UGCA Group Assignment #1

Description

This analysis involved scraping conversation messages among different users on a car forum and analyzing them to see relationship between various car brands and their respective attributes. Also, different brands were analyzed with respect to each other to see which brands are frequently mentioned together.

Notes

- The code has been written used Python3 using Jupyter notebook as interface
- The code needs to be executed in a sequential fashion as in certain case previously created tables are referred to later

Approach

Conversion messages on a car forum were scraped using Selenium. Basic text processing techniques such as word tokenization, stopword removal were applied to get relevant words.

Using tokenized words data, lift ratio between different brands was calculated which is the ratio of number of times 2 brands were mentioned in a particular post divided by number of times each of them were mentioned in the post individually.

The highest lift ratio were then plotted on a Multidimensional scaling plot to see which brands are closely related with one another and mentioned frequently. To see which brands are closely linked with attributes such as performance, price, class, lift ratio was calculated in a similar fashion described above to see how closely these brands are associated with each of these variables. This analysis can be useful while branding a car and understanding close competitors.

The analysis can be further improved by calculating distances between two brands mentioned in a post and also distances between attributes and brands mentioned to see which brands/attributes are more closely associated.

Source Data used

- Start:
  https://forums.edmunds.com/discussion/18576/general/x/edmunds-members-cars-conversations/p2563
- End:
  https://forums.edmunds.com/discussion/18576/general/x/edmunds-members-cars-conversations/p2413
- Number of Comments Needed for Scraping: 150 Pages (30 comments for each page)

Top 10 brands by frequency from the analysis

1. ford
2. bmw
3. toyota
4. nissan
5. audi
6. kia
7. cadillac
8. honda
9. mercedes
10. Chrysler

```
In [23]:  top_10_brand_frequency_list = sortedWordsList[:10]
          top_10_brand_frequency_list

Out[23]:  [('ford', 128),
           ('bmw', 59),
           ('toyota', 50),
           ('nissan', 49),
           ('audi', 48),
           ('kia', 42),
           ('cadillac', 42),
           ('honda', 35),
           ('mercedes', 32),
           ('chrysler', 28)]
```

LIFT calculation in matrix form

```
            ford        bmw      toyota      nissan       audi
kia    \
ford     1.000000    2.950212    4.177500    7.104592    1.450521    9.9
46429
bmw      2.950212    1.000000    4.531525    3.082670   18.881356    1.7
98224
toyota   4.177500    4.531525    1.000000    7.275102    3.713333    6.3
65714
nissan   7.104592    3.082670    7.275102    1.000000    5.683673   25.9
82507
audi     1.450521   18.881356    3.713333    5.683673    1.000000    4.4
20635
kia      9.946429    1.798224    6.365714   25.982507    4.420635    1.0
00000
cadillac 2.486607   10.789346    4.243810    2.165209   11.051587    2.5
26077
honda    6.962500    8.631477   38.194286   12.991254    5.304762   12.1
25170
mercedes 3.263672   33.042373   11.140000    5.683673   23.208333    0.0
00000
chrysler 9.946429    5.394673    9.548571    6.495627    3.315476   11.3
67347

            cadillac       honda     mercedes    chrysler
ford        2.486607    6.962500    3.263672    9.946429
bmw        10.789346    8.631477   33.042373    5.394673
toyota      4.243810   38.194286   11.140000    9.548571
nissan      2.165209   12.991254    5.683673    6.495627
audi       11.051587    5.304762   23.208333    3.315476
kia         2.526077   12.125170    0.000000   11.367347
cadillac    1.000000    6.062585   19.892857    3.789116
honda       6.062585    1.000000   11.935714    9.093878
mercedes   19.892857   11.935714    1.000000    4.973214
chrysler    3.789116    9.093878    4.973214    1.000000
```

