DATA MINING PROJECT

By:

- Parth Gandhi(pg521)
- Bhagyashree Parkar(bp526)
- Aditya Patwardhan(ap1991)
- Saranya Mantena(Sm2592)

Topic: **Adidas vs Nike**

INTRODUCTION

Data Set: adidas vs nike.csv

	Product Name	Product ID	\
0	Women's adidas Originals NMD_Racer Primeknit Shoes	AH2430	
1	Women's adidas Originals Sleek Shoes	G27341	
2	Women's adidas Swim Puka Slippers	CM0081	
3	Women's adidas Sport Inspired Questar Ride Shoes	B44832	
4	Women's adidas Originals Taekwondo Shoes	D98205	

	Listing Price	Sale Price	Discount	Brand	\
0	14999	7499	50	Adidas Adidas ORIGINALS	
1	7599	3799	50	Adidas ORIGINALS	
2	999	599	40	Adidas CORE / NEO	
3	6999	3499	50	Adidas CORE / NEO	
4	7999	3999	50	Adidas ORIGINALS	

The dataset consists of 3268 products from Nike and Adidas with 12 features of information including their ratings, discount, sales price, listed price, product description, and the number of reviews.

DATA CLEANING

- Check for Null values.

```
[7]   df.isnull().sum()
      ✓  0.3s

Product Name      0
Product ID        0
Listing Price     0
Sale Price        0
Discount          0
Brand             0
Description       0
Rating            0
Reviews           0
Last Visited      0
dtype: int64
```

DATA CLEANING

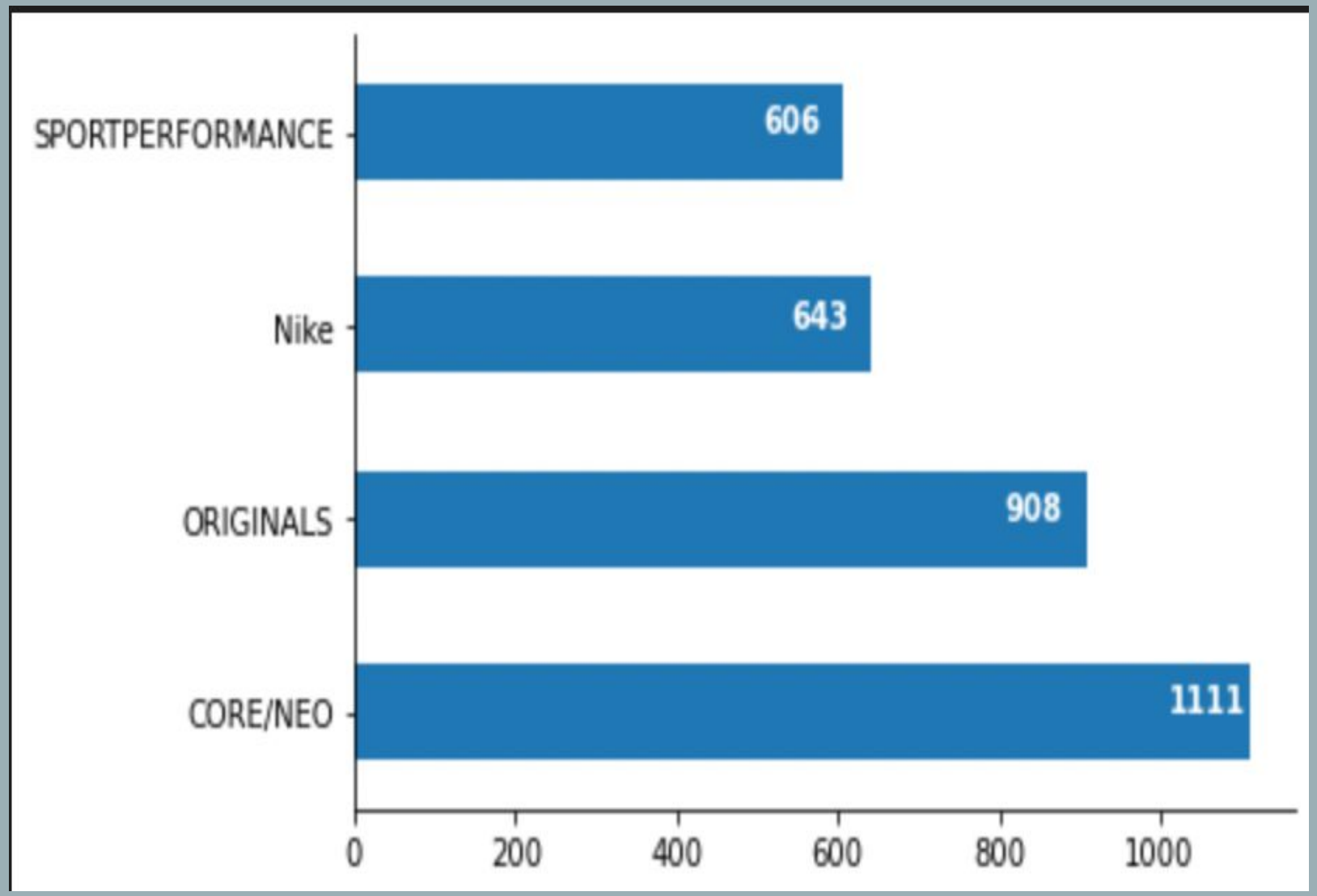
- Remove Redundancy

'Adidas Adidas ORIGINALS' and 'Adidas ORIGINALS' are duplicated, keep one of them

```
df['Brand'] = df['Brand'].str.replace('Adidas','')
df['Brand'] = df['Brand'].str.replace(' ','')

# Merge the brand and description to facilitate subsequent analysis of text word frequency
df['Description'] = df['Description'].astype(str) + ' '+ df['Brand'].astype(str)
df.head()
```

