety ...	4.7	24162	17.15	3.81	https://amazon.com/Purina-Fancy-Feast-Grilled-...	lb	usd
B00JN9IWNG	Purina Fancy Feast Gravy Lovers Poultry & Beef...	4.7	34308	22.10	3.93	https://amazon.com/Purina-Fancy-Feast-Poultry-...	lb	usd
B075QK5SP2	Sheba Perfect Portions Paté Wet Cat Food Tray ...	4.7	33384	20.98	5.38	https://amazon.com/Sheba-Wet-Food-Signature-Wh...	lb	usd
B01BMBPOR4	IAMS PROACTIVE HEALTH Adult Indoor Weight & Ha...	4.8	27281	24.98	1.56	https://amazon.com/Iams-Proactive-Health-Hairb...	lb	usd
B071RK5KZF	Sheba Perfect Portions Cuts in Gravy Wet Cat F...	4.7	28877	20.98	5.38	https://amazon.com/Signature-Delicate-Multipac...	lb	usd

Since the dataset is ready now, I are going to do some basic analysis of sub-dataset using pandas.DataFrame.describe function. This operation will tell us how much products are in each dataset, and some basic descriptive statistics of product rating and price. The results are attached below.

```
[1213] # cat
       cat_rdy_df.describe()
```

	rating	rating_count	price	unit_price
count	441.000000	441.000000	441.000000	441.000000
mean	4.546939	3916.589569	28.515170	5.755442
std	0.303384	6614.657954	13.748041	6.688057
min	0.000000	0.000000	1.090000	0.000000
25%	4.400000	495.000000	18.690000	2.070000
50%	4.600000	1484.000000	25.980000	4.200000
75%	4.700000	3938.000000	36.000000	6.850000
max	5.000000	52859.000000	84.970000	58.990000

In cat data, I have 441 products with the averaged rating of 4.55, averaged product price of 28.52 dollars and averaged unit price of 5.76 dollars per bls.

```
[1214] # dog
       dog_rdy_df.describe()
```

	rating	rating_count	price	unit_price
count	428.000000	428.000000	428.000000	428.000000
mean	4.578271	3627.586449	43.576706	5.054159
std	0.518257	4967.225613	22.646580	9.567197
min	0.000000	0.000000	5.240000	0.000000
25%	4.600000	610.000000	24.987500	1.930000
50%	4.700000	1841.500000	39.635000	2.630000
75%	4.700000	4547.500000	59.890000	4.000000
max	5.000000	41616.000000	144.530000	115.990000

In dog data, I have 428 products with the averaged rating of 4.58, averaged product price of 43.58 dollars and averaged unit price of 5.05 dollars per bls. Here I know dog foods always come larger in size than cat food.

```
[1215] # fish
       fish_rdy_df.describe()
```

	rating	rating_count	price	unit_price
count	209.000000	209.000000	209.000000	209.000000
mean	4.500000	1827.162679	27.473062	10.007703
std	0.411026	4630.782066	25.712747	55.945935
min	1.000000	1.000000	1.790000	-40.000000
25%	4.400000	48.000000	11.990000	0.580000
50%	4.600000	240.000000	18.990000	3.080000
75%	4.700000	1164.000000	31.080000	5.900000
max	5.000000	34867.000000	194.350000	714.290000

In fish data, I have 209 products with the averaged rating of 4.50, averaged product price of 27.47 dollars and averaged unit price of 10.01 dollars per bls.

```
# bird
bird_rdy_df.describe()
```

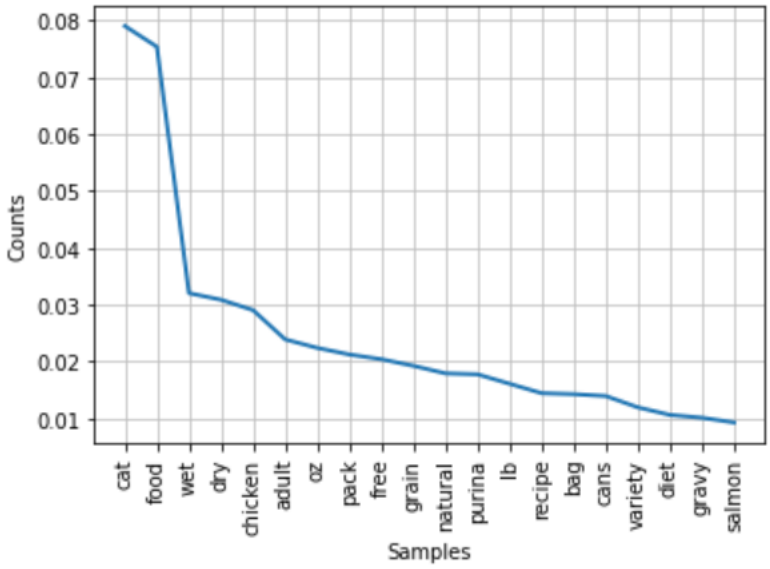
	rating	rating_count	price	unit_price
count	332.000000	332.000000	332.000000	332.000000
mean	4.528614	914.168675	27.239367	4.853313
std	0.425821	2905.817482	24.211773	9.200536
min	0.000000	0.000000	3.190000	-38.000000
25%	4.500000	71.750000	14.787500	0.397500
50%	4.600000	209.500000	20.295000	2.605000
75%	4.700000	704.750000	29.992500	6.420000
max	5.000000	41616.000000	232.990000	79.460000

