CIS 8392 – ADVANCED TOPICS IN BIG DATA ANALYTICS

Examining Market Structure using Perceptual Maps

Submitted By

Sravani Kothi

Abstract

Consumer perceptions are important components of brand equity and therefore marketing strategy. A central analytical tool used to understand consumer perceptions is perceptual mapping, which organizes brands according to how consumers rate them with respect to attributes. Understanding how consumers perceive different brands within a competitive set is integral to many marketing goals. The aim of this paper is to provide clear guidelines for producing these maps so that they are indeed useful and simple aids for the reader. To facilitate this, we present how the data is sourced for the creation of perceptual maps, how the data is stored and the process of data analytics involved in obtaining perceptual map for the given dataset. The focus has been kept on Principal Component Analysis to construct the map. The map will clearly show how different attributes are mapped in the minds of the customers and how the same information can be helpful for managers to take marketing decisions.

Introduction:

Perceptual maps are often used in marketing to visually study relations between two or more attributes. Perceptual mapping is a procedure for determining the perceived relative images of a set of objects, such as products or brands. It turns consumer appraisals of similarity or preference into distances represented in perceptual space. It is an important tool which the marketing managers or the product developers can use before taking up any major decisions in advertisements, brand positioning or even in case of new product development. The same tool can be effectively used for even market segmentation to identify the specific target audience or customers.

By and large, marketers have relied upon surveys or choice tasks administered to a sample of customers to measure brand perceptions. However, complete reliance on acquiring consumer responses hinders measurement capabilities: surveys are costly to administer; respondent pools are often sparse and deplete quickly, particularly for certain demographics; participants may be unable or unwilling to reveal their true beliefs; participant attention may wane in the face of too many questions; and results may become outdated quickly, particularly if there is a shock or campaign to shift brand image. The explosion of social media in the past decade has raised hope among many marketers that the 'Big Data' trail on platforms such as Twitter, Facebook could be mined to uncover richer and more scalable insights about consumer behavior and perceptions. In this paper we are going to look into two data sources Twitter data and Survey data.

Survey Data:

There are four modes of survey data collection that are commonly used.

1. **Face-to-face surveys:** Suitable for locations where telephone or mail are not developed. Potential for interviewer bias. Easy to manipulate by completing multiple times to skew results.
2. **Self-administered computer surveys (typically online):** Online (Internet) surveys are becoming an essential research tool for a variety of research fields, including marketing, social and official statistics research. According to ESOMAR online survey research accounted for 20% of global data-collection expenditure in 2006. They offer capabilities beyond those available for any other type of self-administered questionnaire.
3. **Mobile surveys:** Mobile data collection or mobile surveys is an increasingly popular method of data collection. Over 50% of surveys today are opened on mobile devices. The survey, form, app or collection tool is on a mobile device such as a smart phone or a tablet. These devices offer innovative ways to gather data and eliminate the laborious "data entry" (of paper form data into a computer), which delays data analysis and understanding. By eliminating paper, mobile data collection can also dramatically reduce costs.
4. **Mixed-mode surveys:** Researchers can combine several above methods for the data collection. For example, researchers can invite shoppers at malls, and send willing participants questionnaires by emails. With the introduction of computers to the survey process, survey mode now includes combinations of different approaches or mixed-mode designs.

Social Media:

The recent proliferation of social media use by both marketers and consumers offers a promising data source to understand consumer perceptions. To many marketers, ‘mining social media data’ is synonymous with ‘mining user generated text’. Indeed, there is a lot of value to be gained from looking at what consumers are writing about brands in online spaces. However, there are limitations to relying on user text. On many social media platforms, fewer than half the users write their own content; fewer still write about the brands to be monitored; and even fewer write about brands in conjunction with topics or attributes of interest. Yet, every user who connects with a brand via an online platform (whether by following the brand on Twitter, liking the brand on Facebook, or even by liking or sharing a brand’s post) provides information by their voluntary ‘mere virtual presence’ in that online brand community. While liking or following a brand is not always indicative of affinity for the brand, it appears to be the case most of the time. Furthermore, each member of a brand’s online community is likely to be a part of many other online communities (i.e. to follow other accounts). By tracing network relationships to learn more about who a brand’s fans are — what they value and are interested in — one can gain insights about brand image that remain invisible in user-generated text. Tweepy is an easy-to-use Python library for implementing Twitter API to extract twitter data.

