Automated marketing research using online user reviews

Contents :

Problem statement :

Selecting product attributes for market structure analysis

Problem Description

Market Structure Analysis

- Central to marketing
- Key step in
  - Design and development of new products
  - Repositioning of existing products
- Describes substitution and complementary relationship between brands
- Predicts market responses to:
  - Changes in pricing
  - Market strategy
  - Product introduction

Traditional Approach to Market Structure Analysis

- Uses Surveys
- Uses the thought:
    - “All customers perceive all products the same way with difference in attribute evaluation only”
- Little research on how to choose product attributes i.e. keywords
- Voice of Customer not being used for choosing keywords for marketing

Our Approach

Our approach facilitates Market Structure Analysis in 2 ways:

- Selecting attributes based on Voice of Customer
    - Selecting product attributes for marketing on the basis of what customers are concerned about

- Augmenting Traditional approaches by providing input

Approach

- Data Collection:
  - Web Page scraping to get user reviews
- Clustering:
  - Term-Document Matrix
  - Clustering of terms based on cosine similarity
  - Using k-means
- Correspondence Analysis

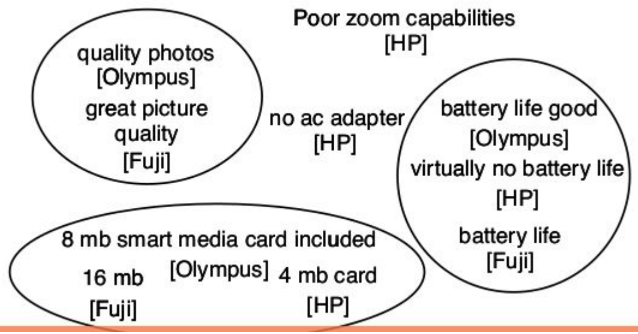
Methodology

Screen Scraping

Olympus: Quality of Photos, …, Battery life (very very good), Only 8 mb Smart media card
HP: …, Only a 4 mb card, virtually no battery life, no AC adapter, Poor zoom capabil
Fuji: Great picture quality, 16 mb, battery life, …

Matrix of Word Vectors

Brand	Original phrase	Stop-words removed	only	life	mb	card	zoom
Olympus	Quality of Photos	quality photos					
Olympus	Battery life (very very good)	battery life good		1			
Olympus	Only 8 mb Smart media card ...	8 mb smart media card inc	1		1	1	
HP	Only a 4MB card	4 mb card			1	1	
HP	Virtually no battery life	virtually battery life		1			
HP	No AC adapter	no ac adapter					
HP	Poor zoom capabilities	zoom capabilities					1
Fuji	Great picture quality	great picture quality					
Fuji	16 mb	16 mb			1		
Fuji	battery life	battery life		1			