✓ 0.4s

Exploratory Data Analysis

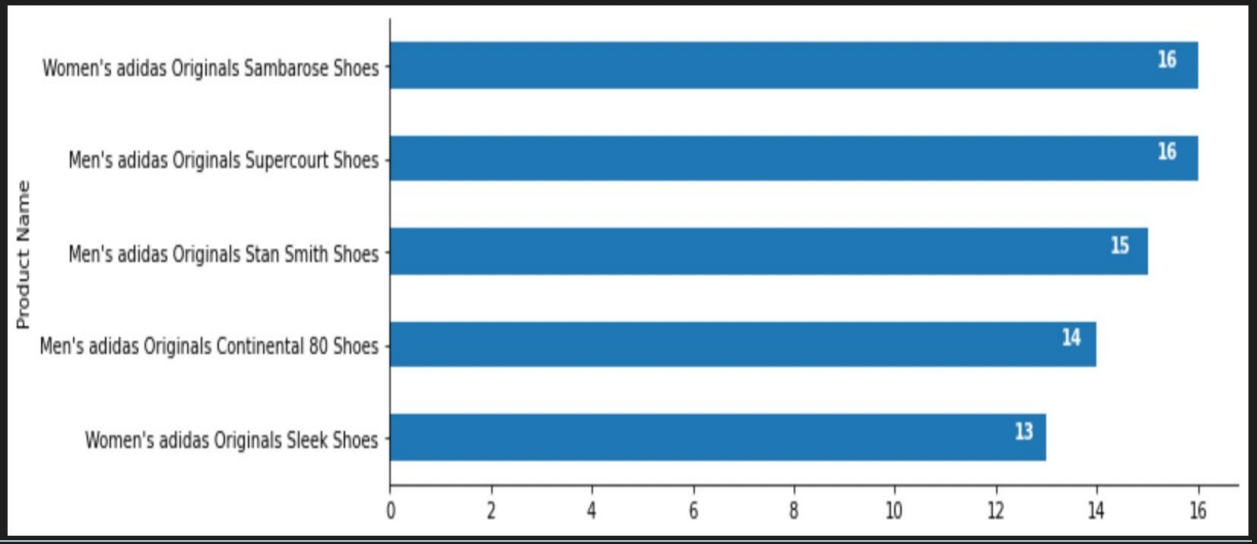


Inference:
Obviously Adidas Core/Neo has the highest sales

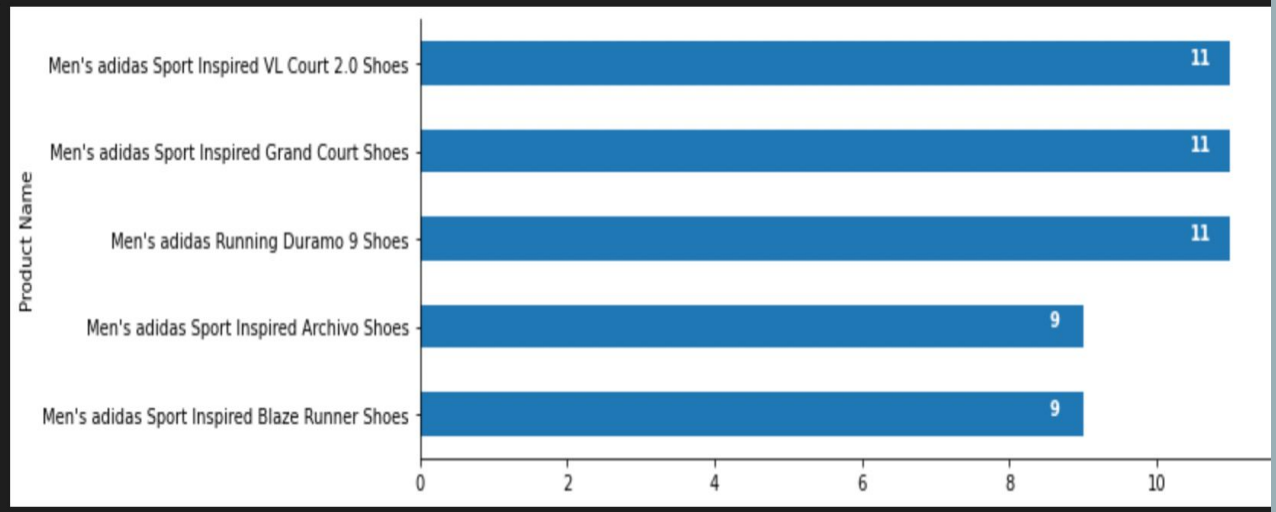
- **Which brand has the highest sales?**