Pre-processing of the datasets:

Social Media:

A tweet contains a lot of opinions about the data which are expressed in different ways by different users. The raw data having polarity is highly susceptible to inconsistency and redundancy. Preprocessing of tweet include following points

- Remove all URLs (e.g. www.xyz.com), hash tags (e.g. #topic), targets (@username)
- Correct the spellings; sequence of repeated characters is to be handled
- Replace all the emoticons with their sentiment.
- Remove all punctuations, symbols, numbers
- Remove Stop Words
- Expand Acronyms (use of an acronym dictionary)
- Remove Non-English Tweets.

Survey data:

- Inconsistent responses: Some questions in a study tap into a similar concept but are phrased with a positive or negative tone.
- Missing data: When not, all questions are required and participants neglect to answer many of them, this non-response is a symptom of poor quality response. Just as concerning with non-response though is how your data might be biased if participants are systematically not responding to some questions. We should consider examining whether the missing data is random or more systematically biased.
- Pattern detection: Participants that respond using conspicuous patterns such as straight lining (all 5's or all 3's) or alternating from 5's to 1's to rating scales also indicate a bot or a disingenuous respondent. But if participants had a good experience on a website, it's not unsurprising for them to rate the experience as exceptional on say 8 or 10 items. There's more concern if we see straight lining or patterns on 20 or 30 questions in a row.
- Session recordings: If the study is task-based with screen-recordings, then we can observe what the participants are doing while completing the study.
- Disqualifying questions: For many studies we look for participants with particular characteristics/criteria. If a participant somehow is admitted into a survey by answering a screening question a certain way, but then reveals in the open-ended answers they are not qualified, they are excluded (e.g., a participant needs to have a particular credit card, but in an open-ended question reveals that he or she doesn't have any credit cards). This can also be done in combination with a cheater question if, for example, participants state they are familiar with fictitious brands or have bought products that don't exist.

Data Storage:

Twitter (www.twitter.com) is an online social networking and micro blogging platform that enables a user to send and read short character text messages, called "tweets". It is ranked

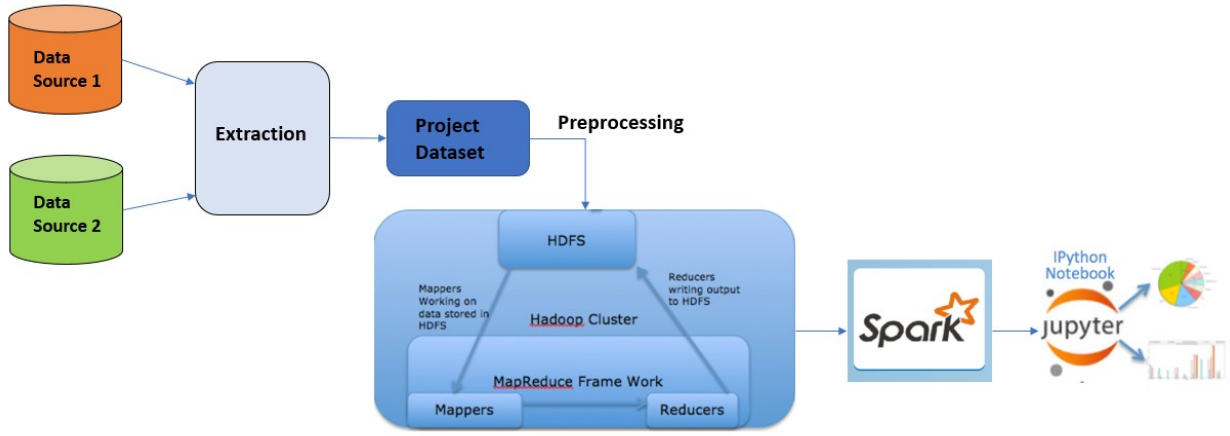
amongst the top 10 most visited websites by Alexa's web traffic analysis for the last 2 years. It has ~250+ million monthly active users who send about ~500+ million tweets per day (ref. Wikipedia). 78% of the twitter users access it on a mobile device and 77% are living outside US. It supports 35+ languages. Its vast global reach and usage has made Twitter into a remarkable platform for analyzing and understanding trends and events shaping the world either on a global or a local scale. The volume, velocity and variety of data that gets generated on twitter fits the description of big data.