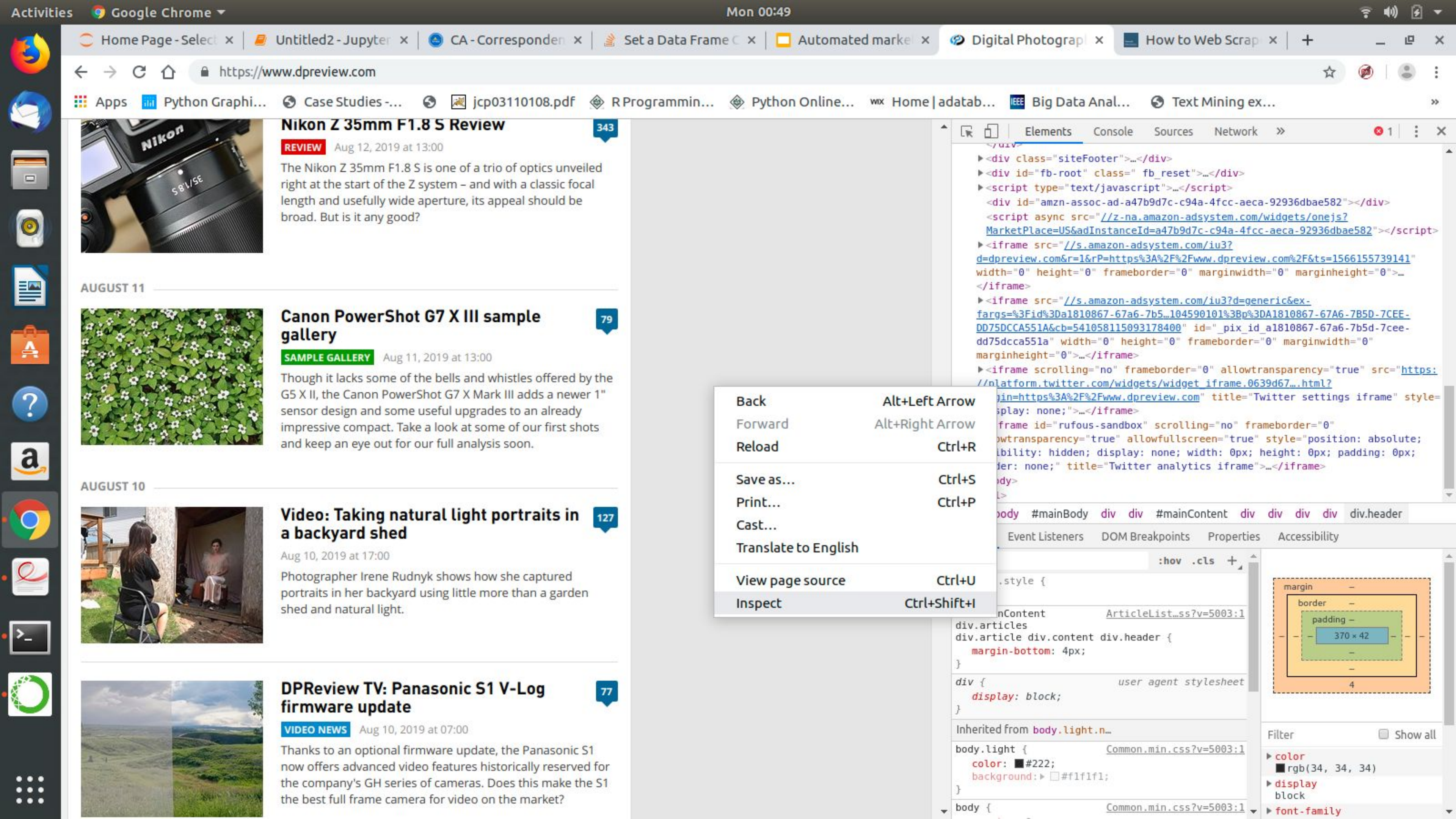


Data Collection :

We collected the online user reviews for digital cameras

Scraping

Web scraping is a technique to automatically access and extract large amounts of information from a website, which can save a huge amount of time and effort.


Activities
Google Chrome
Mon 00:49
Home Page - Select
Untitled2 - Jupyter
CA - Corresponden
Set a Data Frame C
Automated marke
Digital Photograp
How to Web Scrap
https://www.dpreview.com
Apps
Python Graphi...
Case Studies -...
jcp03110108.pdf
R Programmin...
Python Online...
Home | adatab...
Big Data Anal...
Text Mining ex...
Nikon Z 35mm F1.8 S Review
343
REVIEW
Aug 12, 2019 at 13:00
The Nikon Z 35mm F1.8 S is one of a trio of optics unveiled right at the start of the Z system – and with a classic focal length and usefully wide aperture, its appeal should be broad. But is it any good?
AUGUST 11
Canon PowerShot G7 X III sample gallery
79
SAMPLE GALLERY
Aug 11, 2019 at 13:00
Though it lacks some of the bells and whistles offered by the G5 X II, the Canon PowerShot G7 X Mark III adds a newer 1" sensor design and some useful upgrades to an already impressive compact. Take a look at some of our first shots and keep an eye out for our full analysis soon.
AUGUST 10
Video: Taking natural light portraits in a backyard shed
127
Aug 10, 2019 at 17:00
Photographer Irene Rudnyk shows how she captured portraits in her backyard using little more than a garden shed and natural light.
DPReview TV: Panasonic S1 V-Log firmware update
77
VIDEO NEWS
Aug 10, 2019 at 07:00
Thanks to an optional firmware update, the Panasonic S1 now offers advanced video features historically reserved for the company's GH series of cameras. Does this make the S1 the best full frame camera for video on the market?
Back Alt+Left Arrow
Forward Alt+Right Arrow
Reload Ctrl+R
Save as... Ctrl+S
Print... Ctrl+P
Cast...
Translate to English
View page source Ctrl+U
Inspect Ctrl+Shift+I
Elements
Console
Sources
Network
<div class="siteFooter">…</div>
<div id="fb-root" class=" fb_reset">…</div>
<script type="text/javascript">…</script>
<div id="amzn-assoc-ad-a47b9d7c-c94a-4fcc-aeca-92936dbae582"></div>
<script async src="//z-na.amazon-adsystem.com/widgets/onejs?MarketPlace=US&adInstanceId=a47b9d7c-c94a-4fcc-aeca-92936dbae582"></script>
body #mainBody div div #mainContent div div div div div.header
Event Listeners
DOM Breakpoints
Properties
Accessibility
:hov .cls
div.articles
div.article div.content div.header {
margin-bottom: 4px;
}
div {
user agent stylesheet
display: block;
}
Inherited from body.light.n…
body.light {
Common.min.css?v=5003:1
color: #222;
background: #f1f1f1;
}
margin
border
padding
370 × 42
Filter
Show all
color
rgb(34, 34, 34)
display
block
font-family