In bird data, I have 332 products with the averaged rating of 4.53, averaged product price of 27.24 dollars and averaged unit price of 4.85 dollars per bls.

Question one, what are the key words favored by distributor in naming the product.

In this question, I want to have a deep look of product name. The product name here is the title of product made by the distributor. It usually contains more than just the name, but also some highlight information that merchandisers want to draw the customers attention from. First, I need to import nltk to analysis string. Stopword is used to removed unimportant words and collocations is imported for the situation when multiple words commonly co-occur. Before creating the list of product name, pandas.DataFrame.dropna was used to remove missing values, then I can tokenize each element in the list. After reorganizing and filtering using regular expression, I can use nltk.FreqDist.most_common to get the frequently used words, and use nltk.collocations.BigramAssocMeasures for bigram analysis (two words usually used together).

In the cat data list, I have 8519 tokenized words with 795 unique words, and the top 20 tokens are ['cat', 'food', 'wet', 'dry', 'chicken', 'adult', 'oz', 'pack', 'free', 'grain', 'natural', 'purina', 'lb', 'recipe', 'bag', 'cans', 'variety', 'diet', 'gravy', 'salmon']. Using plot.freqMAP, I can create a graph for better visualization. The Y axis shows the percentage of word frequency in the whole dataset.

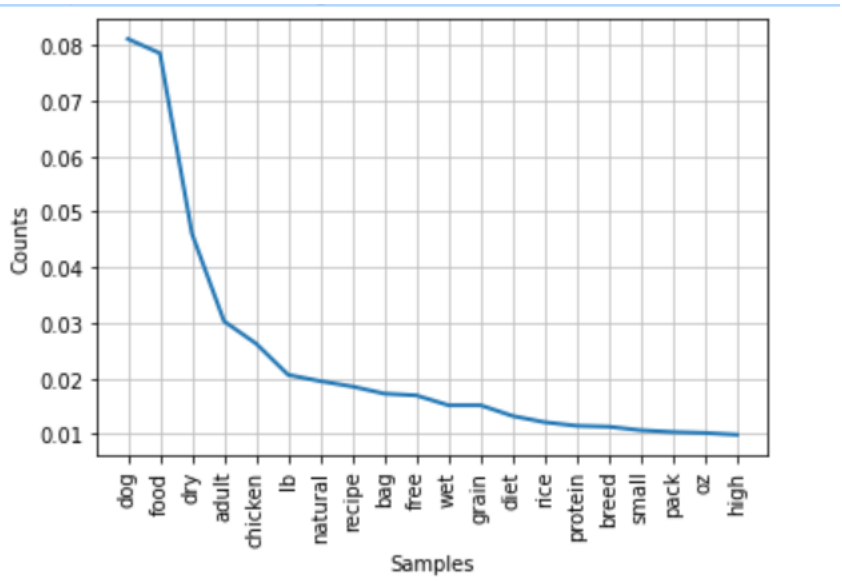


For bigram analysis for cat data list, the result comes as follow.

```
---------- Top 10 Bigram ------------
[(('cat', 'food'), 0.07020746887966806),
 (('wet', 'cat'), 0.028713692946058092),
 (('dry', 'cat'), 0.028049792531120332),
 (('grain', 'free'), 0.018921161825726143),
 (('oz', 'cans'), 0.012116182572614109),
 (('lb', 'bag'), 0.010622406639004149),
 (('variety', 'pack'), 0.008962655601659751),
 (('blue', 'buffalo'), 0.007800829875518672),
 (('high', 'protein'), 0.007302904564315353),
 (('food', 'chicken'), 0.006970954356846473)]
```

From the product name key word analysis for cat food, it's not hard to catch some valuable information like 'chicken' could be favored in recipe since it comes with high frequency, as is 'salmon'. The merchants also like to tag the product as 'natural', as it may help to build a concept of healthy recipe. In bigram analysis, I can see a lot of sellers want customer to know the products are grain free, high protein and chicken-made food.