Which are the top selling products of each brand

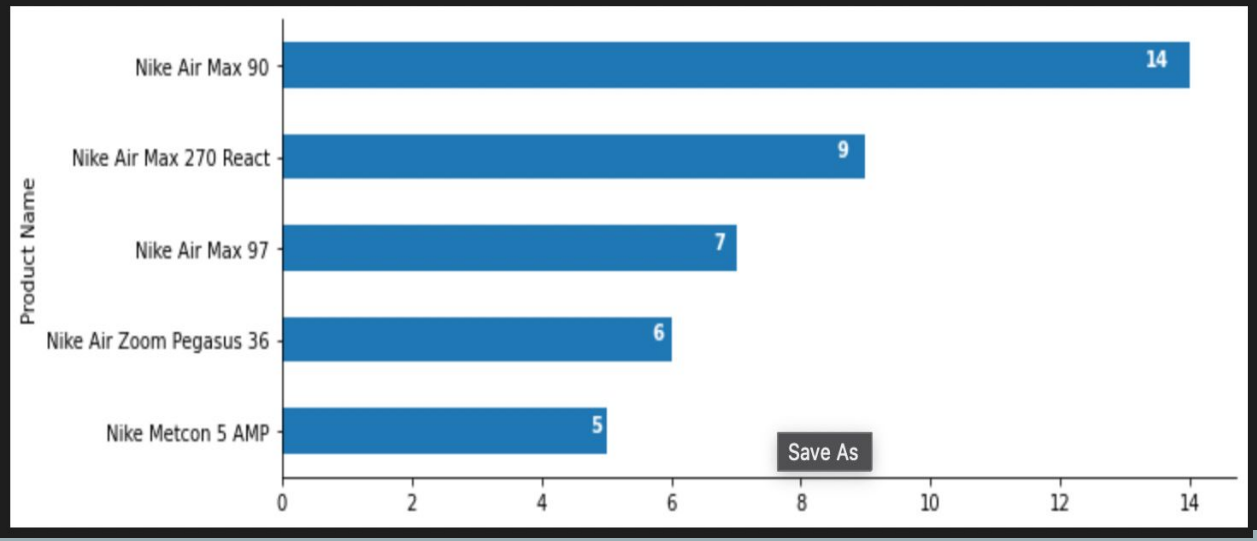
ORIGINALS :



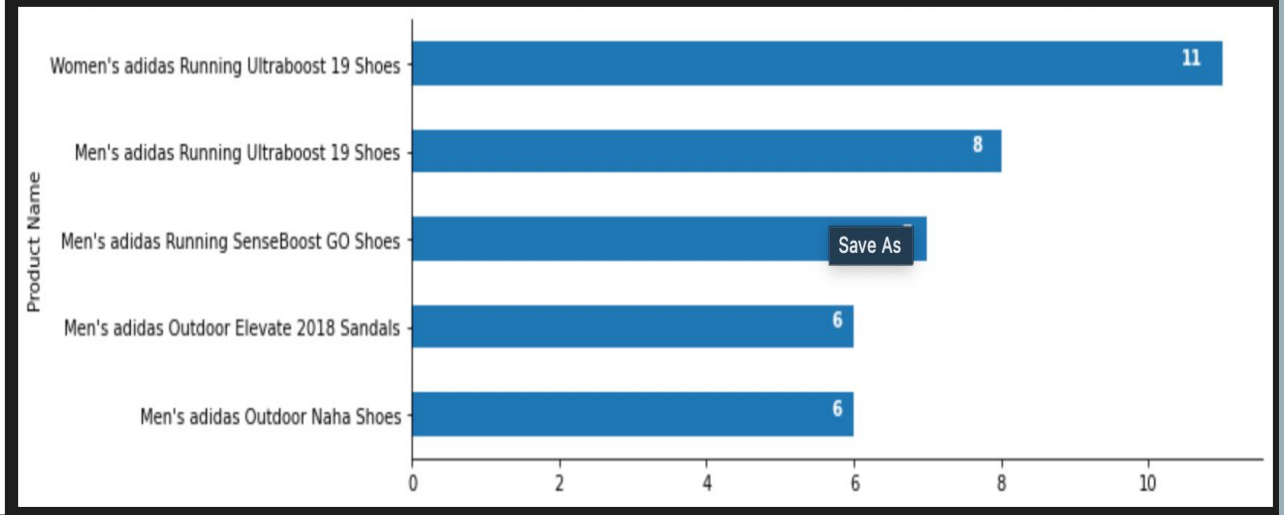
CORE/NEO :



Nike :



SPORTPERFORMANCE :



CLUSTERING

Cluster analysis or clustering is an unsupervised machine learning algorithm that groups unlabeled datasets. It aims to form clusters or groups using the data points in a dataset in such a way that there is high intra-cluster similarity and low inter-cluster similarity.Clustering is used to identify groups of similar objects in datasets with two or more variable quantities.

