INSY669 Text Analytics
Group Assignment

Group member
Yifan Lu
Emily Wu
Rohan Kumar
Kelly Liu
Addison Ji

1. Which forum you chose (provide URL):

The webpage we have selected from the forums is titled 'Chronic Car Buyers Anonymous,' accessible at the URL below. This group consists of individuals who frequently go car shopping and share their thoughts about various cars. Analyzing the discussions and preferences of this group provides valuable insights into consumer trends and preferences, which can significantly benefit automotive businesses in tailoring their marketing strategies and product development.
https://forums.edmunds.com/discussion/4011/general/x/chronic-car-buyers-anonymous

Task A: Identify top 10 brands by frequency. From the posts, calculate lift ratios for associations between the brands. You will have to write a script to do this task. Show the brands on a multi-dimensional scaling (MDS) map.

2. Which 10 brands you chose – provide the frequency table

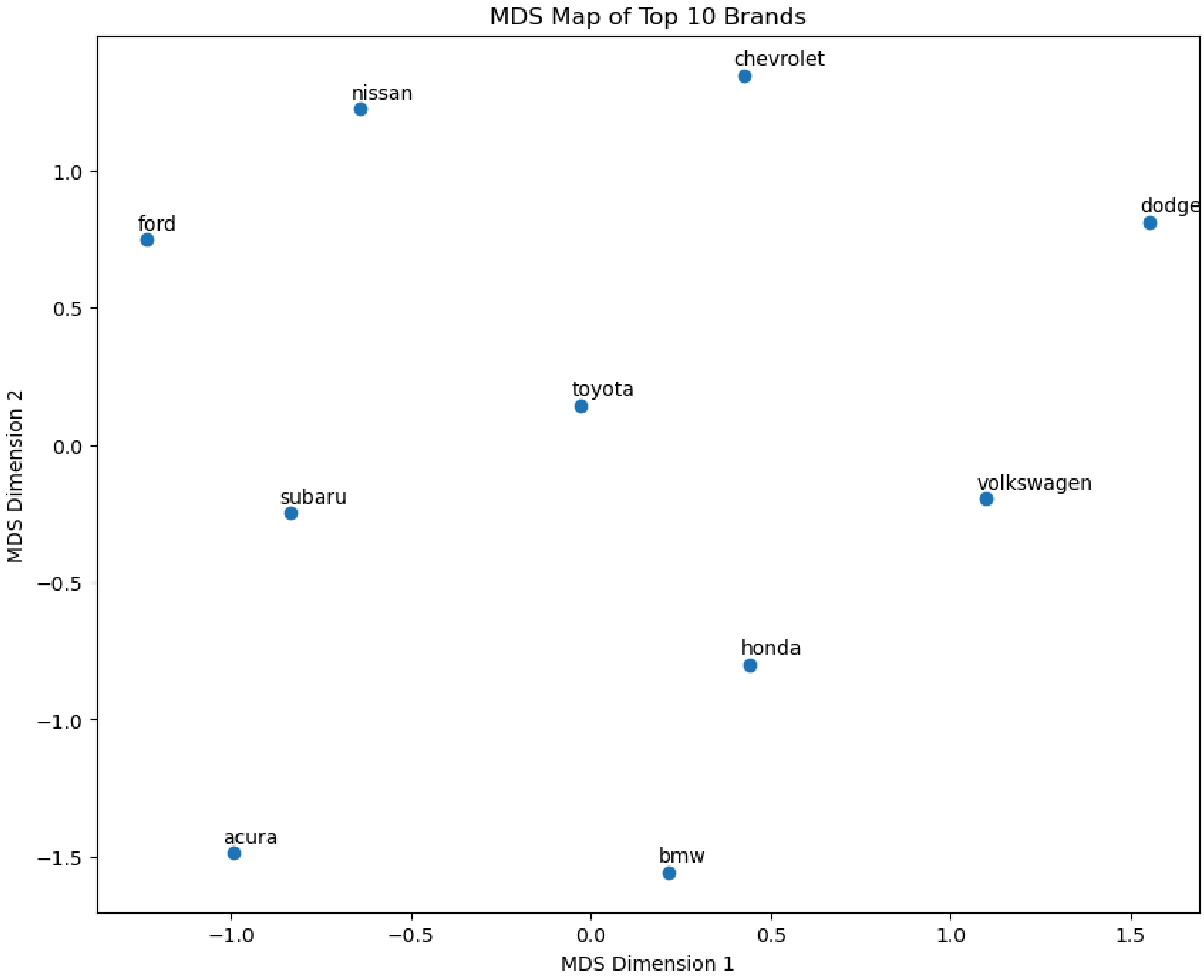
No.	Brand	Count
1	Toyota	2104
2	Honda	1799
3	Ford	741
4	Nissan	724
5	Dodge	657
6	BMW	640
7	Chevrolet	563
8	Acura	437
9	Subaru	397
10	Volkswagen	376