Similarly, for dog food list, I have 791 unique words out of 8602 total words. The top 20 keywords are ['dog', 'food', 'dry', 'adult', 'chicken', 'lb', 'natural', 'recipe', 'bag', 'free', 'wet', 'grain', 'diet', 'rice', 'protein', 'breed', 'small', 'pack', 'oz', 'high'].



Most part is similar with cat, but surprisingly I have rice as a key word. Checking animal health website, it says that vets will recommend a bland diet of white rice (with no seasonings or fats) and boiled chicken for dogs with gastrointestinal issues. White rice is easy to digest and helps their digestive system relax and regain its balance so your dog can eat their regular kibble again. I also have 'breed' in the key word list which may suggest different breed of dogs are suggested with different food. And the top 10 bigram words are as below.

```
---------- Top 10 Bigram ------------
[(('dog', 'food'), 0.07392494765662748),
 (('dry', 'dog'), 0.044290545981639555),
 (('grain', 'free'), 0.01449508777580931),
 (('lb', 'bag'), 0.013528748590755355),
 (('wet', 'dog'), 0.013206635529070703),
 (('adult', 'dry'), 0.011596070220647446),
 (('high', 'protein'), 0.008697052665485585),
 (('food', 'chicken'), 0.007891770011273957),
 (("hill's", 'science'), 0.006120148172008375),
 (('science', 'diet'), 0.006120148172008375)]
```

There information from bigram analysis for dog food is very similar with dog food, grain free, high protein and chicken-made food is popular.

With the same logit, I have the top 10 key words in bird food name are 'food', 'bird', 'lb', 'birds', 'pound', 'wild', 'kaytee', 'pet', 'blend', 'parrot', 'daily', 'bag', 'pack', 'seed', 'natural', 'lbs', 'small',

'higgins', 'mix' and 'non-gmo'. Top 10 key words in fish food name are 'fish', 'food', 'tropical', 'oz', 'pellets', 'foods', 'koi', 'natural', 'aquatic', 'lb', 'flakes', 'dog', 'formula', 'protein', 'color', 'premium', 'floating', 'goldfish', 'dried', 'shrimp'.

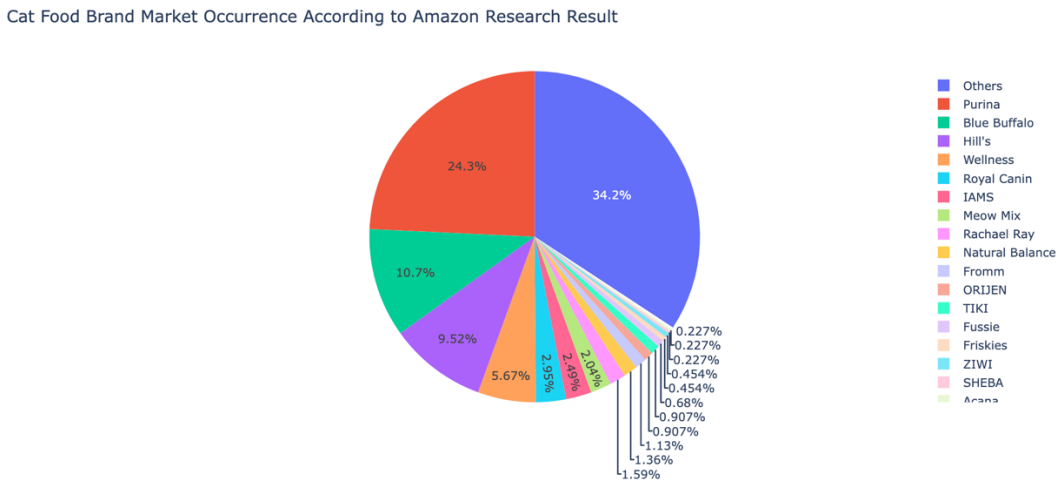
Question 2 Look at the market share of each pet brands in terms of their occurrence.

Unlike the Chinese ecommerce platform TAOBAO, Amazon is not showing number of transactions happening for each product. So, it is impossible to calculate their market share by definition. (Note: Market share is the percent of total sales in an industry generated by a particular company. Market share is calculated by taking the company's sales over the period and dividing it by the total sales of the industry over the same period) Instead, I am going to look at how many products of a brand is out there for selling on Amazon, which is given the name of brand occurrence here. Meanwhile, I can also look at the total number of ratings of each brand which could be a good indicator of times of purchase.