CLUSTER DATA IN PYTHON

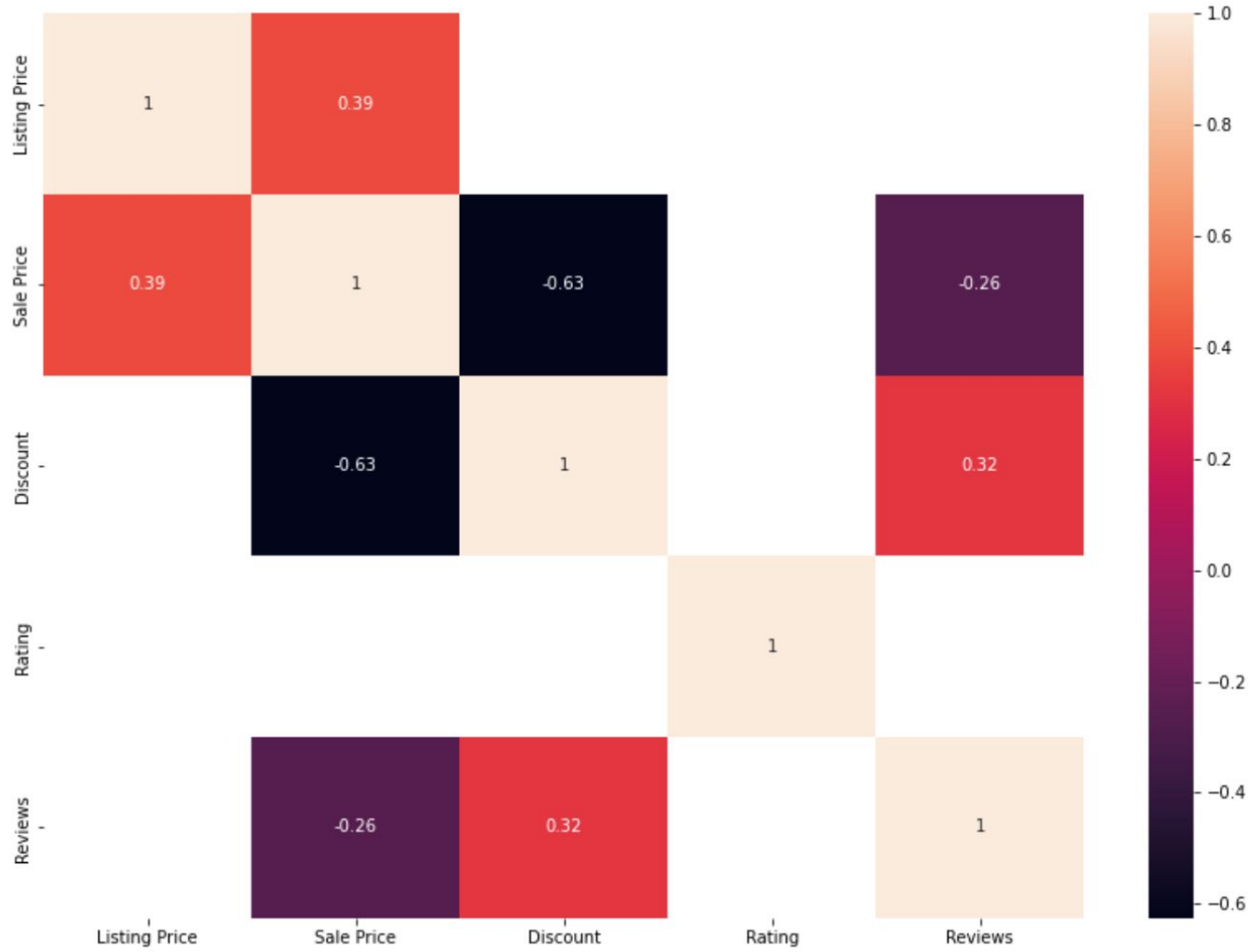
Steps:

1. Choose some values of k and run the clustering algorithm.
2. For each cluster, compute the within-cluster sum-of-squares between the centroid and each data point.
3. Sum up for all clusters, plot on a graph.
4. Repeat for different values of k, keep plotting on the graph.
5. Then pick the elbow of the graph.

CORRELATION COEFFICIENT

```
In [26]: plt.figure(figsize=(14,10))
         sns.heatmap(df.corr(method='spearman'),annot=True,mask=(df.corr()**2<0.04))
```