3. Show all lift values in a table.

No.	Brand 1	Brand 2	Lift Ratio
1	Nissan	Dodge	2.535802
2	Ford	Chevrolet	2.341419
3	Honda	Acura	2.296725
4	Subaru	Volkswagen	2.181253
5	Ford	Dodge	2.097622
6	Dodge	Chevrolet	2.000589
7	Ford	Volkswagen	1.991989
8	Chevrolet	Subaru	1.887155
9	Nissan	Chevrolet	1.797297
10	Ford	Nissan	1.793157
11	Acura	Subaru	1.791468
12	Chevrolet	Volkswagen	1.677941
13	BMW	Acura	1.666905
14	BMW	Volkswagen	1.660572
15	Toyota	Dodge	1.595280
16	Toyota	Nissan	1.583671
17	Toyota	Chevrolet	1.574266
18	Nissan	Volkswagen	1.549459
19	Ford	Subaru	1.534450
20	Honda	Subaru	1.533458
21	Honda	Nissan	1.499912
22	Toyota	Ford	1.495128
23	Toyota	Subaru	1.488349
24	Toyota	Volkswagen	1.440519

25	Acura	Volkswagen	1.396124
26	Toyota	Honda	1.390031
27	Toyota	Acura	1.319922
28	Honda	Ford	1.315621
29	Honda	Volkswagen	1.301846
30	Toyota	BMW	1.230989
31	Honda	Dodge	1.220871
32	Dodge	Subaru	1.219956
33	Nissan	Acura	1.216228
34	Nissan	Subaru	1.184297
35	Ford	BMW	1.154690
36	Honda	Chevrolet	1.117850
37	BMW	Chevrolet	1.067940
38	Honda	BMW	1.041203
39	Ford	Acura	0.936949
40	Dodge	Volkswagen	0.928625
41	BMW	Subaru	0.786366
42	Dodge	Acura	0.773225
43	Nissan	BMW	0.750604
44	Chevrolet	Acura	0.691783
45	Dodge	BMW	0.563166

4.MDS map.



Task B: What insights can you offer brand managers from your analysis in Task A? Choose two brands that you can offer the most interesting/useful insights for.

Nissan and Dodge have a lift ratio of 2.535802. This is the highest lift ratio in the list, suggesting that conversations about Nissan are more than 2.5 times as likely to also mention Dodge than we would expect if there were no association between the brands. This could mean that consumers are comparing these brands often, could be due to similarities in product offerings, price range, or marketing strategies. Brand managers at both Nissan and Dodge could delve deeper to understand the context of these mentions, identify the attributes being compared, and tailor their competitive positioning accordingly.

Ford and Chevrolet have a lift ratio of 2.341419. Ford and Chevrolet show the second strongest positive association. It suggests that discussions about Ford and Chevrolet occur together about 2.34 times more frequently than if their occurrences were statistically independent of each other. As historical competitors, particularly in the truck and muscle car segments, it's not surprising that they are frequently mentioned together. Brand managers could investigate the specific features, models, or campaigns that most often lead to these joint mentions. Understanding whether the sentiment of these mentions is positive or negative could also provide actionable insights for product development, marketing strategies, and customer relationship management.

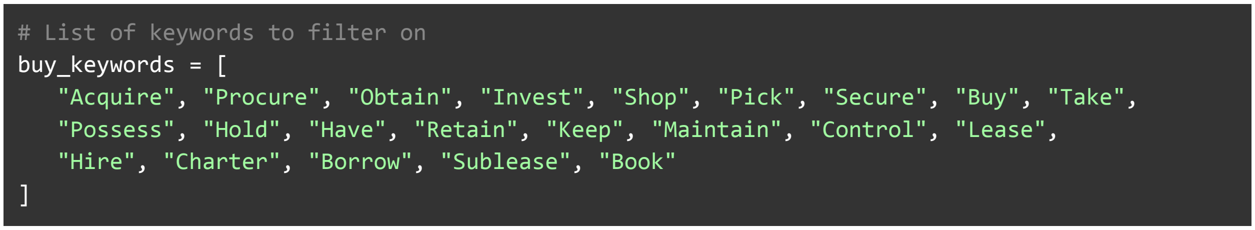
Task C: What are the 5 most frequently mentioned attributes of cars in the discussions? Note that the same attribute may be described by different words – e.g., pick-up and acceleration may both refer to a more general attribute, "performance". You have to make suitable replacements. Now pick the 5 most frequently mentioned brands. Which attributes are most strongly associated with which of these 5 brands?

5. State the 5 attributes you chose (again, a table is good here).

No.	Attribute
1	seat
2	mileage
3	tire
4	engine
5	interior

6. For task E, provide all details of your analysis – e.g., how you measured "aspirational" and how you found the most aspirational brand.