The first step to make 4 lists of popular brands nowadays on market for cat food, dog food, bird food and fish food individually and then I want to split the brand from its full product name by making a new column for it. It’s possible since the brand is usually the first word of the string. Next I are going to use some pandas.DataFrame functions to read the data grouped by brand. Here’s the result read for cat food brand and ranked by total rating counts. Note that Others are the collection of all the unlisted cat foods. Besides that, I could see Purina, Blue Buffalo and Hill’s are top 3 brands with most ratings which is highly possible that they are also the brands with the most purchase.

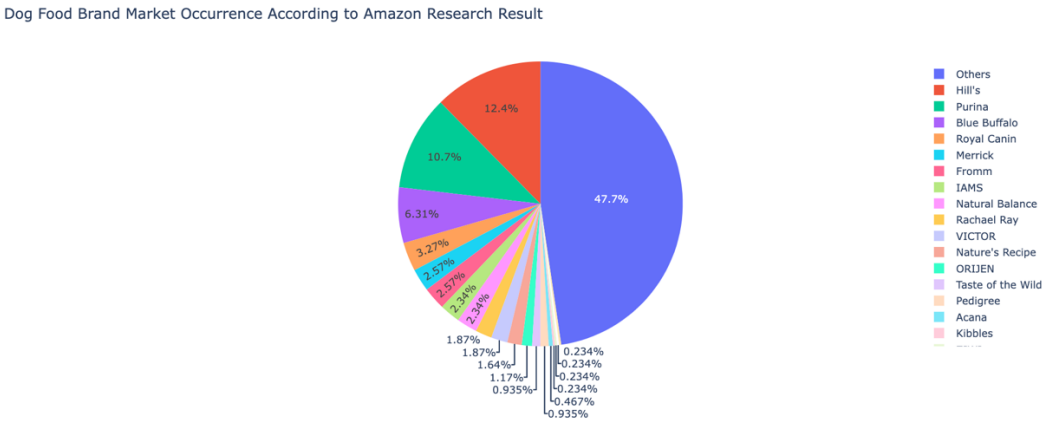
	brand	mean_rating	mean_rating_count	mean_price	mean_unit_price	occurrence	total_rating_count
0	Purina	4.629907	6383.177570	24.610561	4.816542	107	683000
1	Others	4.443046	3205.251656	26.571126	7.025232	151	483993
2	Blue Buffalo	4.519149	2601.914894	30.985106	4.625957	47	122290
3	Hill's	4.652381	2883.047619	39.414048	4.145714	42	121088
4	IAMS	4.727273	7780.818182	23.129091	3.209091	11	85589
5	Meow Mix	4.655556	7091.222222	13.901111	2.458889	9	63821
6	Rachael Ray	4.457143	6272.428571	17.462857	5.438571	7	43907
7	Friskies	4.700000	14778.500000	9.885000	6.860000	2	29557
8	Royal Canin	4.707692	1865.000000	40.752308	7.534615	13	24245
9	Wellness	4.440000	919.080000	32.518000	6.198000	25	22977
10	ORIJEN	4.625000	2680.000000	54.735000	6.892500	4	10720
11	SHEBA	4.700000	9377.000000	22.560000	5.780000	1	9377
12	Natural Balance	4.600000	1325.500000	29.896667	4.148333	6	7953
13	TIKI	4.500000	1918.500000	27.687500	8.467500	4	7674
14	Fussie	4.533333	1636.000000	36.070000	7.293333	3	4908
15	ZIWI	4.300000	1306.500000	41.285000	6.625000	2	2613
16	Acana	4.600000	2046.000000	19.990000	5.000000	1	2046
17	Halo	4.600000	973.000000	29.990000	5.000000	1	973
18	Fromm	4.840000	97.000000	36.254000	14.214000	5	485

Now using plotly I can pie chart the cat food brand by market occurrence. Besides Others, the top 3 places go to Purina, Blue Buffalo and Hill's. The percentage of the pie chart represent the total number of a brand's product over all the product in cat food product data list. For example, 24.3% for Purina means that 24.3 percentage of cat food product selling on Amazon comes from Purina. Surprised to see, the top 4 brands take almost half of the market on Amazon.