Big Data:

Big data is a term for collection of data sets so large and complex that are difficult to manipulate or interrogate with standard database management tools. The 4Vs have now become the defining characteristics of big data which distinguish it from other types of Business Intelligence concepts.

- Volume:
  Refers to the vast amount of data that gets generated every second. Government and Enterprises are now flooded with data coming in from various sources and channels and need to make sense of it. Data is being generated from mobile devices, social media (Facebook posts, twitter, Instagram etc.), traditional business interactions (like Walmart / online purchases), financial transactions, sensors in everyday objects like cars, airplanes, refrigerators, etc.
- Velocity:
  Speed at which data is generated / delivered / captured / analyzed. Twitter users generate on average ~500 million tweets / day. Or for catching credit card frauds, companies must analyze 5 million trade events daily to identify fraud. Big data technologies allow us to analyze data while it is being generated without ever putting it into a database.
- Variety:
  Data can come in any form / type. Streaming / non-streaming. Structured / Unstructured / Semi structured. For example, emails, videos, audio, tweets, clicks, log files, sensor data, transaction data etc.
- Veracity:
  Refers to uncertainty in the data. With many forms of data (for example twitter posts, Facebook, websites a metric like social sentiment – quality and accuracy of which can be questioned. The volumes involved often make up for lack of quality / accuracy.

Architecture diagram:



Hadoop:

Hadoop is a highly scalable analytics platform for processing large volumes of structured and unstructured data. It is a distributed computing framework with two main components: a distributed file system and a map-reduce implementation.

Apache Hadoop is High-availability distributed framework that offers a distributed storage system via its HDFS (Hadoop Distributed File System) and processing management system. Hadoop provides possibility to store data offers in the duplicating. On the other hand, Hadoop offers data processing framework on large data volumes called MapReduce. The MapReduce architecture is composed of two phases: the map and reduce phase. Initially, the input data can be divided into several copies as where the key is the word and the value indicate how many times the word has occurred and assigns to each underemployment the task trackers. Finally, in the phase interruption, the results of each job tracker are combined to produce the finale results.

While dealing with Big data, we need a system which is highly fault tolerant, which is designed for low-cost hardware, holds up very large amount of data. In this case we are dealing with Twitter data and Survey data which includes variety of data including user generated text and images and volume of data. Hadoop can store and process any file data: large or small, be it plain text files or binary files like images, even multiple different version of some particular data format across different time periods. Some of the important characteristics of HDFS is the storage and processing in a distributed environment, streaming access to data files. At the level of data security, Hadoop provides itself with file permissions and user authentication thereby ensuring utmost security of the data.

Spark:

Apache Spark is a fast, in-memory data processing engine that is suitable for use in a wide range of circumstances. We incorporate Spark into our application to rapidly query, analyze, and transform data at scale. Spark provides an interface for programming entire clusters with implicit data parallelism and fault tolerance. It is a unified analytics engine for big data and machine learning. The data in the local drive on the Cluster is copied to HDFS for Spark to access.