Beautiful soup :

- It is a python library for pulling out data from Html and xml files
- Beautiful soup parses the document using the best available parser . (we have used html parser).
- Beautiful Soup transforms a complex HTML document into a complex tree of Python objects.

Identifying the useful links :

- Fetch all tags <a href='...' >

- Use regular expressions to extract links of products

Examples :

- https://www.dpreview.com/samples/2514555088/canon-rf-24-240mm-f4-6-3-is-sample-gallery

- https://www.dpreview.com/articles/5022781382/is-the-panasonic-lumix-dc-s1r-right-for-you

Extract reviews :

- Iterate over all the links we got .
- Find all elements <div class = ‘message >
- Iterate over all div tags and fetch <p>...<p>

Data Preprocessing

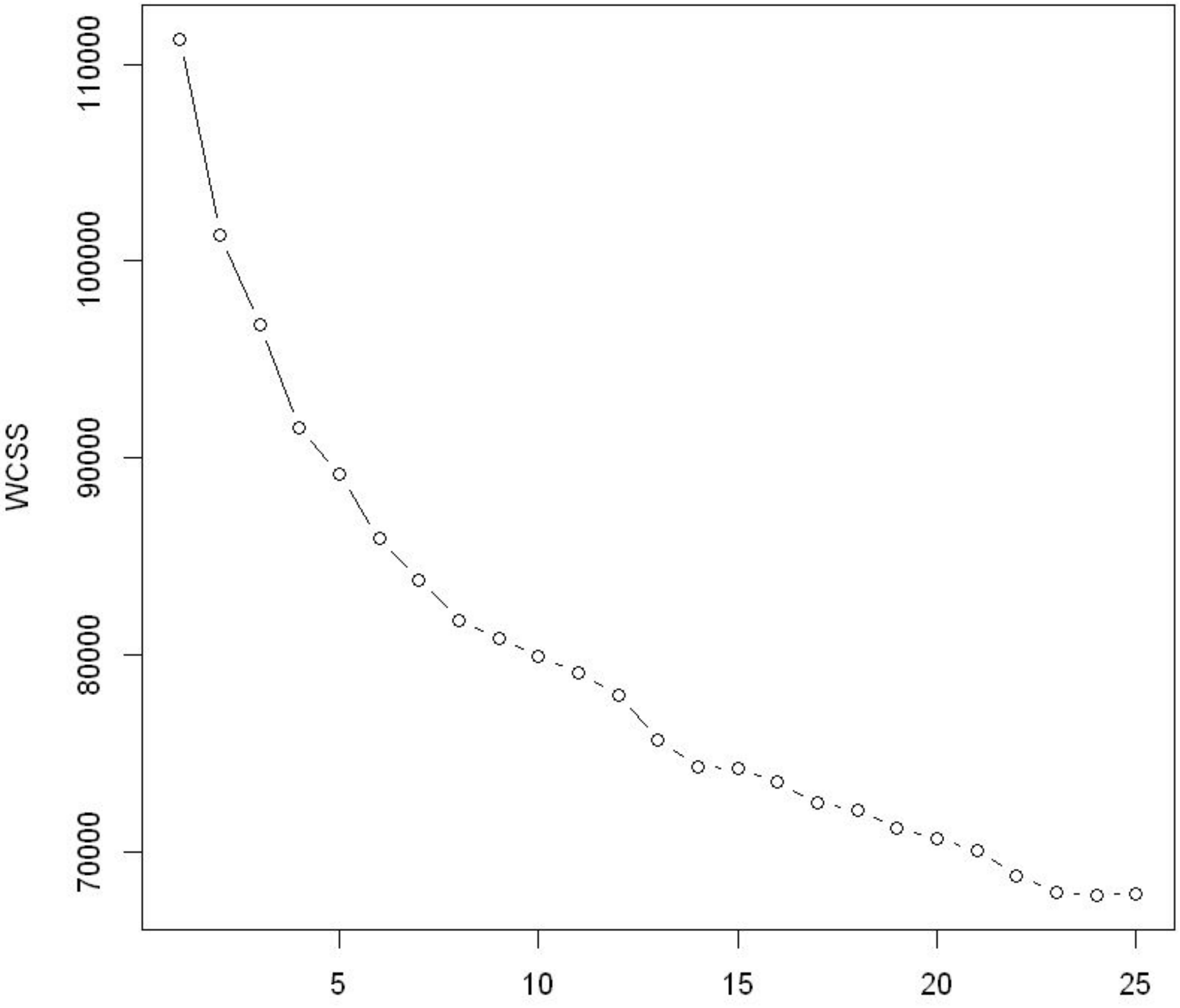
1. Convert in lowercase
2. Remove stopword
3. Remove punctuation
4. Remove url
5. Stemming