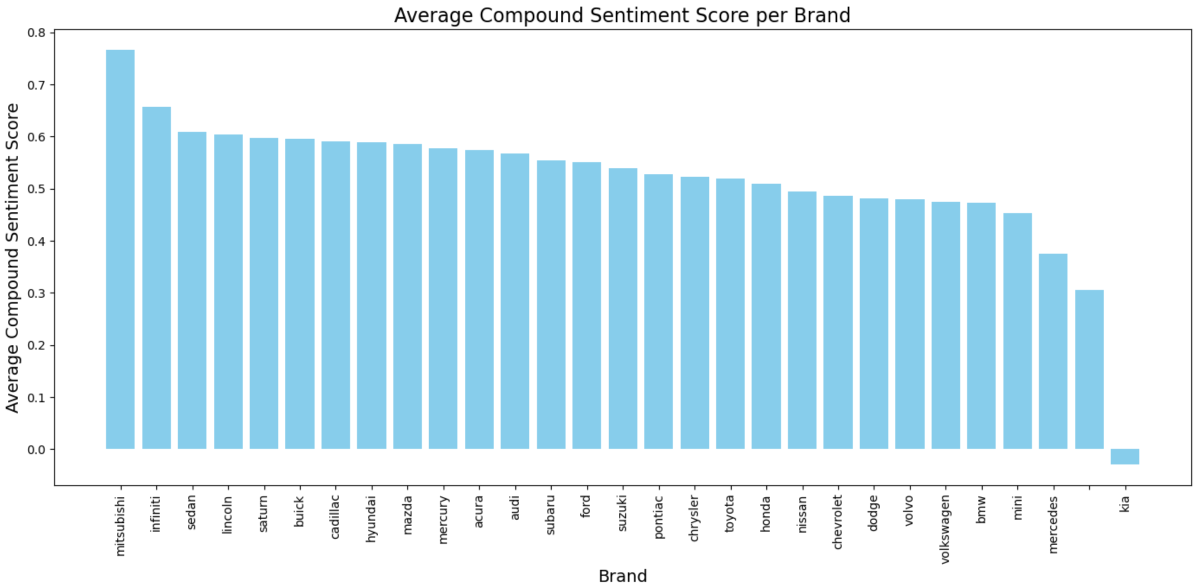
To capture the post related to the most aspirational brand in your data in terms of people actually wanting to buy or own. Our team first filtered the posts with keywords that represent buy or own. And then performed sentiment analysis to analyze the posts. The keywords that we used are listed as below. We did the analysis in 2 approaches as described below to measured "aspirational":



```
# List of keywords to filter on
buy_keywords = [
   "Acquire", "Procure", "Obtain", "Invest", "Shop", "Pick", "Secure", "Buy", "Take",
   "Possess", "Hold", "Have", "Retain", "Keep", "Maintain", "Control", "Lease",
   "Hire", "Charter", "Borrow", "Sublease", "Book"
]
```

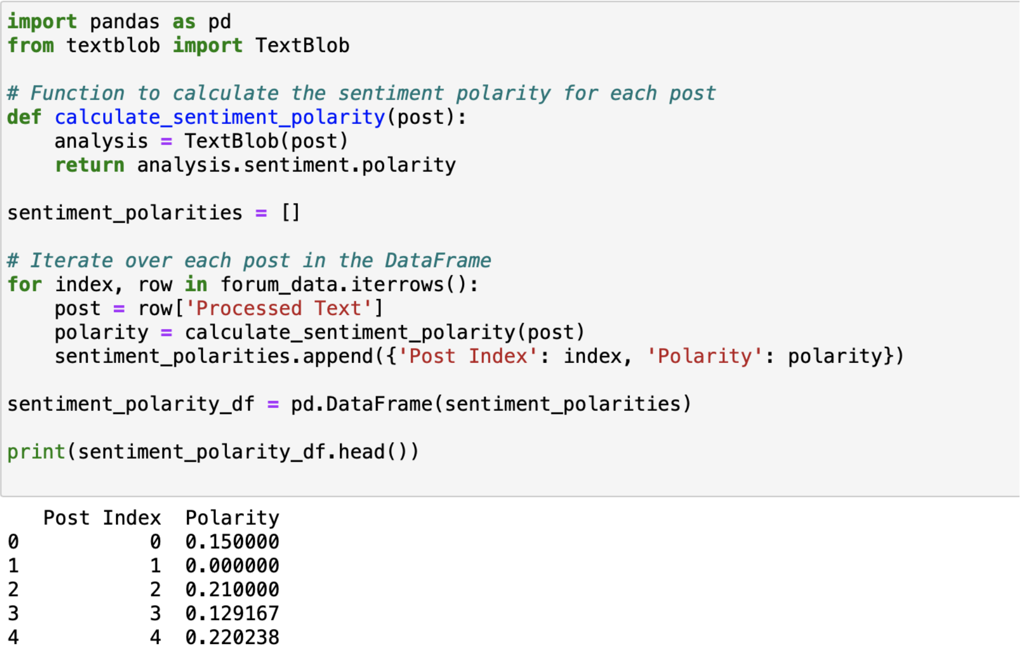
Approach 1: Compare the average of overall buy/own related post sentiment score per brand