Data Analysis:

Perceptual mapping:

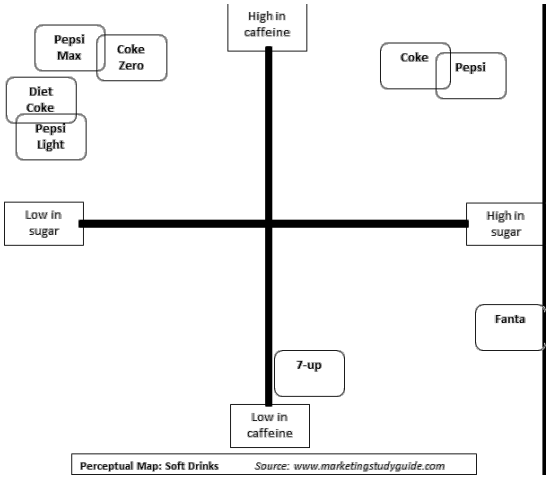
Perceptual mapping is a diagrammatic technique used by asset marketers that attempts to visually display the perceptions of customers or potential customers. Typically, the position of a company's product, product line, or brand is displayed relative to their competition. Perceptual maps, also known as market maps, usually have two dimensions but can be multi-dimensional. The word 'perceptual' comes from the word 'perception', which basically refers to the consumers' understanding of the competing products and their associated attributes.

Main types of Perceptual Maps:

There are two main formats for a presenting a perceptual map.

1. Using two determinant attributes

   The first format simply uses two determinant attributes on the graph. Below is a simple example of a perceptual map for soft drinks in this format.



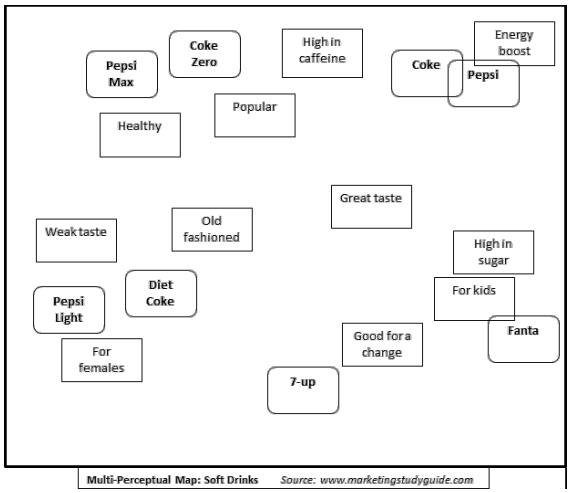
   The main advantage of this presentation format is that it is very simple to construct and interpret. In this case only two product attributes have been considered, they are 'to what extent does the consumer consider the product to be high/low in sugar' and 'to what extent is a product considered high/low in caffeine'. The simple combination of these two scores

(probably obtained from a consumer survey) places the product offering onto the map. For example, on this map, the 7-Up product offering is perceived as having a moderate level of sugar and being relatively low in caffeine'.

2. Using many product attributes

The second approach to perceptual mapping has the capacity to map multiple product attributes at the same time. This type of map is a little bit more confusing and difficult to interpret, but it does provide a good overview of how the target market views and connects the various attributes.

It is to be noted that there are no defined axes in this type of perceptual map. Instead the various product attributes are scattered throughout the map, along with the perceived positioning of the various product offerings.



Data:

The survey dataset contains customer ratings of 9 brands A, B, C, D, E, F, G, H, I, J on various attributes namely:

- perform: Brand has strong performance
- leader: Brand is a leader in the field
- latest: Brand has the latest products
- fun: Brand is fun
- serious: Brand is serious
- bargain: Brand products are a bargain
- value: Brand products are a good value
- trendy: Brand is trendy
- rebuy: would buy from Brand again

Since there are multiple variables, we are going to follow the second approach i.e., multi-attribute perceptual mapping.