K-means Clustering

- ❖ It is partition based algorithm. It is most popular algorithms for text mining.
- ❖ It is efficient on the large data.
- ❖ It work on the numerical data.

Elbow Method

It is used to choose optimal no of cluster.This method cannot give you the optimal number of clusters as an exact point, it can give you an optimal range .



Some Clusters :

Size ,
mb,
speed,etc
.

Memory

Zoom,
focus,
lag,delay,
etc.

Lens

Heat,low
,
good,lon
g life,

Battery

Well, time
,light,gre
at

Brightness

better,ey
e
,get,resol
ution

Display

size,weig
ht,light,he
avy,etc.

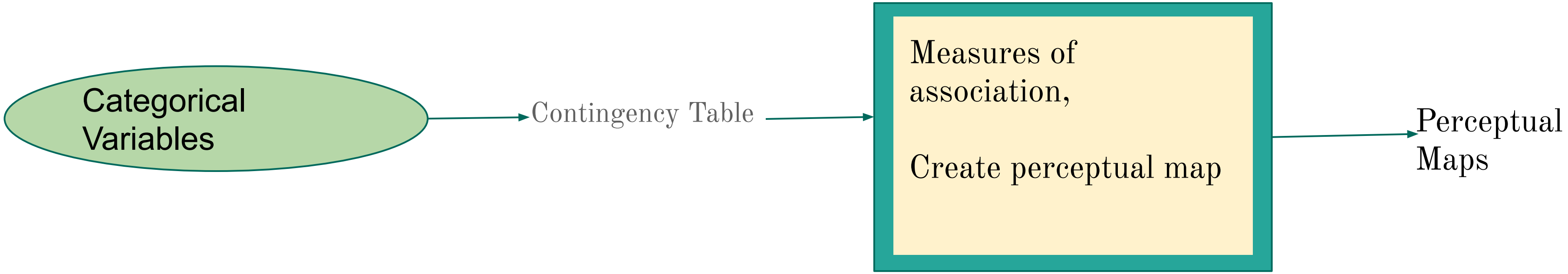
Design

Correspondence Analysis

Correspondence Analysis

- Multivariate statistical technique
- Geometric approach to categorical data analysis
- Deals with categorical data
- Perceptual maps are plotted for extracted components

Correspondence Analysis process



Contingency table

	Sony	nikon	casio	fuji	canon	kodak	olympus	panasonic
Battery	31	12	2	61	22	2	71	24
Memory	22	14	1	1	3	1	5	1
Size	31	28	4	0	4	4	1	4
Control	4	9	0	11	26	3	20	0
Zoom	58	61	3	7	1	2	8	3
Lens	43	15	2	31	6	62	4	2
Focus	32	6	5	4	2	5	12	5
Flash	6	72	0	10	8	1	0	0
Disk	7	5	3	3	5	0	41	37
Video	82	3	2	12	10	6	1	2
Brightness	8	16	7	7	2	3	7	78
Viewfind	1	2	1	3	0	1	3	3