Out[26]: <AxesSubplot:>



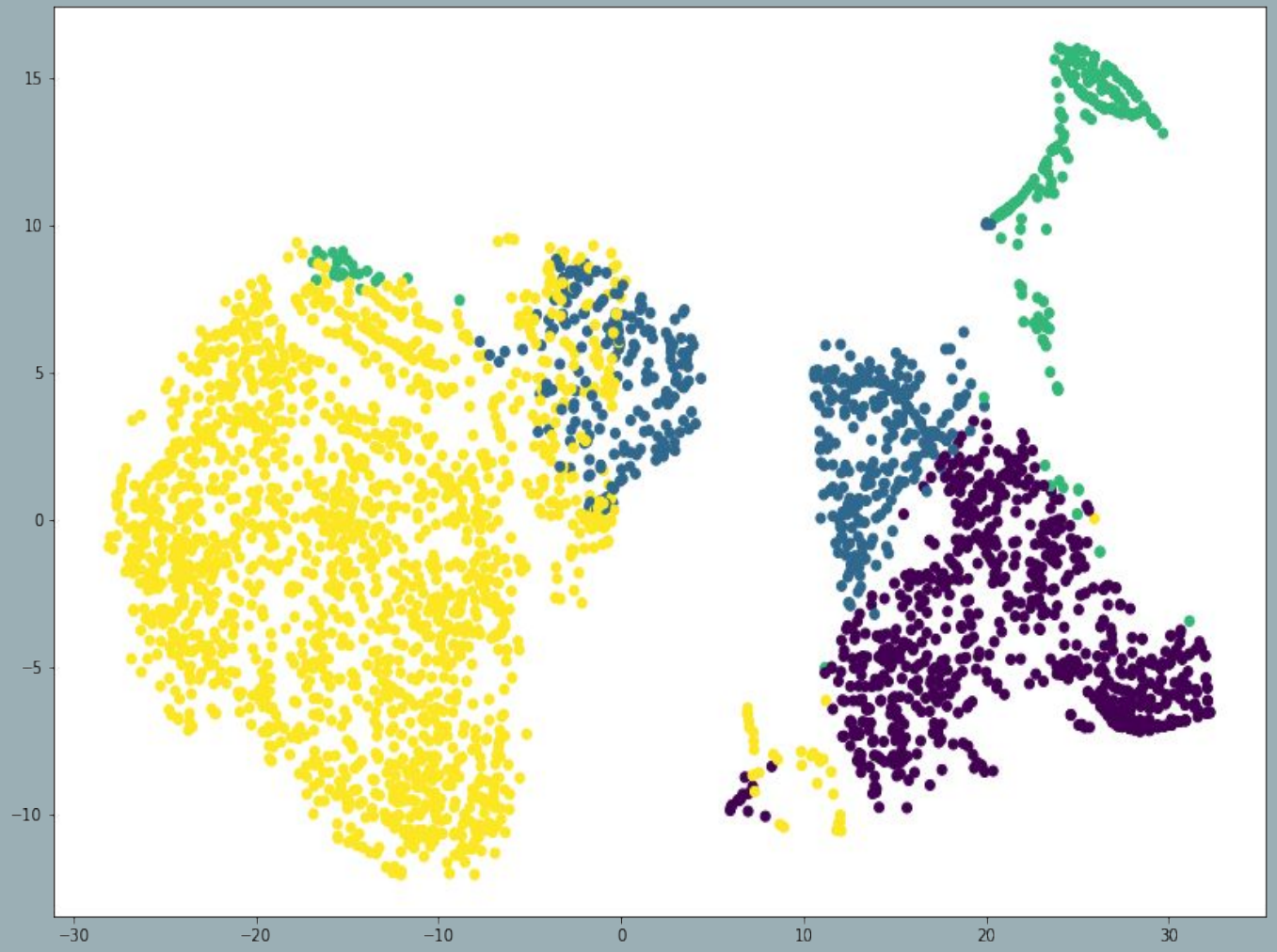
K-MEANS

- We have determined the number of categories by using the elbow method and also the silhouette coefficient.
- By combining the Silhouette coefficient and the elbow graph, the number of clusters that we obtained was 4.

KMeans

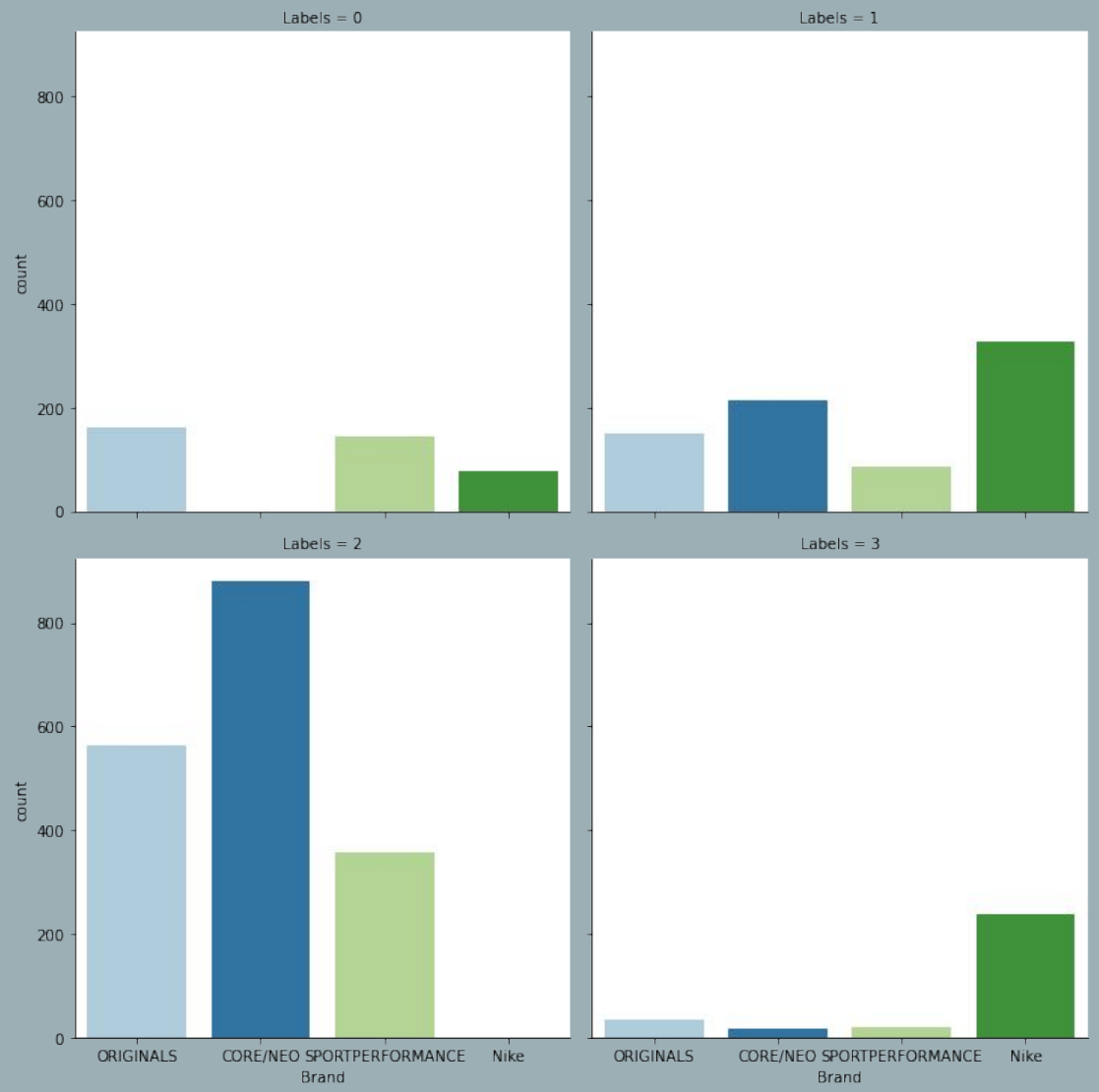
```
num_clusters = range(2,11)
inertias =[]
sil_scores = []
for k in num_clusters:
    model = KMeans(n_clusters=k)
    model.fit(df_4d)
    inertias.append(model.inertia_)
    sil_scores.append(silhouette_score(df_4d,model.labels_))
plt.plot(num_clusters,inertias,'-o')
```

K-MEANS



- Then we used the Tsne tool in python to visualize the the data.
- The data itself was reduced and still had 4 dimensions.
- Then we added the result calculated by K-means to the table.

