MT5757 Applied Multivariate Statistics
Group 3

Researchers – **150010758, 180025784, 180024762**

The purpose of this report is to identify relationships between variables in a multi-dimensional dataset. Since the boom of social media and its subsequent impact on consumer advertisement, promotional activity for business clients has significantly grown on the web platform. Platforms such as Facebook, Twitter and Tumblr are often used as risk free alternatives to high profile advertisement options to better extend the reach of consumer attention. Applications of which extends for both small and large companies.

We have identified a successful cosmetics company who are utilising Facebook to promote their activities to a worldwide audience. In particular, the firm concentrates on promoting a single brand by posting specialised content between 1st of January to 31st December of 2014. This provides a unique comparison between the reaches of a traditional paid advertisement campaign and a social media approach which removes the commercialised paid element. Furthermore, data across the year could be analysed for seasonal impacts, to identify optimised marketing trends. The interesting aspect of this data is that it can be analysed to answer pressing management questions with regards to the correct usage of company's capital resources towards marketing.

The dataset is comprised of 500 observations of social media metrics used by a cosmetics company. The metrics involves the following 19 variables in the table below and has been modified to analytical purposes, thus satisfying the multi-dimensional constraint for this report.

Variable	Output
Page Total likes	Integer, total number of likes for the page at time of post publication
Type	Categorisation, 1 - Status; 2 - Photo; 3 - Video; 4 - Link
Category	Categorisation, 1 - Action; 2 - Product; 3 - Inspiration
Post Month	Integer, month of the post publication
Post Weekday	Integer, day of the week of the post publication
Post Hour	Integer, hour of the day of post publication in 24hr time
Paid	Binary, 0 - not paid promotion; 1 - paid promotion
Lifetime Post Total Reach	Integer, the number of unique users who saw the post
Lifetime Post Total Impressions	Integer, the number of times the post is displayed
Lifetime Engaged Users	Integer, the number of unique users who clicked anywhere on the post
Lifetime Post Consumers	Integer, the number of users who clicked anywhere on the post
Lifetime Post Consumptions	Integer, the number of clicks anywhere on the post
Lifetime Post Impressions by people who have liked your Page	Integer, the number of impressions of users who have liked the page
Lifetime Post reach by people who have liked your Page	Integer, the number of unique users who have seen the post because they liked the page
Lifetime People who have liked your Page and engaged with your post	Integer, the number of unique users who have liked the page and clicked on the post
Comment	Integer, the number of comments on the post
Like	Integer, the number of likes of the post
Share	Integer, the number of shares of the post
Total Interactions	Integer, sum total of likes, comments and shares

From the table above, it can be seen that the dataset involves 16 integer variables, 2 categorical variables and 1 binary variable. The Type variable was initially displayed as a character variable but for the purposes of Principle Component Analysis was transformed into a categorical variable as explained above. There were 5 missing values in the dataset and therefore required additional cleaning of the dataset.

One important aspect to be considered is the difference between reach, interactions and impressions. Impressions is defined to be the number of times a particular content has been displayed. Thus, one

individual could have multiple impressions if the post has displayed to the individual more than once. This is contrasted with reach which counts the total number of unique users who has seen the post. Interactions is thus defined as the total number of interactions made with the post by 'liking', commenting on or sharing the post.

As discussed above, the purpose of this report is to observe relationships between variables of interest and thus identify patterns in our dataset which could be utilised to answer a commercial issue. As a group, it was identified that the Type of content posted, which follows a categorical variable of 1-4 (1=Status, 2=Photo, 3=Video, 4=Link) would be an appropriate response variable for our purposes. Our aim is to identify patterns in consumer behaviour which would lead the Cosmetic company to devote resources between the 4 types of social media content. We would seek to understand which type of content has the biggest consumer reach and the underlying mechanisms that governs this behaviour.

Similarly, Category was also seen as an appropriate response variable of interest which follows categorical variables between 1-3 (1=Action, 2=Product, 3=Inspiration). This combined with Type of content will provide invaluable information for consumer analysis. Using the other variables as metrics for evaluation, we could potentially unravel underlying trends in marketing outputs and coupled with the Type of Content would provide the cosmetics company an optimised marketing strategy.

Furthermore, the Paid variable which follows a binary output of 0=Unpaid, 1=Paid can be utilised as a cost benefit mechanism but this will have second priority compared to the first two response variables.