Tasks:

- Relationship between Attributes
- Relationship between Brands
- Relationship between Attribute and Brands
- Representing these relationships in a low dimensional space

Table

	BD	D	DM	FI	HM	P	S	W	Total
A1	150	137	207	91	76	210	185	20	**1076**
A2	142	139	200	120	105	221	185	29	**1141**
A3	146	130	193	114	87	205	148	20	**1043**
A4	57	68	269	260	87	159	239	42	**1181**
Total	**495**	**474**	**869**	**585**	**355**	**795**	**757**	**111**	**4441**

Correspondence Matrix

	BD	D	DM	FI	HM	P	S	W	Row mass
A1	0.034	0.031	0.047	0.020	0.017	0.047	0.042	0.005	**0.242**
A2	0.032	0.031	0.045	0.027	0.024	0.050	0.042	0.007	**0.257**
A3	0.033	0.029	0.043	0.026	0.020	0.046	0.033	0.005	**0.235**
A4	0.013	0.015	0.061	0.059	0.020	0.036	0.054	0.009	**0.266**
Col. Mass	**0.111**	**0.107**	**0.196**	**0.132**	**0.080**	**0.179**	**0.170**	**0.025**	**1.000**

$$z_{ij} = x_{ij}/N$$

Row Profiles

	BD	D	DM	FI	HM	P	S	W	Row mass
A1	0.139	0.127	0.192	0.085	0.071	0.195	0.172	0.019	**0.242**
A2	0.124	0.122	0.175	0.105	0.092	0.194	0.162	0.025	**0.257**
A3	0.140	0.125	0.185	0.109	0.083	0.197	0.142	0.019	**0.235**
A4	0.048	0.058	0.228	0.220	0.074	0.135	0.202	0.036	**0.266**
Col. Mass	**0.111**	**0.107**	**0.196**	**0.132**	**0.080**	**0.179**	**0.170**	**0.025**	**1.000**

$z_{ij} = z_{ij}/\text{Rowmass}[i]$

Row Profile

1. $z_{ij} = x_{ij}/N$

   Z = Correspondence Matrix

2. $z_{ij} = z_{ij}/\text{Rowmass}[i]$; where Rowmass[i]=Rowsum[i]/N

   N= Total sum

   X_{ij}= ij^{th} element in contingency table

   Resulting matrix can be used to find similarity/ dissimilarity between attriibutes

Correspondence Matrix

	BD	D	DM	FI	HM	P	S	W	Row mass
A1	0.034	0.031	0.047	0.020	0.017	0.047	0.042	0.005	**0.242**
A2	0.032	0.031	0.045	0.027	0.024	0.050	0.042	0.007	**0.257**
A3	0.033	0.029	0.043	0.026	0.020	0.046	0.033	0.005	**0.235**
A4	0.013	0.015	0.061	0.059	0.020	0.036	0.054	0.009	**0.266**
Col. Mass	**0.111**	**0.107**	**0.196**	**0.132**	**0.080**	**0.179**	**0.170**	**0.025**	**1.000**

$$z_{ij} = x_{ij}/N$$

Column Profiles

	BD	D	DM	FI	HM	P	S	W	Row mass
A1	0.303	0.289	0.238	0.156	0.214	0.264	0.244	0.180	**0.242**
A2	0.287	0.293	0.230	0.205	0.296	0.278	0.244	0.261	**0.257**
A3	0.295	0.274	0.222	0.195	0.245	0.258	0.196	0.180	**0.235**
A4	0.115	0.143	0.310	0.444	0.245	0.200	0.316	0.378	**0.266**
Col. Mass	**0.111**	**0.107**	**0.196**	**0.132**	**0.080**	**0.179**	**0.170**	**0.025**	**1.000**

$z_{ij} = z_{ij}/\text{Column\_mass}[j]$

Column Profile

- Relationship between Brands

➢ Column Profile

1. $z_{ij} = x_{ij}/N$
2. $z_{ij} = z_{ij}/ColumnMass[j]$; where $ColumnMass[j] = Columnsum[j]/N$