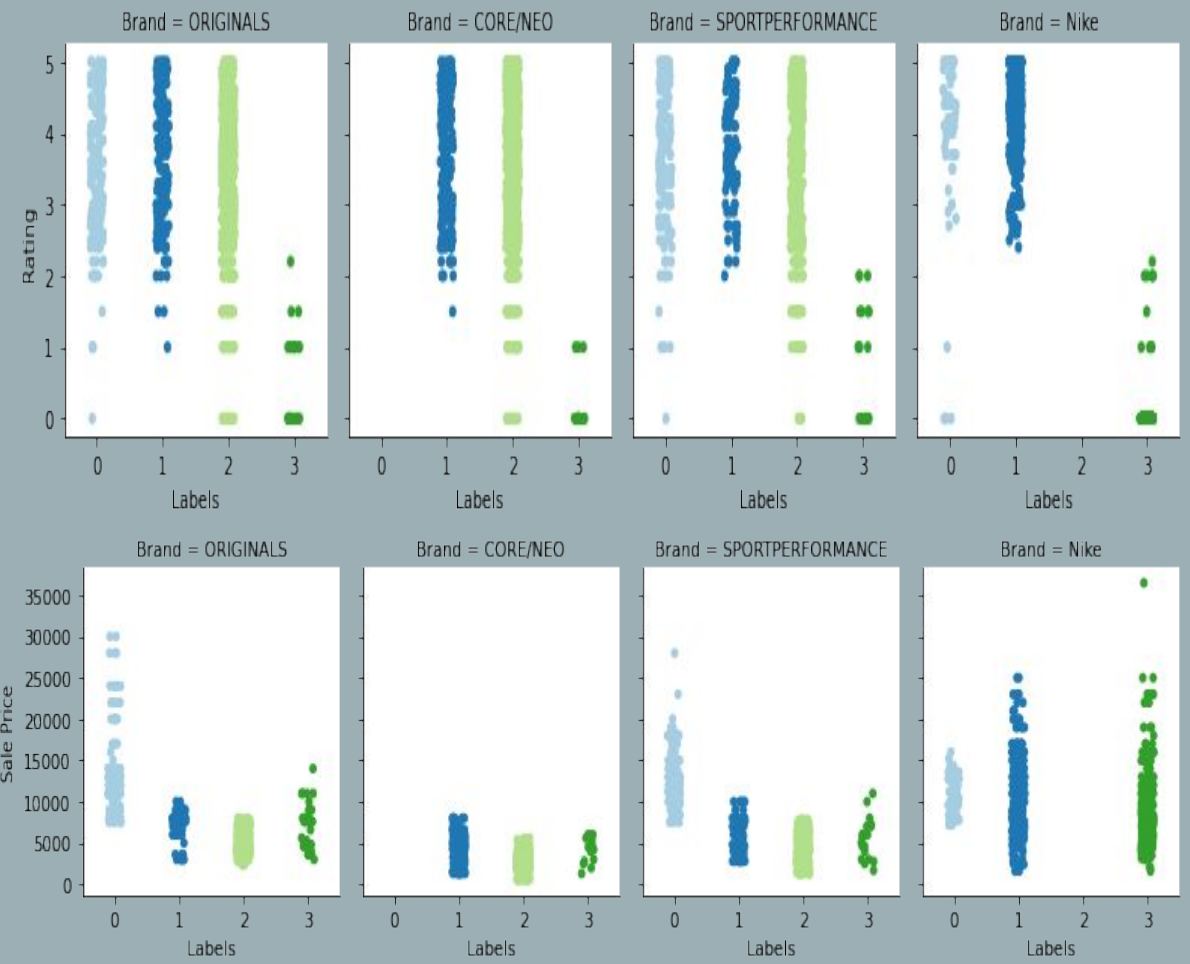
K-MEANS



- Then we used the python seaborn library to to find out which of the brands were included in each category.
- We got a total of 4 graphs according to each of the clusters for all 4 of the brands.
- The results we got were:
  - No Adidas Core/Neo in category 0.
  - The main brand in category 3 is Nike.
  - The number of Adidas (sub-brands) in category 2 is significantly higher than in the other three categories, and there is no Nike in category 3.

K-MEANS

- We then created a scatterplot for each of the features, 'Listing Price','Sale Price','Discount','Rating','Reviews' against each of the labels.
- We analyzed 20 graphs in total (4 for each category), and the results we obtained were:
  - Category 0 has the highest order price and does not include the Adidas Core/Neo brand
  - Category 2 has the largest discount and the lowest discounted price
  - Nike has the highest discount since there is no discount
  - Category 3 has a lower rating
  - Category 0: High-priced products
  - Category 1: Cheap and low discount products
  - 2 categories: cost-effective products (medium price high discount)
  - Category 3: Low Rated Products