Measures:

In real world data analysis tasks, we analyze complex data i.e. multi-dimensional data. We plot the data and find various patterns in it or use it to train some machine learning models. As the dimensions of data increases, the difficulty to visualize it and perform computations on it also increases. So, we have to reduce the dimensions of a data
1. Remove the redundant dimensions
2. Only keep the most important dimensions

Let's try to understand some terms

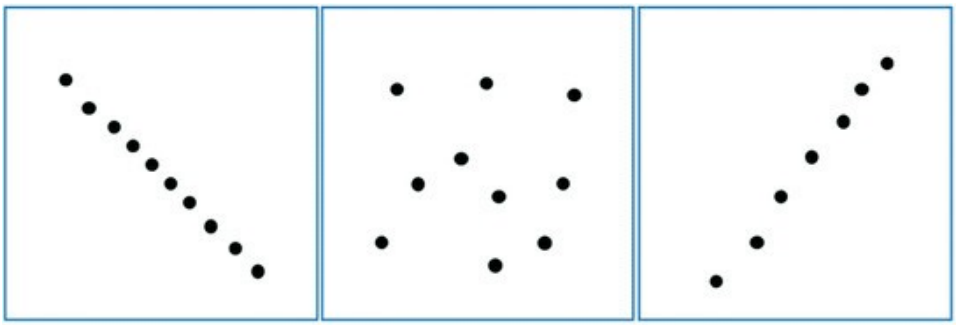
Variance:

It is a measure of the variability or it simply measures how spread the data set is. Mathematically, it is the average squared deviation from the mean score. We use the following formula to compute variance *var(x)*.

$$var(x) = \frac{\sum(x_i - \bar{x})^2}{N} \qquad cov(x,y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N}$$

Covariance:

It is a measure of the extent to which corresponding elements from two sets of ordered data move in the same direction. Formula is shown above denoted by *cov(x,y)* as the covariance of *x* and *y*.
Here, *xi* is the value of x in *ith* dimension.
x bar and *y bar* denote the corresponding mean values.
One way to observe the covariance is how interrelated two data sets are.



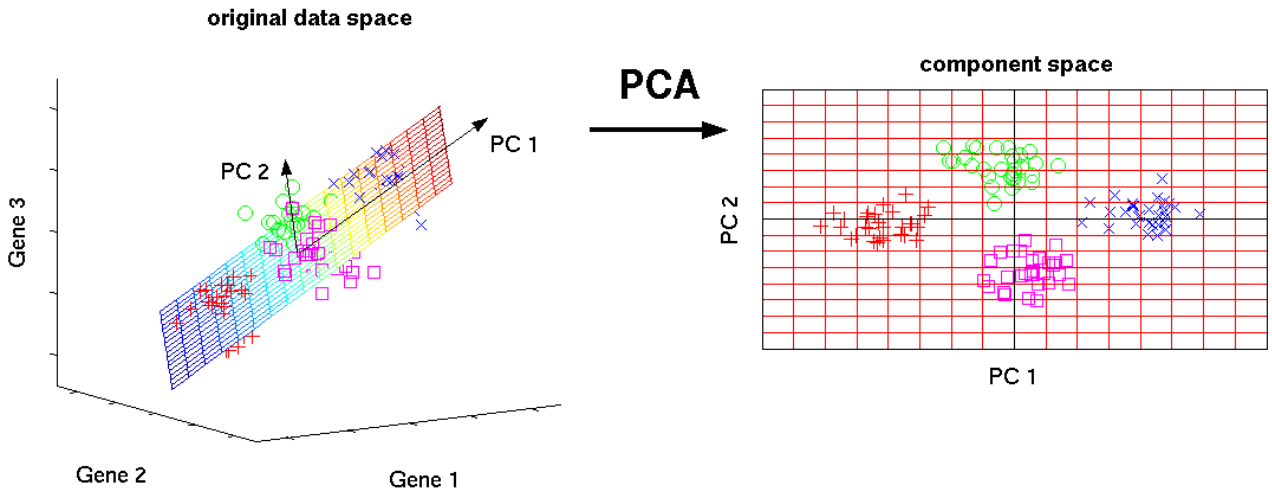
Positive covariance means X and Y are positively related i.e. as X increases Y also increases. Negative covariance depicts the exact opposite relation. However, zero covariance means X and Y are not related.