$N =$ Total sum

$x_{ij} = ij^{th}$ element in contingency table

Resulting matrix can be used to find similarity/ dissimilarity between Brands

Relationship between attribute and brand: Weighted χ^2 Distance

$D = (D_r^{-1})^{½}(Z - rc^T)(D_c^{-1})^{½}$

Z = m x n Correspondence matrix

r = m x1 Rowmass vector

c= n x 1 Columnmass vector

D_r = m x m diag(r) matrix

D_c = n x n diag(c) matrix

Interpretation

- The vectors **r** and **c** give the marginal probabilities of being the row and column classes, respectively, while **Z** gives the joint probability distribution of rows and columns.
- $\mathbf{Z}\text{-}\mathbf{rc}^T$ gives deviation from independence.
- **D**: chi-squared statistic, yielded from summing the deviations, squared and appropriately scaled.

If independent=> Z-rc = 0

If there is some non-zero distance=> attribute and brands are not independent

Reducing Dimensions : SVD

$D = U\Sigma V^T$: Doing SVD of chi-sq distance matrix

Find U

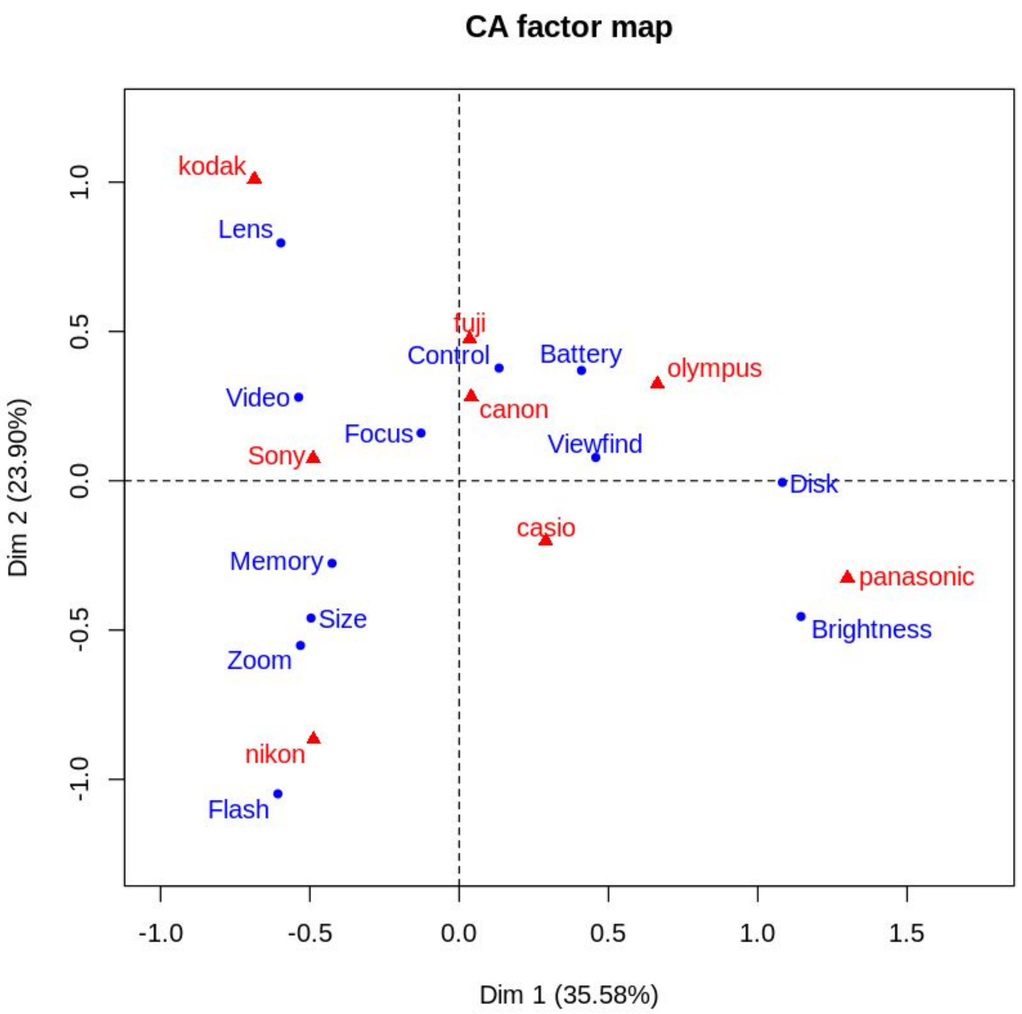
Obtain row(Attribute) PCs: $P = (D_r^{-1})^{½}UD$