Multidimensional Scaling MAP
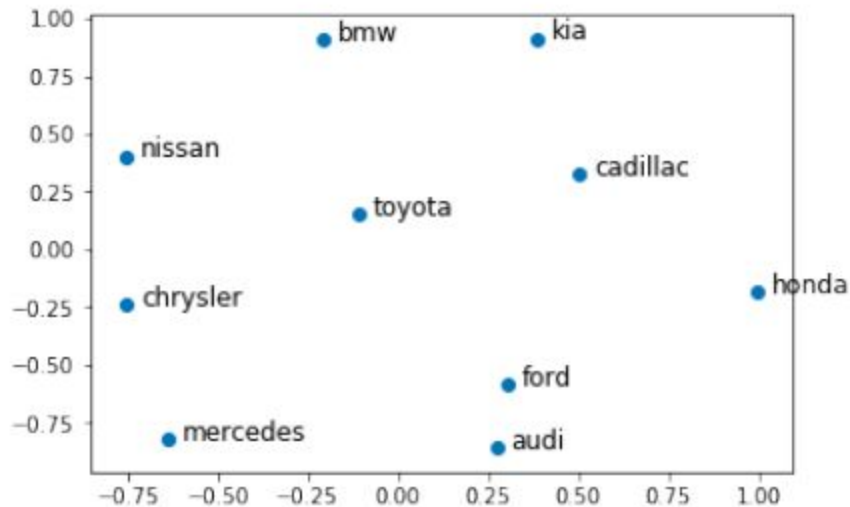
Insights for Ford brand manager:
Base on our "market basket analysis", the highest lift ratio for association (LRA) of ford is with Chrysler, and the second highest LRA is with Kia. The LRA indicates the likelihood of two brands being compared when one of the two brands is mentioned in a comment. In the past, Ford has always treated Chevron as their main competitor, but, in this specific analysis, Chevron is not ranked as one of the most popular brands. So, instead of focusing on creating competitive advantages against chevron, Ford should allocate more resources to create competitive advantages again Kia and Chrysler.

Insights for BMW brand manager:
Base on our "market basket analysis", the highest lift ratio for association (LRA) of BMW is with Mercedes. In fact, the LRA between BMW and Mercedes is the highest comparing to any other brand couples. The LRA indicates the likelihood of two brands being compared when one of the two brands is mentioned in a comment. In conclusion, Mercedes is the main competitor for BMW and the competition between these two brands are intense. The BMW's market share is going to increase significantly If BMW creates more competitive advantages against Mercedes.

Five mostly mentioned attributes of the car

1. Price (synonym for cheap,cheaper,costs,cost,economy,value,prices)
2. Class(synonym for luxury,interior,leather,brand,premium,seats,looks,seat)
3. Performance(synonym for engine,engines,power,acceleration,quality,sport,miles,torque,mileage,speed,sports,weight,suspension)
4. Trans(synonym for manual,auto,automatic,automatics,transmission)
5. Wd(synonym for awd,rwd,fwd,wheel,wheels)

```
In [61]:   # add the synonyms of the attributes to the current count

           for item in word_list:
               if(item[0]=="engine" or item[0]=='engines' or item[0]=='power' or
           item[0]=='acceleration' or item[0]=='quality'
                   or item[0]=='sport' or item[0]=='miles' or item[0]=='torque' o
           r item[0]=='mileage'
                   or item[0]=='speed' or item[0]=='sports'  or item[0]=='weight'
           or item[0]=='suspension'):
                   attribute_count['performance'] += 1
               elif(item[0]=='awd' or item[0]=='rwd' or item[0]=='fwd' or item[0]
           =='wheel' or item[0]=='wheels'):
                   attribute_count['wd'] += 1
               elif(item[0]=='manual'or item[0]=='auto'or item[0]=='automatic'or
           item[0]=='automatics'or item[0]=='transmission'):
                   attribute_count['trans'] += 1
               elif(item[0]=='cheap'or item[0]=='cheaper'or item[0]=='costs'or it
           em[0]=='cost'or item[0]=='economy'or item[0]=='value' or item[0] == 'p
           rices'):
                   attribute_count['price'] += 1
               elif(item[0]=='luxury'or item[0]=='interior'or item[0]=='leather'o
           r item[0]=='brand'or item[0]=='premium'
                   or item[0]=='seats'or item[0]=='looks'or item[0]=='seat'):
                   attribute_count['class'] += 1
```

Out[63]:

	attribute	freq
0	price	39572
1	class	3608
2	performance	941
3	trans	419
4	wd	143

<u>Advices for Ford product manager:</u>

Compare to other brands, Ford has relatively low lift ratio with the attribute performance. So, in order to create more competitive advantages, it is a good idea to improve the performance of Ford's product in general. On the other hand, the lift ratio between Ford and attribute luxury (class) is relatively low meaning customers are not associating Ford with luxury. In order to improve the brand image, it may be a good idea to improve the quality of Ford's product.