In the same way, for dog food product, I can have the ranking list by total rating count and pie chart for the market occurrence.

	brand	mean_rating	mean_rating_count	mean_price	mean_unit_price	occurrence	total_rating_count
0	Others	4.474020	2911.627451	43.046127	5.787990	204	593972
1	Purina	4.652174	6273.413043	32.931522	4.027826	46	288577
2	Hill's	4.698113	3484.264151	55.960189	5.569245	53	184666
3	Blue Buffalo	4.644444	5993.148148	33.721852	3.559630	27	161815
4	IAMS	4.720000	6055.200000	31.024000	4.867000	10	60552
5	Rachael Ray	4.600000	5605.750000	27.725000	2.158750	8	44846
6	Royal Canin	4.764286	3197.214286	59.196429	3.985000	14	44761
7	Taste of the Wild	4.700000	7756.000000	50.240000	2.112500	4	31024
8	Nature's Recipe	4.628571	4372.428571	18.018571	3.741429	7	30607
9	Pedigree	4.700000	7374.750000	19.332500	8.817500	4	29499
10	Natural Balance	4.620000	1970.200000	46.754000	2.793000	10	19702
11	Merrick	4.654545	1755.636364	51.412727	3.814545	11	19312
12	VICTOR	4.675000	2171.000000	59.240000	1.516250	8	17368
13	ORIJEN	4.660000	2205.200000	86.390000	4.638000	5	11026
14	Kibbles	4.700000	5141.000000	21.000000	0.680000	1	5141
15	ZIWI	4.600000	3892.000000	49.980000	22.720000	1	3892
16	Fromm	4.700000	173.090909	49.217273	6.445455	11	1904
17	Cesar	4.700000	1671.000000	45.690000	1.140000	1	1671
18	Acana	4.600000	784.500000	47.990000	4.290000	2	1569
19	Halo	4.700000	703.000000	19.990000	5.000000	1	703



Same for dog food, I could see Purina, Blue Buffalo and Hill's are top 3 brands with most ratings and possibly the brands with the most purchase. However, by market occurrence analysis I could know that the market for dog market is much more diversified. About half of the market is divided by 16 brands I have on the list, and other relatively small brands add up to the second half of market. It could be an information for an entrepreneur who is planning marching to pet market. Dog food market could be a better cut-in point since the market is not as monopolized as cat food market yet. The market size is favorable, and customer is willing to try various products.

The result for bird and fish market is quite different from dog and cat. For bird market, the top 3 brand with most rating count are Kaytee, ZuPreem and Lyric and the top 3 brand with most market occurrence are Kaytee, Harrison and ZuPreem. While in fish market, the top 3 brand with most rating count are Aqueon, Fluval and Omega One, and the top 3 brand with most market occurrence is Aqueon, Fluval and Zoo Med.

Question 3. User Review Key Words analysis

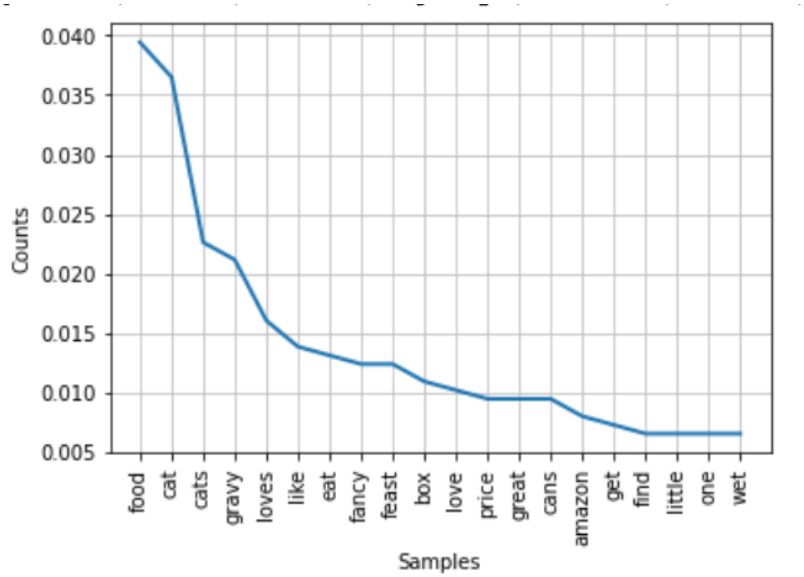
In this question, I am going to analyze the reviews made for selected product from 4 category of pet food. The selected product comes from question 2 with highest number of ratings of the brands with the most rating counts. For example, the chosen brand for Cat Food is the most rated product of Purina, which the brands with the most rating counts calculated from question 2.