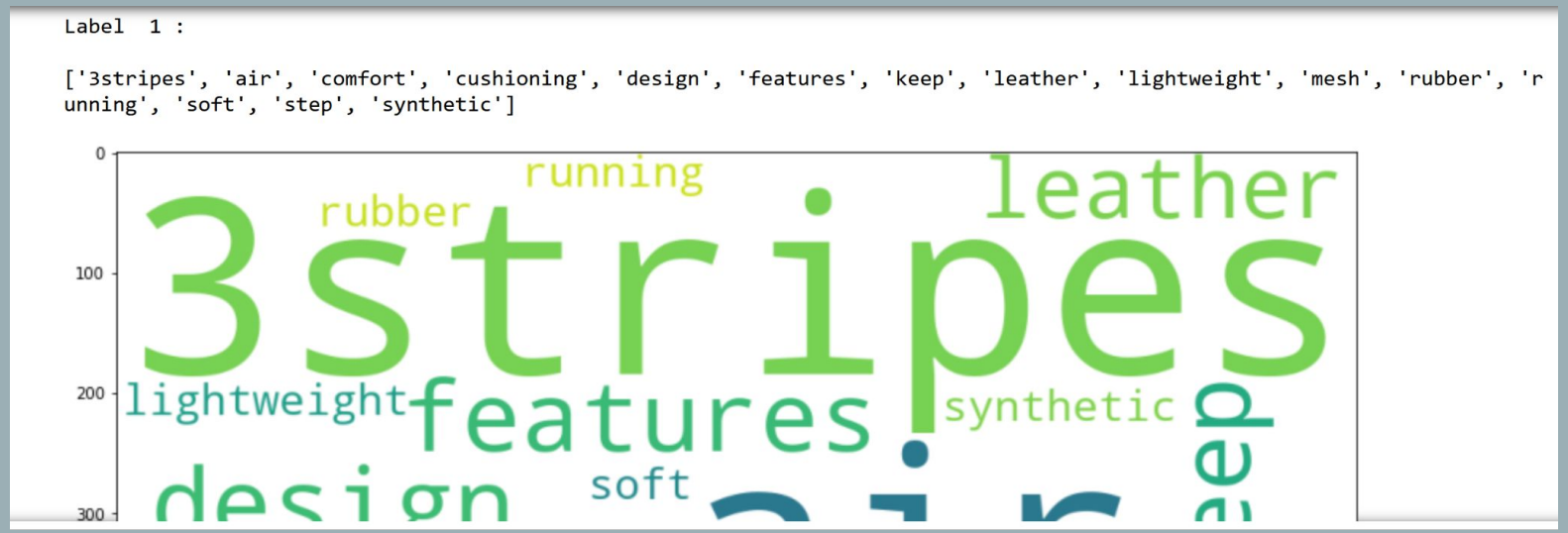
WORD CLOUD AND WORD TOKENIZER

```
from nltk import word_tokenize
import re

# filter unimportant words
stop_words = ['of','an','a','are','is','with','the','adidas','on','in','this','by'
              ,'to','and','as','for','have','has','at','in','its','these','it','you','your'
               ,'that','look','shoe','shoes','outsole','midsole','feel','feet','every','from'
              ,'they','while','upper','style','foot','provides','nike','originals','coreneo'
              ,'sportperformance','comfortable','run','new']

def remove_noise(text,stop_words = stop_words):
    words = word_tokenize(text)
    cleaned_words = []
    for word in words:
        word = re.sub('\W','',word)
        if len(word) > 1 and word.lower() not in stop_words:
            cleaned_words.append(word.lower())
    return cleaned_words
```

```
# Calculate word frequency to get keywords
from sklearn.feature_extraction.text import TfidfVectorizer
tfidf = TfidfVectorizer(max_df=0.9,max_features=15,min_df=0.1,ngram_range=(1,2),tokenizer=remove_noise)
```

- We have used the Word Tokenizer to filter out the unnecessary words such as ‘of, an, a, are’.
- We have also used the TfidfVectorizer to get the frequency of the words, used the above word tokenizer to filter the data, to get the keywords.

