Machine Learning

Kisuc Kim

A3: Unsupervised Learning Case Study

Analyzing App Download Behavior – Market Segmentation

There are millions of mobile apps. But there are also different people, devices, and purposes. Demographically and geographically, the patterns of application usage behavior vary based on individuals' preferences. The ultimate goal of this analysis is to understand the mobile app landscape in United States, find meaningful insights, and to recommend actionable suggestions. In order to achieve the goal, firstly our market needs to be specified in detail. Given dataset from an online research was conducted with 1522 respondents. Developing an effective automated machine learning model to segment the sample respondents will be the first step. The sample was needed to be analyzed demographically before clustering based on their behavior and preference. There are some findings. The majority of samples is for iOS users (50% of total respondents is using iPhone, 24% is iPod, and some of 30% from tablet users). To normalize the age feature, the original age group merged to 10s, 20s, 30s, 40s, 50s, and 60s (See Appendix 1, 3). Race of respondents were imbalanced (See Appendix 2). There are 60% of single device users, 40% Multi devices users. Using many apps are considered to understand the respondent's usage pattern. But, just features related to the number of apps were disregarded. Having many apps and using many apps are different. According to the correlation analysis based on given data, it is found that the proportion of free apps to download is not strongly correlated to users' behavior. As a rule of thumb, if an absolute value of correlation is 0.6 or higher, it is generally considered to be a strong correlation (BMJ, 2019). For behavior factors, a correlation between luxury brands and designer brands are also found.

There was preference information regarding users' behavior amongst 88 variables. In this analysis, many methodologies are used. Fist, agglomerative clustering was a great method to segment all users based on unsupervised machine learning. This hierarchical clustering can automatically segment and give us a dendrogram visualization. It is easy to segment all users but not easy to understand what you can do for each cluster (See Appendix 4). Instead of segmenting 100% all users, using only 4 components, focusing on 40% of users with interpretable behavior and preferences will be more helpful to develop effective marketing activities. Also, 5 was the optimal number of clusters based on inertia plot (See Appendix 4, 5). PCA and K-Means clustering gave me 4 groups (Conservative, Followers, YOLO (Carpe Diem), Smart Consumers). Conservative group seemed too skeptical. YOLO group is good market area but they will not be easy to predict their future preference, Smart Consumer group tends like a cherry picker. Follower group would be a best and safer choice. After clustering, an additional correlation analysis gives me the fifth cluster are highly correlated to music and entertainment (See Appendix 9).

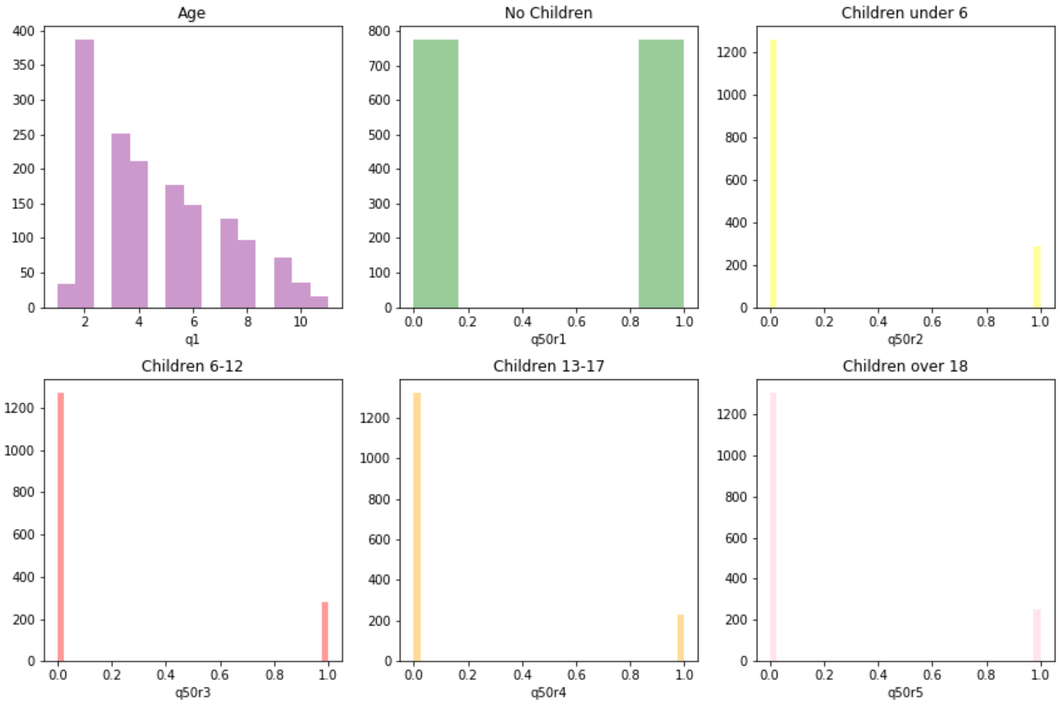
Here is an actionable insights and recommendations based on this analysis. Focusing on specific group and cluster can lead to success of new app marketing activities. Music or entertainment application mainly focusing on cluster 5 and follower group can be profitable. Because the follower group tends to buy when they are recommended a good deal. Even, they are open to be updated and waiting for a good recommendation. They are overwhelmed in the world with too much information. If you are targeting the 5th cluster in the follower group, you will be able to target all of age range, and all genders almost equally. I would highly recommend conducting a focus group meeting for them as the secondary research. Then, you will be able to know what specific entertainments have in more potential area.