In terms of technique, Selectorlib was introduced to scrap the Amazon product user reviews. I set the interval between each scrapping to 3 seconds, as Amazon web is protected with anti-scrapping service, to prevent the website from DDS attack. Also, I need to declare the review pages' urls in the urls.txt file, which I give a sample in the uploaded submission file. This is to allow the scrapper looping through the URLs to scrap reviews from each page. After then, I are using pandas.DataFrame to search in the product dataset, find the results belongs to the brand and get the URL for the one with the highest rating count.

After getting the url for our most rated product for cat, dog, fish and bird, I write the URL into a txt file and by using the review scraping technique published on Github, I will read the review from Amazon and store the information in CVS files for further analysis.

When interpreting the CVS files in jupyter notebook, normal tidy-up process is necessary, like remove invalid data. Using similar technique for question 1, NLTK is selected for string analysis. The tokenizers I made in question 1 will make this process much more efficient. Not only can I look at the review dataset as a whole, but also put the reviews for 3 categories, positive (rating in 4-5 stars), neutral(rating in 3 stars) and negative(rating in 1-2 stars) to look at the difference in these groups.

Look at the review for cat product. The program read 3158 tokenized words with 787 of them are unique. The top 20 words are 'food', 'cat', 'cats', 'gravy', 'loves', 'like', 'eat', 'fancy', 'feast', 'box', 'love', 'price', 'great', 'cans', 'amazon', 'get', 'find', 'little', 'one', 'wet'.



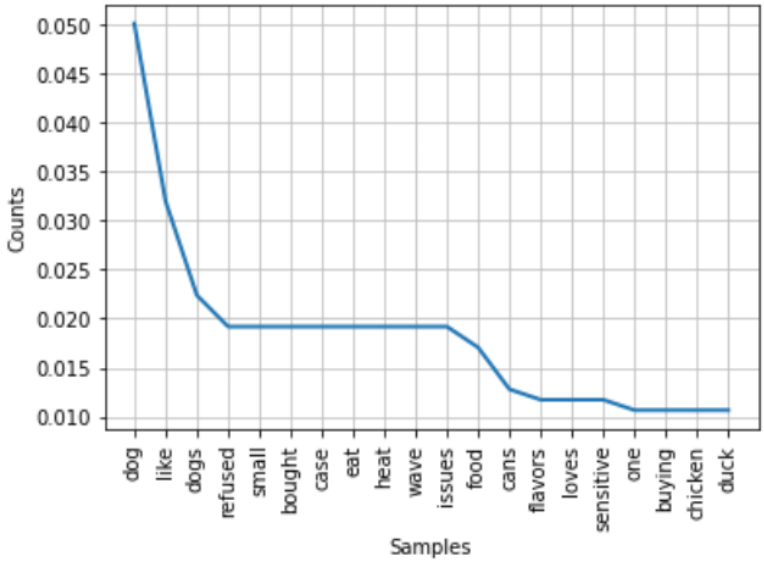
For bigram analysis, I have the top 10 paired word for the reviews of Purina can food.

```
---------- Top 10 Bigram ------------
[(('2x', 'normal'), 0.001610305958132045),
 (('3rd', 'parties'), 0.001610305958132045),
 (('a-lot', 'fussy'), 0.001610305958132045),
 (('able', 'order'), 0.001610305958132045),
 (('absolutely', 'stray'), 0.001610305958132045),
 (('add', 'thin'), 0.001610305958132045),
 (('address', 'olaf'), 0.001610305958132045),
 (('ago', 'would'), 0.001610305958132045),
 (('alone', 'works'), 0.001610305958132045),
 (('along', 'cup'), 0.001610305958132045)]
```

From the keyword, it is not hard to guess one of the reason this product so popular is because its reasonable price.

(The chosen brand for Cat Food is Purina, with the product https://amazon.com/Purina-Friskies-Seafood-Chicken-Variety/dp/B0777NRJ2N/ref=sr_1_16?keywords=cat+food&qid=1651895503&s=pet-supplies&sprefix=cat+food%2Cpets%2C84&sr=1-16)

For dog product, I have 2177 tokenized word with 243 of them are unique. The top 20 words are 'dog', 'like', 'dogs', 'refused', 'small', 'bought', 'case', 'eat', 'heat', 'wave', 'issues', 'food', 'cans', 'flavors', 'loves', 'sensitive', 'one', 'buying', 'chicken', 'duck'.