Obtain column(Brand) PCs : $Q = (D_c^{-1})^{½}VD$

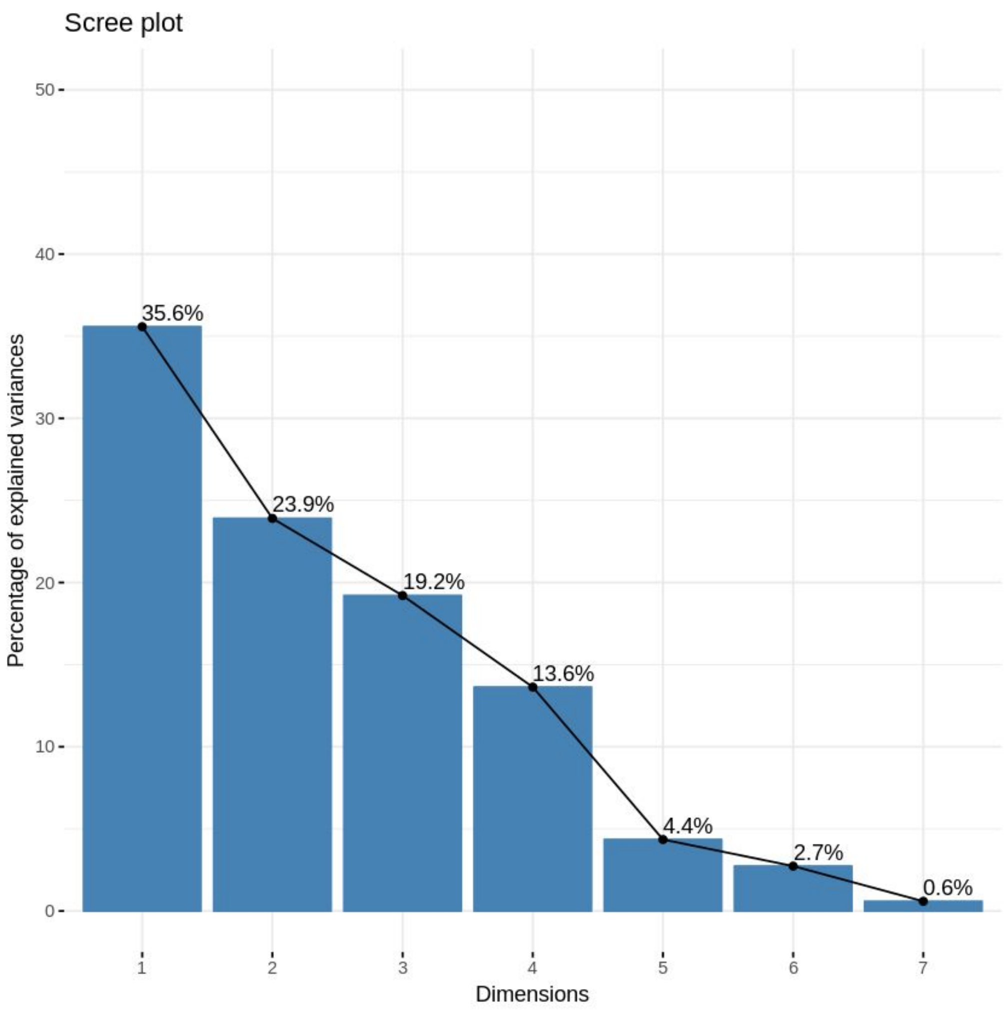
	PC1	PC2..
A1		
A2		
..		
B1		
B2..		

Perceptual Map

Implemented using FactoMineR and factoextra in R



Scree plot


Scree plot
50
40
30
20
10
0
Percentage of explained variances
35.6%
23.9%
19.2%
13.6%
4.4%
2.7%
0.6%
1
2
3
4
5
6
7
Dimensions

References

- Automated Marketing Research Using Online Customer Reviews THOMAS Y. LEE and ERIC T. BRADLOW
- https://www.crummy.com/software/BeautifulSoup/bs4/doc/
- https://en.wikipedia.org/wiki/Correspondence_analysis