1. Calculated the sentiment compound score per post
2. Took the average of the compound score and grouped by brands



Approach 2: Set up regression to see how does the attributes in buy/own related post contribute to the overall post sentiment score

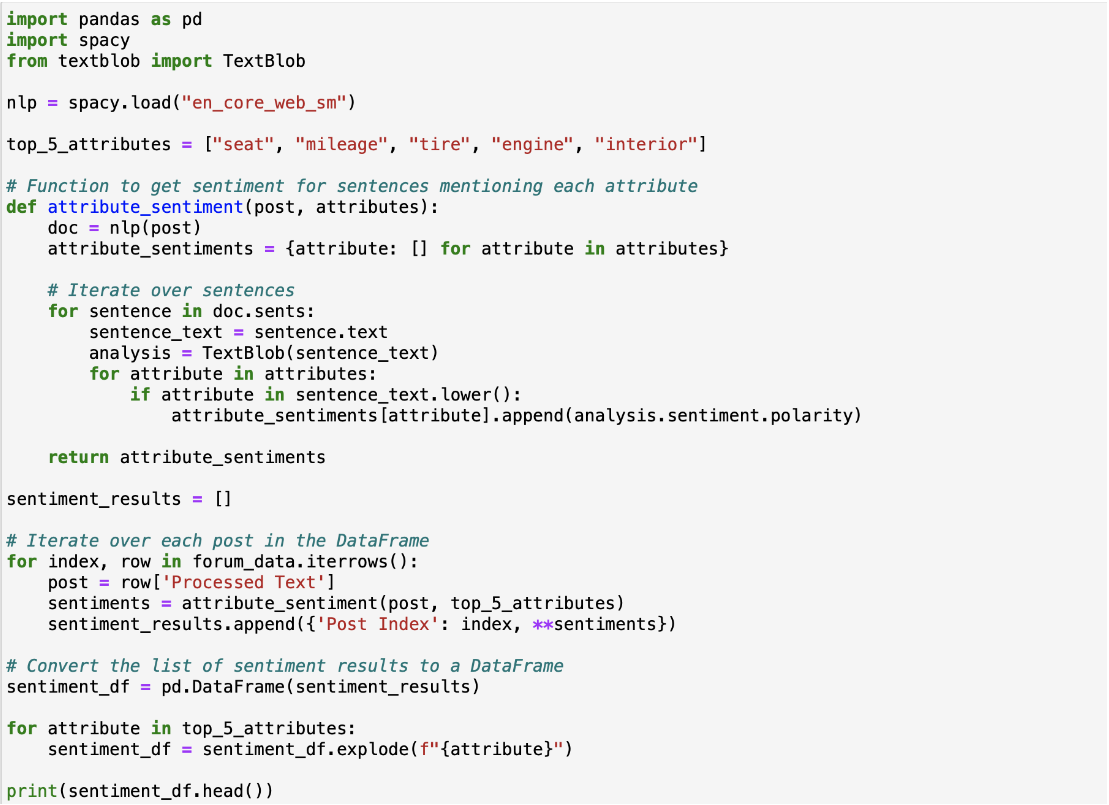
1. Calculated the sentiment compound score per post



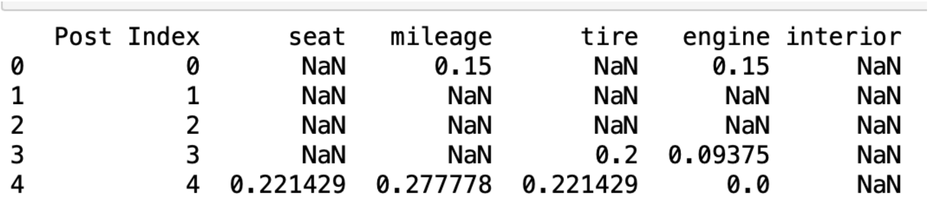
```
import pandas as pd
from textblob import TextBlob

# Function to calculate the sentiment polarity for each post
def calculate_sentiment_polarity(post):
    analysis = TextBlob(post)
    return analysis.sentiment.polarity

sentiment_polarities = []

# Iterate over each post in the DataFrame
for index, row in forum_data.iterrows():
    post = row['Processed Text']
    polarity = calculate_sentiment_polarity(post)
    sentiment_polarities.append({'Post Index': index, 'Polarity': polarity})

sentiment_polarity_df = pd.DataFrame(sentiment_polarities)

print(sentiment_polarity_df.head())
```

```
   Post Index  Polarity
0           0  0.150000
1           1  0.000000
2           2  0.210000
3           3  0.129167
4           4  0.220238
```