For bigram analysis, I have the top 10 paired word for the reviews of Purina ONE SmartBlend True Instinct Adult Canned Wet Dog Food.

```
---------- Top 10 Bigram ------------
[(('dog', 'eat'), 0.019189765458422176),
 (('heat', 'wave'), 0.019189765458422176),
 (('chicken', 'duck'), 0.010660980810234541),
 (('turkey', 'venison'), 0.010660980810234541),
 (('agrees', 'dogs'), 0.009594882729211088),
 (('also', 'refused'), 0.009594882729211088),
 (('amount', 'local'), 0.009594882729211088),
 (('assume', 'exposed'), 0.009594882729211088),
 (('awful', 'dog'), 0.009594882729211088),
 (('beef', 'bison'), 0.009594882729211088)]
```

From the key words analysis for this product, I can see consumers are happy about the recipe, since they are talking a lot for the ingredients like ‘chicken’, ‘duck’ and ‘beef’.

(The chosen brand for Dog Food is Purina, with the product https://amazon.com/Purina-Smartblend-Instinct-Tender-Chicken/dp/B01EYB3K1Q/ref=sr_1_121?keywords=dog+food&qid=1651895782&s=pet-supplies&sprefix=dog+food%2Cpets%2C84&sr=1-121)

Similarly, top 20 tokenized words for bird are 'millet', 'bird', 'package', 'loves', 'birds', 'love', 'open', 'seed', 'el', 'llegó', 'las', 'con', 'treat', 'sprays', 'old', 'like', 'knows', 'order', 'one' and 'well', and words for fish are 'fish', 'like', 'bottom', 'top', 'food', 'good', 'sink', 'pellets', 'get', 'love', 'right', 'around', 'catfish', 'loves', 'seem', 'excited', 'mess', 'open', 'pretty' and 'quickly'. For more information, please check out the coding result.

In the last step, I will look at the key words for high, neutral and low ratings. In this analysis, I put all the high rated review of cat, dog, bird and fish food together, so is for neutral and low rated reviews, to make it more efficient. But it could be a direction of future effect to break down the high, neutral and low rated reviews for different categories to make it more effective.
The result says the top 20 words for low rated reviews are 'millet', 'like', 'package', 'food', 'one', 'case', 'heat', 'open', 'dog', 'even', 'bought', 'seed', 'refused', 'small', 'wave', 'el', 'llegó', 'las', 'con' and 'cats'. Cooperate the knowledge learned from question 1, I could see a lot of low rating and reviews are coming from bird product. The reason that it is not providing a lot of valuable information could be there is not enough data or people who are so disappointed about a product do not even want a write a detailed review.

The top 20 words for neutral rated reviews are 'eat', 'dog', 'really', 'hungry', 'box', 'missing', 'cans', 'cats', 'delivery', 'damage', 'get', 'flavor', 'getting', 'would', 'chicken', 'turkey', 'price', 'replacement', 'need' and 'cat'. Here I see a lot of complains are resulted from missing items, delivery issue and product damage.

The top 20 words for highly rated reviews are 'food', 'fish', 'loves', 'like', 'love', 'cat', 'good', 'eat', 'bottom', 'gravy', 'right', 'cats', 'product', 'open', 'others', 'around', 'top', 'bird', 'dog' and 'issues'. Gravy is also welcomed in pet food.

As warp-up, this program is designed to be a data analysis script for analyzing merchandise on e-commerce platform. Though I walk it through with Amazon, but it is also applicable to other online shopping website like Amazon and BestBuy with necessary modification. With all the data reading and structuring process for pet food products here, I would have a deep understanding on how flexible and powerful Pandas could be for as a data analysis / manipulation library in Python.

What's more, I are also using pandas.DataFrame.to_csv function to store the data I get from question two as the summative information is very useful and presentative. I can almost tell a story about different brands' market performance and their competition based on that. Another technical highlight is NLTK which is leading platform for building Python programs to work with human language data by its authentic definition. Beside of simple summative data like sales or count of ratings, review could be more valuable as it contains with more information, like personal preference or explanation, but it is impossible to read thousands of reviews. Therefore, NLTK does a good job filtering and tokenizing word and extract the common characteristic of the tokens. Plotly is selected here for data visualization. There are of course more tools out there could do visualization, like Matplotlib and Seaborn. But Plotly is favored here because it is better at creating web-based data visualizations that can be displayed in Jupyter notebooks especially or web applications using Dash or saved as individual HTML files.

With the all the help from tools I mentioned above, I can answer question one, what are the key words favored by distributor in naming the product. Based on key word in top 20 list, I know the distributor usually advertise for their protein recipe if there is, like 'chicken' and 'salmon' for cat food, same 'chicken' for dog. For fish food it just says 'protein' and for bird food, they want to tell the consumer they are 'non-gmo'. However, I see a trend of 'grain free' in cat and dog food. After consulting the cat health care website, the information is 'Be aware that although gluten allergies and sensitivities are a widely discussed topic in human nutrition, these allergies are exceedingly rare in pet, and food allergies of any sort are also uncommon. Therefore, pet foods labeled as gluten-free or grain-free are not intrinsically healthier or better for your pet.' What's more, veterinarians recommended that dogs be fed a grain-inclusive diet unless there is a reason that makes it unsuitable. Could 'grain free' just a marketing trick? This could leave more space for research in the future.