<u>Advices for BMW product manager:</u>

Among the five brands, BMW has the highest lift ratio with the attribute "class". The "class" attribute measures the luxury of a brand's product. In other words, BMW currently has the reputation as the most luxury brand among the five chosen brands. It is important for BMW to maintain the high quality of its products, so BMW can keep the competitive advantage against its competitors.

Aspirational Brand

The aspirational attributes we took for the analysis are :

1. Cost
2. Reliability
3. Safety
4. Mileage
5. Reputation

Suitable replacements made :

```
In [98]: for item in word_list:
             if(item[0]=="accuracy" or item[0]=='accurate' or item[0]=='authent
         icity' or item[0]=='authentic' or item[0]=='quality'
                 or item[0]=='honesty' or item[0]=='honest' or item[0]=='loyalt
         y' or item[0]=='loyal'
                 or item[0]=='trust' or item[0]=='trustworthy'  or item[0]=='re
         liable'):
                 aspiration_count['reliability'] += 1
             elif(item[0]=='safe' or item[0]=='secure' or item[0]=='assurance'
         or item[0]=='cover' or item[0]=='safeness'):
                 aspiration_count['safety'] += 1
             elif(item[0]=='amount'or item[0]=='expenditure'or item[0]=='rate'o
         r item[0]=='value'or item[0]=='charge'):
                 aspiration_count['cost'] += 1
             elif(item[0]=='miles'or item[0]=='miles/gallon'or item[0]=='distan
         ce'or item[0]=='economy'or item[0]=='value' or item[0] == 'prices'):
                 aspiration_count['mileage'] += 1
             elif(item[0]=='status'or item[0]=='stature'or item[0]=='esteem'or
         item[0]=='prestige'or item[0]=='renown'
                 or item[0]=='repute'):
                 aspiration_count['reputation'] += 1
```

Frequency of brand-aspirational-attribute :

```
In [106]: sorted_aspiration_d = sorted(aspiration_d.items(), key = itemgetter(1)
          , reverse = True)
          sorted_aspiration_d

Out[106]: [(('ford', 'price'), 19),
           (('nissan', 'price'), 12),
           (('bmw', 'price'), 10),
           (('bmw', 'class'), 6),
           (('ford', 'class'), 6),
           (('audi', 'price'), 5),
           (('toyota', 'trans'), 5),
           (('nissan', 'class'), 4),
           (('toyota', 'price'), 4),
           (('nissan', 'trans'), 4),
           (('ford', 'trans'), 4),
           (('ford', 'wd'), 4),
           (('audi', 'class'), 3),
           (('toyota', 'performance'), 3),
           (('ford', 'performance'), 3),
           (('toyota', 'class'), 2),
           (('bmw', 'trans'), 2),
           (('bmw', 'wd'), 2),
           (('audi', 'wd'), 2),
           (('bmw', 'performance'), 2),
           (('nissan', 'performance'), 1),
           (('audi', 'trans'), 1),
           (('nissan', 'wd'), 1)]
```

We use the top 5 brands to do this analysis. We want to find out which one is the best brand by measuring the cost, the mileage, the safety, the reliability and the reputation for each brand. From the result, we can see that Ford is the best brand. For safety and reputation, it remains the same as other brands which indicates a similar customer impression on all these brands. But Ford has the largest frequency for cost, mileage and reliability.Also if we take the cumulative sum of all the aspirational attributes for each brand, Ford is getting highest sum of attributes compared to the other brands. Hence we came to conclusion that FORD is best brand for the aspirational attributes point of view