2. Calculated the sentiment compound score per attribute

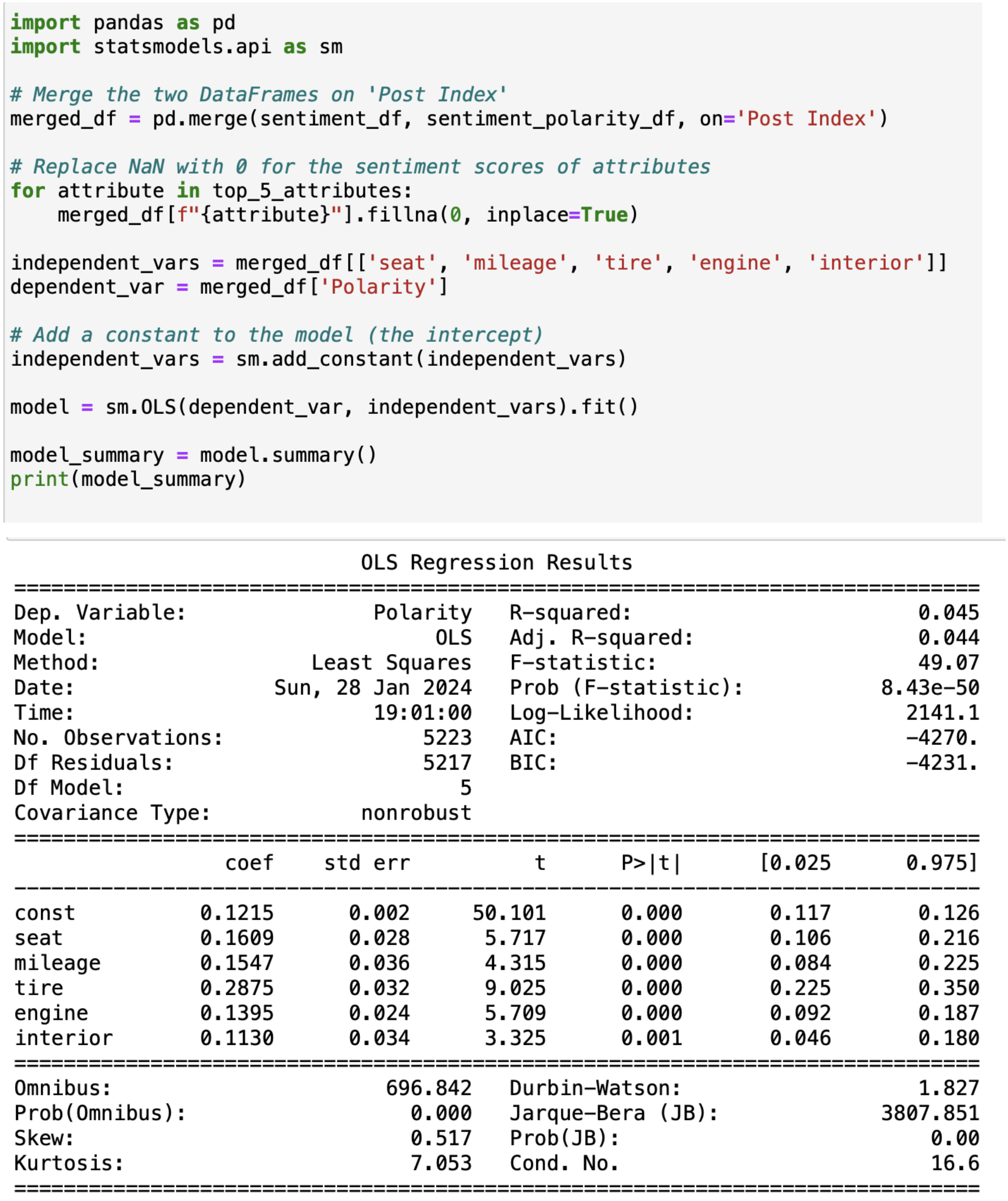


```
import pandas as pd
import spacy
from textblob import TextBlob

nlp = spacy.load("en_core_web_sm")

top_5_attributes = ["seat", "mileage", "tire", "engine", "interior"]

# Function to get sentiment for sentences mentioning each attribute
def attribute_sentiment(post, attributes):
    doc = nlp(post)
    attribute_sentiments = {attribute: [] for attribute in attributes}

    # Iterate over sentences
    for sentence in doc.sents:
        sentence_text = sentence.text
        analysis = TextBlob(sentence_text)
        for attribute in attributes:
            if attribute in sentence_text.lower():
                attribute_sentiments[attribute].append(analysis.sentiment.polarity)

    return attribute_sentiments

sentiment_results = []

# Iterate over each post in the DataFrame
for index, row in forum_data.iterrows():
    post = row['Processed Text']
    sentiments = attribute_sentiment(post, top_5_attributes)
    sentiment_results.append({'Post Index': index, **sentiments})

# Convert the list of sentiment results to a DataFrame
sentiment_df = pd.DataFrame(sentiment_results)

for attribute in top_5_attributes:
    sentiment_df = sentiment_df.explode(f"{attribute}")

print(sentiment_df.head())
```



```
   Post Index      seat   mileage      tire   engine interior
0           0       NaN      0.15       NaN     0.15      NaN
1           1       NaN       NaN       NaN      NaN      NaN
2           2       NaN       NaN       NaN      NaN      NaN
3           3       NaN       NaN       0.2  0.09375      NaN
4           4  0.221429  0.277778  0.221429      0.0      NaN
```