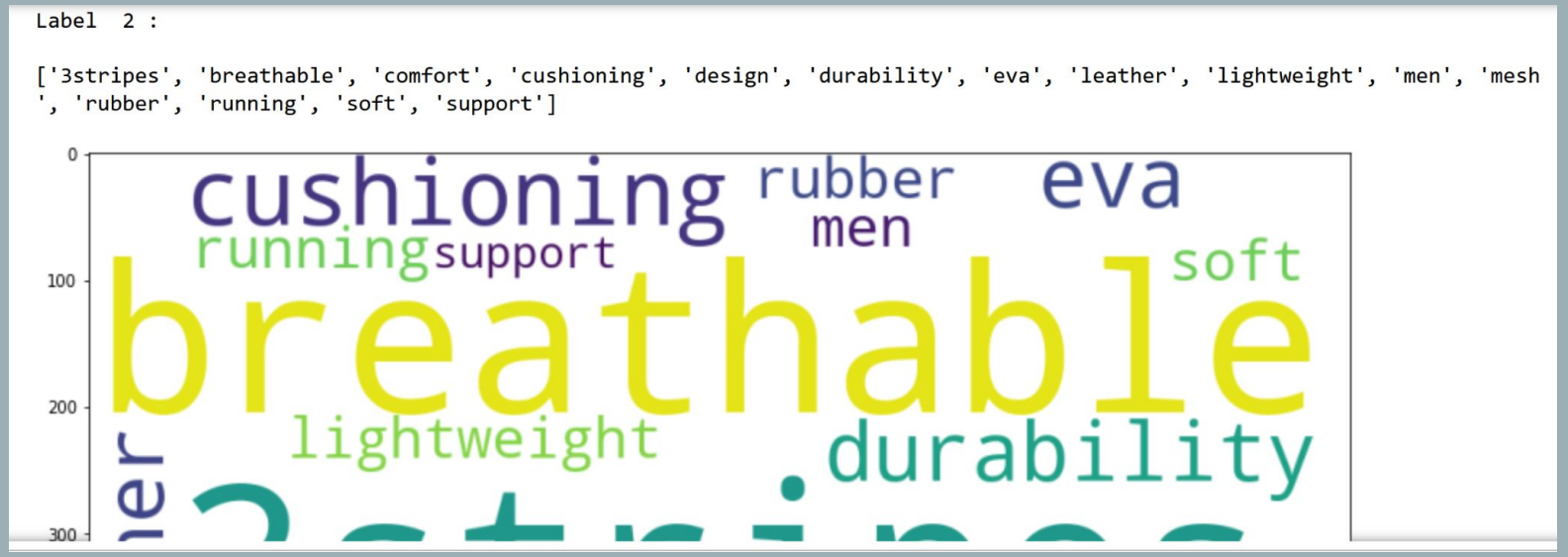
WORD TOKENIZER AND WORDCLOUD

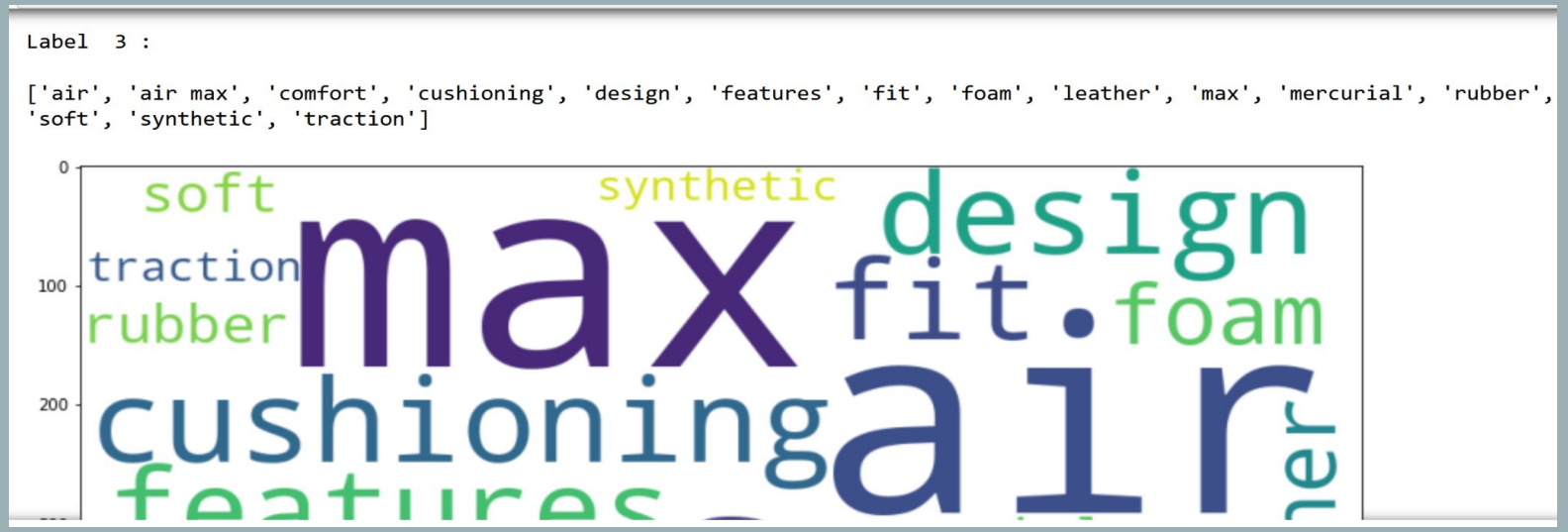


- We have also displayed each of the labels from the frequency count as an individual WordCloud along with the words.
- The reviews in Label 0 were: 'air', 'boost', 'comfort', 'cushioning', 'design', 'energy', 'fit', 'inspired', 'knit', 'lightweight', 'mesh', 'responsive', 'responsive cushioning', 'running', 'support'.
- The reviews in Label 1 were: '3stripes', 'air', 'comfort', 'cushioning', 'design', 'features', 'keep', 'leather', 'lightweight', 'mesh', 'rubber', 'running', 'soft', 'step', 'synthetic'.

WORD TOKENIZER AND WORDCLOUD

- The reviews in Label 2 were: 3stripes', 'breathable', 'comfort', 'cushioning', 'design', 'durability', 'eva', 'leather', 'lightweight', 'men', 'mesh', 'rubber', 'running', 'soft', 'support'.
- The reviews in Label 3 were: 'air', 'air max', 'comfort', 'cushioning', 'design', 'features', 'fit', 'foam', 'leather', 'max', 'mercurial', 'rubber', 'soft', 'synthetic', 'traction'.





Inference

- **When comparing the number of products, Adidas has a lot more that Nike. Additionally, Adidas is categorized in 3 brands: Adidas CORE/NEO, Adidas SPORT PERFORMANCE and Adidas ORIGINALS, while Nike just has the one brand.**
- **From the wrod_cloud we can infer that high price products which belong to label 0(1st Cluster) are reviewed as better was “cushioning” and “responsive” whereas label 1(cluster 2) for “stripes” and “comfort” ; label 2(cluster 3) for “breathable” and “stripes” and label 3(cluster 4) for “air max” and “comfort”.**
- **Furthermore, overall ratings and reviews are higher for Adidas than Nike with this dataset. All these are indicative of Adidas's huge popularity with particular pieces while Nike has a nicer overall range of products.**

THANK YOU