# IS507 - Final Project Methods Draft

## Section BC - Group 16: Online News Popularity

The dataset chosen for the research is obtained from the UCI Machine Learning Repository and is titled Online News Popularity Dataset. It has 39,797 observations and 61 features. We have performed Data Cleaning by removing potential outliers and imputing the observation values with the mean in place of 0 after Exploratory Data Analysis.

Before performing any research techniques on the data, we intend to perform dimensionality reduction using Principal Component Analysis (PCA) and identify the reduced components. PCA assumes linearity among the data and is sensitive to the scale of features, so we need to standardize the data. Multivariate normality is an assumption of PCA that is not strict and can be tested with a version of Shapiro-Wilk. We check for factorability using the KMO test for sampling adequacy, which ideally should be > 0.7. Bartlett's test of sphericity tests whether the correlation matrix is similar to the identity matrix. Cronbach's alpha coefficient describes the internal consistency among the variables, and the higher the value, the more the variables within a component are reliable. We will predict the number of shares/popularity of online news using appropriate machine learning methods. XGBoost, a popular boosting technique can be utilized to ascertain this, as this technique provides regularized learning that helps reduce the loss function, preventing overfitting. Boosting may assume an ordinal relationship between the encoded values for the input variables. Additionally, XGBoost can only handle numeric vectors. If any of these assumptions are found in the dataset, we shall encode the categorical variables to make them numerical. We hypothesize that we can predict the shares of online news accurately. We will find similar articles based on the polarity of the words in articles, which may be related to positive, negative, or neutral sentiments. This will be done by using the Hierarchical Clustering on Principal Components (HCPC) technique. PCA will help us choose the number of dimensions to be retained in the output. Hierarchical clustering will be performed using Ward's criterion on the selected principal components. We will choose the number of clusters using the hierarchical tree and then perform K-means Clustering. This method is suited as HCPC is very useful as the dataset is large and has 58 continuous variables. Furthermore, the PCA step can be considered a denoising step, leading to a more stable clustering.

A hypothesis that we will work on is whether the type of data channel (entertainment, lifestyle) influences the number of shares of news articles. Combining the data channel categories, we can get a single variable whose attributes will be categorical. Since our outcome variable is continuous, we can use a one-way ANOVA or Kruskal-Wallis test to determine the validity of the hypothesis. The data will be checked for normality and run the test. Another hypothesis would be to check any association between when the online article is published and its number of shares. We can run unpaired T-Test or Mann-Whitney tests to extract a potential pattern by comparing the outcome variable with various binary outcomes that we have in our hands. The data will be checked for normality and variances, upon which we will determine the final test we will proceed with.

We will use the R statistical program with version 4.2.2, including the following packages: plyr, dplyr, readr, tidyr, stringr, DescTools, xgboost, caret, psych, FactoMineR, factoextra, HCPCshiny, RVAdeMoire, ggplot2