Majority of respondents identified themselves as White or Caucasian. It can refer that this research was conducted in some local areas where White or Caucasian people are mainly living. Mobile app can be downloaded in everywhere regardless of specific location. Unless you are targeting only for the specific area or people, online survey in many areas or considering racial and ethnic stratification would be a good idea to avoid this kind of disproportion issue

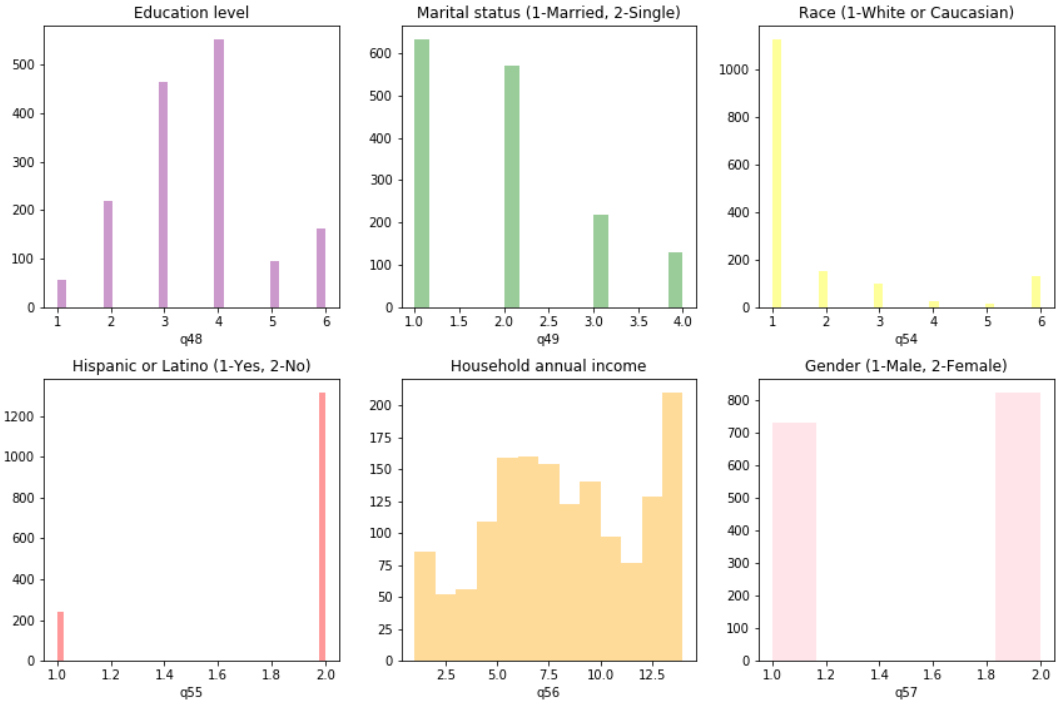
< References >

BMJ. (2019, March 28). “11. Correlation and regression”. The bmj. Retrieved from: https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/11-correlation-and-regression