Now let's think about the requirement of data analysis.
Since we try to find the patterns among the data sets so we want the data to be spread out across each dimension. Also, we want the dimensions to be independent. Such that if data has high covariance when represented in some *n* number of dimensions then we replace those dimensions with linear combination of those n dimensions. Now that data will only be dependent on linear combination of those related n dimensions. *(related = have high covariance)*

Principal Component Analysis (PCA):

PCA finds a new set of dimensions (or a set of basis of views) such that all the dimensions are orthogonal (and hence linearly independent) and ranked according to the variance of data along them. It means more important principle axis occurs first. (more important = more variance/more spread out data)

The image below shows the transformation of a high dimensional data (3 dimension) to low dimensional data (2 dimension) using PCA.



Principal components:

A principal component is a normalized linear combination of the original predictors in a data set. In image above, *PC1* and *PC2* are the principal components. Let's say we have a set of predictors as X^1, X^2...,X^p

The principal component can be written as:

$$Z^1 = \Phi^{11}X^1 + \Phi^{21}X^2 + \Phi^{31}X^3 + ... + \Phi^{p1}X^p$$

where,

- Z^1 is first principal component

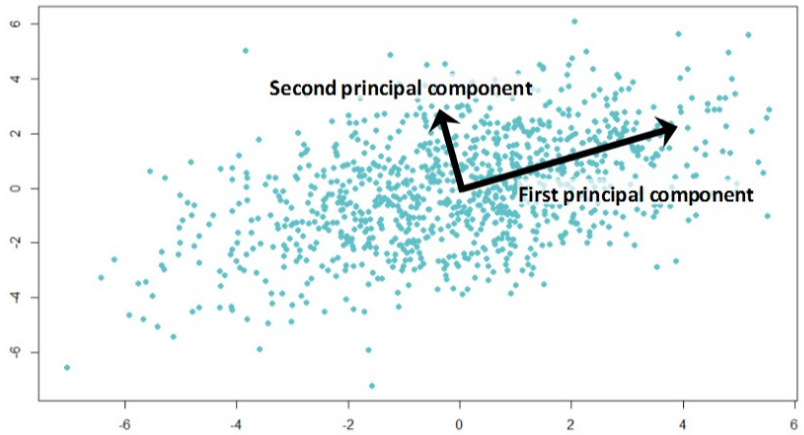
- Φ^{p1} is the loading vector comprising of loadings (Φ^1, Φ^2..) of first principal component. The loadings are constrained to a sum of square equals to 1. This is because large magnitude of loadings may lead to large variance. It also defines the direction of the principal component (Z^1) along which data varies the most. It results in a line in *p* dimensional space which is closest to the *n* observations. Closeness is measured using average squared euclidean distance.
- $X^1..X^p$ are normalized predictors. Normalized predictors have mean equals to zero and standard deviation equals to one.

Therefore, first principal component is a linear combination of original predictor variables which captures the maximum variance in the data set. It determines the direction of highest variability in the data. Larger the variability captured in first component, larger the information captured by component. No other component can have variability higher than first principal component. The first principal component results in a line which is closest to the data i.e. it minimizes the sum of squared distance between a data point and the line. Similarly, we can compute the second principal component also.

Second principal component (Z^2) is also a linear combination of original predictors which captures the remaining variance in the data set and is uncorrelated with Z^1. In other words, the correlation between first and second component should is zero. It can be represented as:

$$Z^2 = \Phi^{12}X^1 + \Phi^{22}X^2 + \Phi^{32}X^3 + ... + \Phi^{p2}X^p$$

If the two components are uncorrelated, their directions should be orthogonal (image below). This image is based on a simulated data with 2 predictors. Notice the direction of the components, as expected they are orthogonal. This suggests the correlation b/w these components in zero.