3. Setup a regression with the independent variable as the attribute sentiment scores and the dependent variable as the overall post's sentiment score



```python
import pandas as pd
import statsmodels.api as sm

# Merge the two DataFrames on 'Post Index'
merged_df = pd.merge(sentiment_df, sentiment_polarity_df, on='Post Index')

# Replace NaN with 0 for the sentiment scores of attributes
for attribute in top_5_attributes:
    merged_df[f"{attribute}"].fillna(0, inplace=True)

independent_vars = merged_df[['seat', 'mileage', 'tire', 'engine', 'interior']]
dependent_var = merged_df['Polarity']

# Add a constant to the model (the intercept)
independent_vars = sm.add_constant(independent_vars)

model = sm.OLS(dependent_var, independent_vars).fit()

model_summary = model.summary()
print(model_summary)
```

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:               Polarity   R-squared:                       0.045
Model:                            OLS   Adj. R-squared:                  0.044
Method:                 Least Squares   F-statistic:                     49.07
Date:                Sun, 28 Jan 2024   Prob (F-statistic):           8.43e-50
Time:                        19:01:00   Log-Likelihood:                 2141.1
No. Observations:                5223   AIC:                            -4270.
Df Residuals:                    5217   BIC:                            -4231.
Df Model:                           5                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.1215      0.002     50.101      0.000       0.117       0.126
seat           0.1609      0.028      5.717      0.000       0.106       0.216
mileage        0.1547      0.036      4.315      0.000       0.084       0.225
tire           0.2875      0.032      9.025      0.000       0.225       0.350
engine         0.1395      0.024      5.709      0.000       0.092       0.187
interior       0.1130      0.034      3.325      0.001       0.046       0.180
==============================================================================
Omnibus:                      696.842   Durbin-Watson:                   1.827
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3807.851
Skew:                           0.517   Prob(JB):                         0.00
Kurtosis:                       7.053   Cond. No.                         16.6
==============================================================================
```

With the 2 approaches, we can tell the overall sentiment per brand for buying and owning related posts. We can also determine which attributes contribute to the overall post's sentiment score the most.

7. Advice/insights based on your analysis for brand, product and advertising managers.

Task D: What advice will you give to a (i) product manager, and (ii) marketing/advertising manager of these brands based on your analysis in Task C?

For Product Managers:

Toyota:

- Focus on maintaining 'seat' quality and comfort, as it has a high sentiment score and lift.
- Carry out a customer satisfaction survey focused on tires, and devise strategies for enhancement based on the feedback received.

Nissan:

- The engine, despite its higher lift, exhibits a less competitive sentiment score to other attributes. Product managers should re-evaluate the product positioning (a comfort car or performance car) and explore potential improvements for the engine.

Honda:

- Seat and interior have similar sentiment scores and are frequently mentioned according to lift. The quality of these attributes must be consistently maintained, with ongoing investment in the budget.
- Consider improvements in 'tire', the attribute with the lowest sentiment score.

Ford:

- Emphasize 'interior' quality, as it has the highest sentiment score and is frequently mentioned.
- Ford's tires and engines are often mentioned but have much lower competitive sentiment scores across all brands. Product managers should plan improvements for these aspects.

Dodge:

- Focus on 'seat', as it's a strong point in terms of sentiment.
- Dodge's tires are less competitive among the brands according to the sentiment score. To increase sales growth and market share, PMs should consider investing more budget on these attributes to improve car performance and driving experience.

For Marketing/Advertising Managers:

Toyota:

- Highlight the comfort and quality of Toyota's 'seat' in campaigns.
- Develop strategies to improve the perception of 'tire'.

Nissan:

- Advertise the comfort of 'seat' and Counteract the lower sentiment for 'tire' through marketing.

Honda:

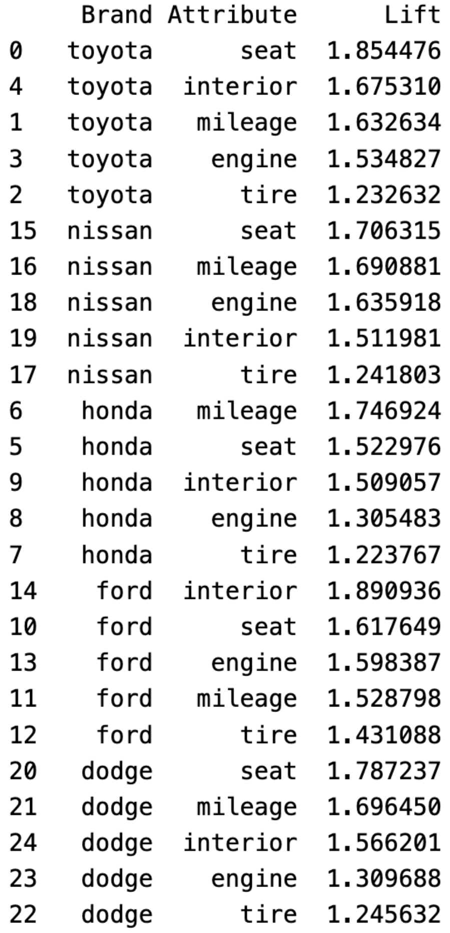
- Market the comfort ('interior').
- Create positive narratives around 'tire' and "engine" to enhance its perception.

Ford:

- Focus on the luxurious and comfortable 'interior' in advertising.
- Develop campaigns to address and improve the perception of 'tire' and "engine".

Dodge:

- Emphasize the interior in marketing campaigns.
- Address the lower sentiment in “tire” through strategic marketing efforts.



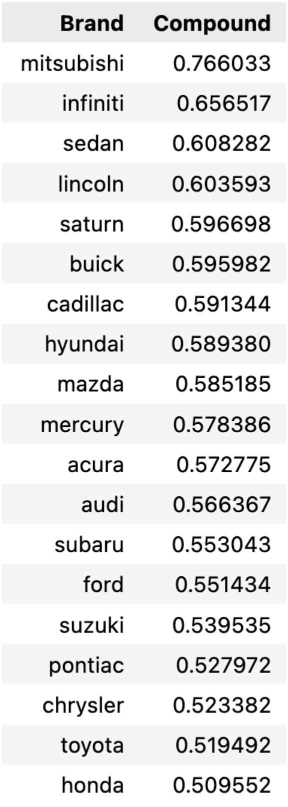
```
     Brand Attribute      Lift
0   toyota      seat  1.854476
4   toyota  interior  1.675310
1   toyota   mileage  1.632634
3   toyota    engine  1.534827
2   toyota      tire  1.232632
15  nissan      seat  1.706315
16  nissan   mileage  1.690881
18  nissan    engine  1.635918
19  nissan  interior  1.511981
17  nissan      tire  1.241803
6    honda   mileage  1.746924
5    honda      seat  1.522976
9    honda  interior  1.509057
8    honda    engine  1.305483
7    honda      tire  1.223767
14    ford  interior  1.890936
10    ford      seat  1.617649
13    ford    engine  1.598387
11    ford   mileage  1.528798
12    ford      tire  1.431088
20   dodge      seat  1.787237
21   dodge   mileage  1.696450
24   dodge  interior  1.566201
23   dodge    engine  1.309688
22   dodge      tire  1.245632
```

Task E: Which is the most aspirational brand in your data in terms of people actually wanting to buy or own? Describe your analysis. What are the business implications for this brand?

According to the highest compound sentiment score for the posts related to buy/own as listed below, **Mitsubishi** has the highest score, which suggests that it is the most aspirational brand in our data in terms of people actually wanting to buy or own.

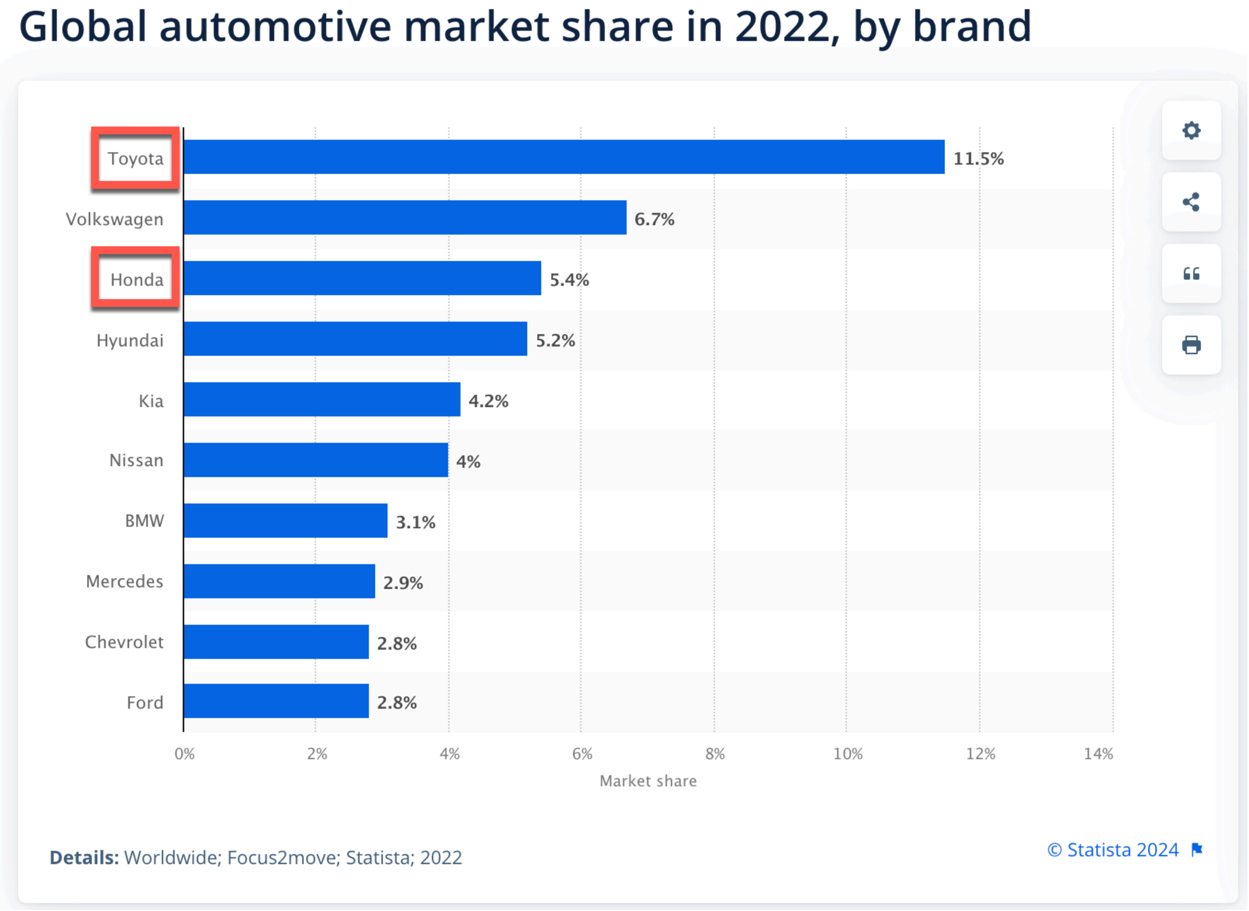