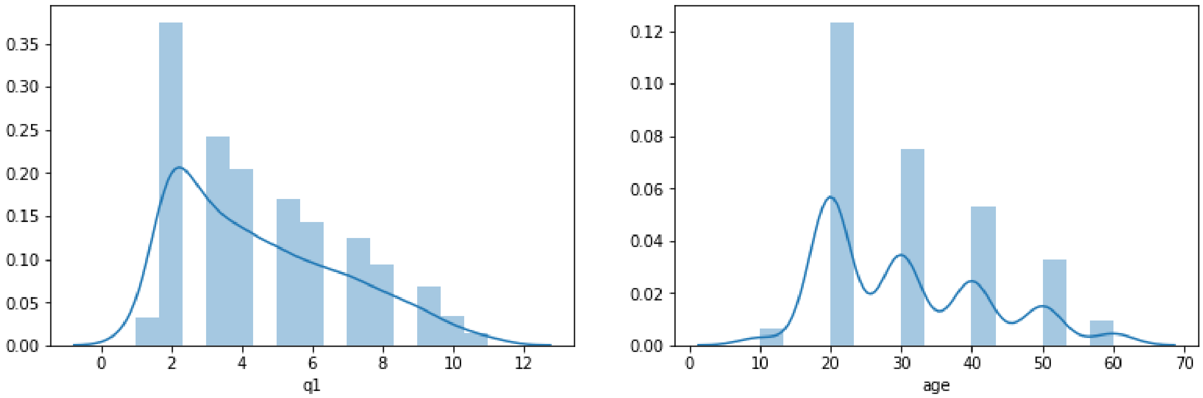
< Appendix >



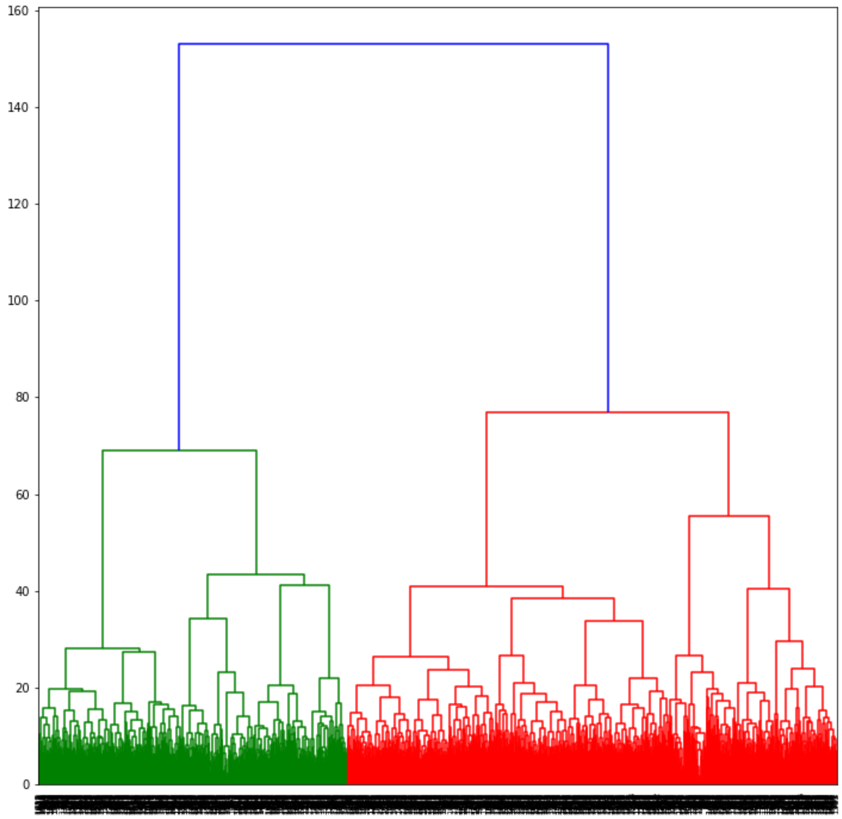
< Appendix 1 – Histogram about age information (Respondents, respondents' children) >



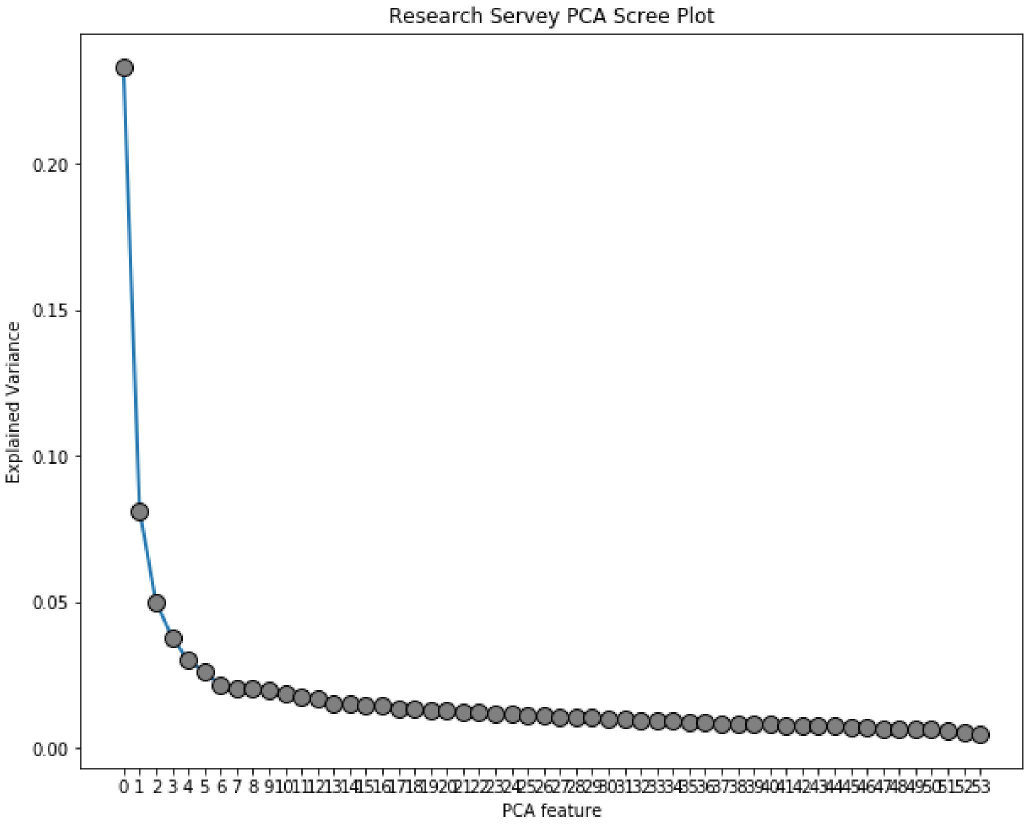
< Appendix 2 – Histogram for Comprehensive Demographic Information >