All succeeding principal component follows a similar concept i.e. they capture the remaining variation without being correlated with the previous component. In general, for *n* × *p* dimensional data, min (*n-1, p)* principal component can be constructed.

Steps to generate Perceptual Map:

1. **Initial data cleaning**

   PCA can be applied only on numerical data. Therefore, if the data has categorical variables they must be converted to numerical. The basic data cleaning prior to implementing this technique needs to be done.

2. **Load and Standardize Data**

   We'll load the data and store it in a pandas data frame. The data set contains ratings from 1210 users (instances) for 9 features of 9 brands. Even though all of the features in the dataset are measured on the same scale (a 0 through 5 rating), we must make sure that we standardize the data by transforming it onto a unit scale (mean=0 and variance=1). Also, all null (NaN) values were converted to 0. It is necessary to transform data because PCA can only be applied on numerical data.

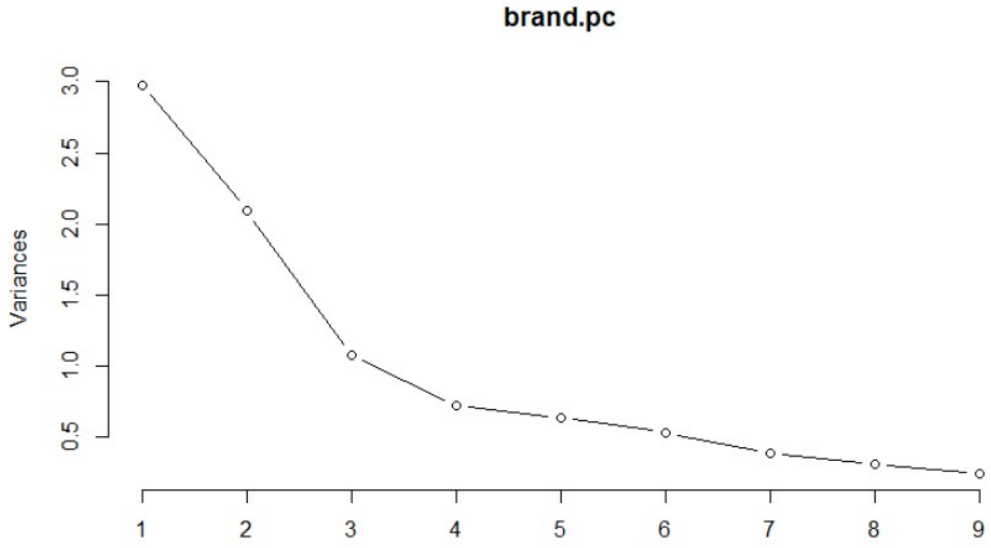
3. **Covariance Matrix and Eigen decomposition**

   A covariance matrix is created based on the standardized data. The covariance matrix is a representation of the covariance between each feature in the original dataset. After the covariance matrix is generated, eigen decomposition is performed on the covariance matrix. Eigenvectors and eigenvalues are found as a result of the eigen decomposition. Each eigenvector has a corresponding eigenvalue, and the sum of the eigenvalues represents all of the variance within the entire dataset.