For the second question, how's the market share of each pet brands. Purina, Blue Buffalo and Hill's are top three most popular brands measured by market occurrence and count of rating. Kaytee, ZuPreem, Harrison and Lyric are popular brands for bird food and Aqueon, Fluval, Zoo Med and Omega One are main players in fish food market. Based on my personal experience, I am quite satisfied with the result I have for cat food. One of my cat use to have eating problem and I consult the vet for food suggestion. Purina, Hill's Science and Royal Canin are highly

recommended. Some of the high-end brands advertise with high protein recipe but it could lose the nutrition balance. According to our vet, these three brands have well balanced recipe and they have been tested by a lot of consumers. This data analysis result partially verified this information which makes me very happy.

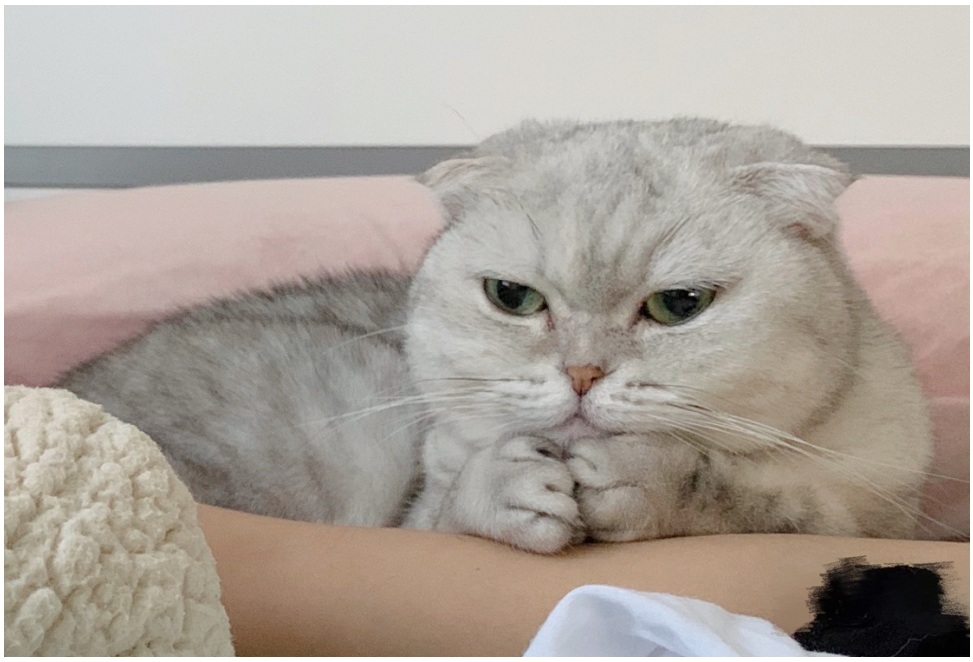
The third question is about key words analysis for review. For cat food buyer, they are talking about can food which come with gravy. Fancy Feast means to be popular choice especially for cat wet food. And consumer would very like to express their satisfaction if they come with reasonable price. For dog food buyer, the word sensitive indicates there could be the common concern among dog lovers for their pet's stomach sensitivity issue, and chicken and duck are favorable ingredients here. Millet is the most welcomed ingredient for bird food. And catfish might a popular breed among fish lover as it is one of the top 20 key words from fish food review. Finally, to improve the rating, the seller needs to pay more attention on missing items, delivery issue and product damage problems. Since the delivery is highly possible fulfilled by Amazon, the seller needs to cooperate with Amazon for a solution or improve the packaging from their side.

Reference:

https://github.com/scrapehero-code/amazon-review-scraper
https://financesonline.com/amazon-statistics/

FEEDBACK:

This final project is hard for me, but it is also a very chance of practice. I tried to engage the knowledge I learnt from the second half of the class and learned more from outside resource like Github. I really appreciate the freedom of the project as we can select whatever interested topic to conduct the analysis. Pets are adorable, so it is also pleasant for me to doing some research for me. It is for this course, and also for my cats.



(Her name is TWO-POINT-FIVE. She is my first cat, and she is 8 years old now)

(The name of white cat is SIXTEEN. He is 4 years old now)

(Hope all pets could enjoy healthy food and lead a happy life)