Something interesting is that Mitsubishi has the lowest frequency mentioned in the posts, but ends up with the highest sentiment score. Toyota and Honda have the exceptional highest frequency mentioned in the posts, but they both have around the average level of 0.5 sentiment score. Though we see the high skewness in frequency mentioned per brand data, the frequency mentioned in posts should be correlated with the market share as seen in the “Figure - Market Share” below. Therefore, the sentiment score of the posts should be reflecting the market aspiration on the cars.

Figure - Compound Sentiment Per Brand



Brand	Compound
mitsubishi	0.766033
infiniti	0.656517
sedan	0.608282
lincoln	0.603593
saturn	0.596698
buick	0.595982
cadillac	0.591344
hyundai	0.589380
mazda	0.585185
mercury	0.578386
acura	0.572775
audi	0.566367
subaru	0.553043
ford	0.551434
suzuki	0.539535
pontiac	0.527972
chrysler	0.523382
toyota	0.519492
honda	0.509552

Figure - Market Share



Global automotive market share in 2022, by brand

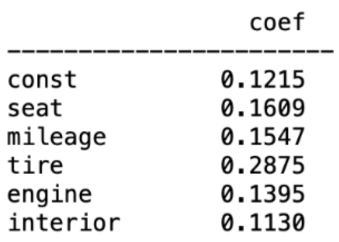
Details: Worldwide; Focus2move; Statista; 2022

© Statista 2024

https://www.statista.com/statistics/316786/global-market-share-of-the-leading-automakers/

To examine the influence of specific attributes on the overall sentiment of the posts, the tire attribute stands out with the most significant positive impact. This indicates that discussions related to buy/own actions and mentioned tires are associated with more positive sentiment scores. It may suggest that forum users place a high value on tire-related features or that conversations about tires are typically positive. Other attributes like seats, mileage, engines, and interiors also show positive associations with sentiment, though to a lesser extent. The strength of these associations, as reflected by the coefficients from the regression as listed in the chart below, provides insight into what forum users might prioritize or have stronger opinions about when discussing vehicles.

Figure - Coefficient Per Selected Attribute



```
                 coef
------------------------
const          0.1215
seat           0.1609
mileage        0.1547
tire           0.2875
engine         0.1395
interior       0.1130
```

Advice for Mitsubishi Brand Managers:

1. **Pay Attention to Tires and Seats:** Because the perception of "tire", "seat" has a big influence on the sentiment as a whole, make sure these components are comfortable and of excellent quality. Take into account to investigate more in Mitsubishi customer needs and satisfaction, and to improve according to the demand such as to improve the longevity and comfort of seats and tires.
2. **Emphasise Mileage:** Since many customers are concerned about mileage, perhaps emphasise cost savings and fuel efficiency in marketing initiatives.

Advice for Mitsubishi Product Managers:

1. **Product Development**: Prioritise tires and seats in the product development and enhancement strategies because these characteristics have a significant impact on customer sentiment.

Advice for Mitsubishi Advertising Managers:

1. **Marketing Campaigns**: In advertising campaigns, highlight the comfort and quality of the seats and tires. Goodwill in these areas contributes to a favourable view of the brand as a whole.