4. **Selecting Principal Components**

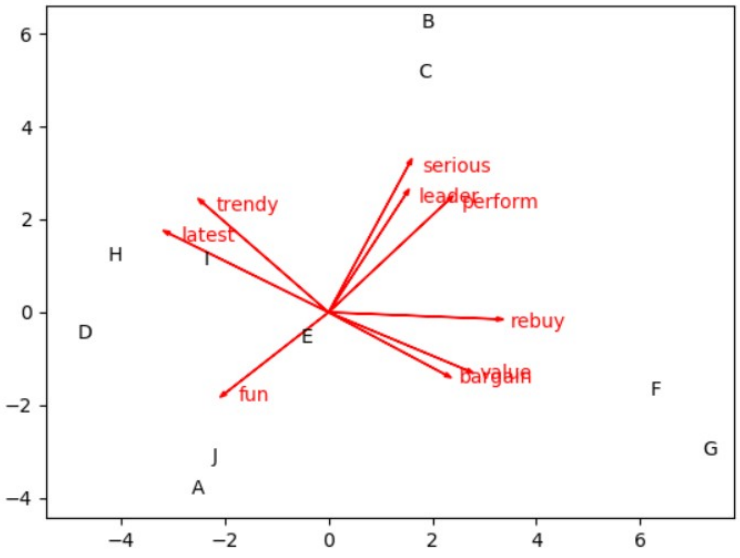
   Scree plot:

   A scree plot plots the variances against the number of the principal components. It is often interpreted as indicating where additional components are not worth the complexity; this occurs where the line has an elbow, a kink in the angle of bending, a somewhat subjective determination. The elbow occurs at either component three or four, depending on interpretation; and this suggests that the first two or three components explain most of the variation in the observed brand ratings.



5. Visualizing PCA

A good way to examine the results of PCA is to map the first few components, which allows us to visualize the data in a lower-dimensional space. A common visualization is a biplot, a two-dimensional plot of data points with respect to the first two PCA components, overlaid with a projection of the variables on the components.



We see the result in above figure, where adjectives map in four regions: category leadership (serious, leader, and perform in the upper right), value (rebuy, value, and bargain), trendiness

(trendy and latest), and finally fun on its own. We interpret the adjective clusters and relationships and see four areas with well differentiated sets of adjectives and brands that are positioned in proximity. Brands F and G are high on value, for instance, while A and J are relatively high on fun, which is opposite in direction from leadership adjectives (leader and serious).

Marketing Insights from Perceptual Maps:

Case Study 1:

From the map, it can be observed that brand E doesn't have distinct brand image in customer perception. It depends on the strategic goals to take a decision about repositioning the brand. If the brand manager wishes to increase differentiation, one possibility would be to take action to shift the brand in some direction on the map. Suppose they wanted to move in the direction of brand C. After the analysis from the perceptual map, it can be observed that the brand E can be positioned as brand C by strengthening perform and serious attributes and dialing down value and fun.

Case Study 2:

Another option would be not to follow another brand but to aim for differentiated space where no brand is positioned. There is a large gap between the group B and C, versus F and G. This area might be described as the "value leader" area or similar. This suggests that brand E could target the gap by increasing its emphasis on performance while reducing emphasis on latest and fun.

Conclusion:

To summarize, when we wish to compare several brands across many dimensions, it can be helpful to focus on just the first two or three principal components that explain variation in the data. We can select how many components to focus on using a scree plot, which shows how much variation in the data is explained by each principal component. A perceptual map plots the brands on the first two principal components, revealing how the observations relate to the underlying dimensions (the components).

PCA may be performed using survey ratings of the brands (as we have done here) or with objective data such as price and physical measurements, or with a combination of the two. In any case, when you are confronted with multidimensional data on brands or products, PCA visualization is a useful tool for understanding differences in the market.

References:

1. https://www.freelancer.com/community/articles/twitter-data-mining-a-guide-to-big-data-analytics-using-python
2. http://www.digitaljournal.com/tech-and-science/science/using-twitter-for-big-data-analytics-to-analyze-disasters/article/500364
3. https://www.maritzcx.com/blog/survey-data-cleansing-five-steps-for-cleaning-up-your-data/
4. https://www.bernardmarr.com/default.asp?contentID=766
5. https://www.computerweekly.com/feature/Big-data-storage-Hadoop-storage-basics
6. https://databricks.com/spark/about
7. https://en.wikipedia.org/wiki/Multidimensional_scaling
8. https://en.wikipedia.org/wiki/Perceptual_mapping
9. http://www.segmentationstudyguide.com/understanding-perceptual-maps/