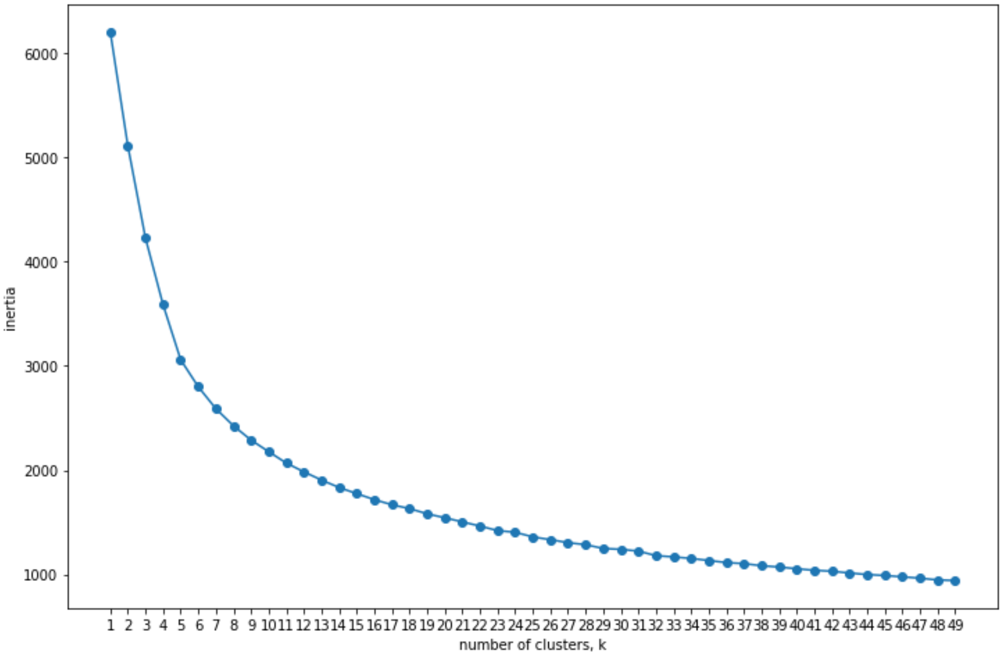
< Appendix 3 – Normalizing age group (Left – Before, Right - After >



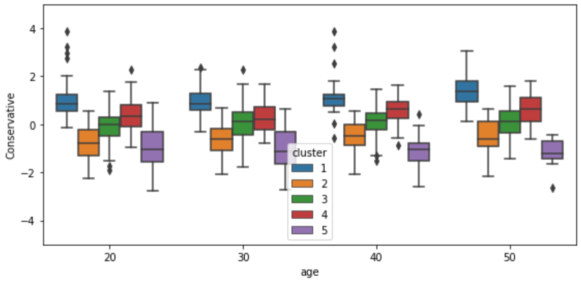
< Appendix 4 – Dendrogram >



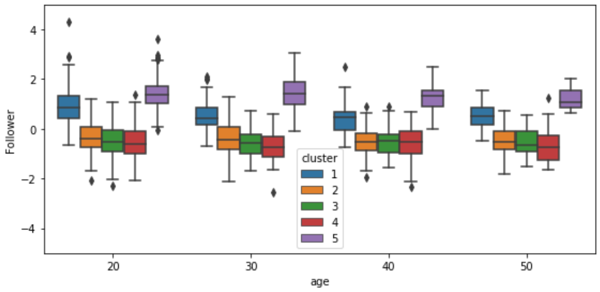
< Appendix 5 –Scree Plot for Principal Component Analysis >



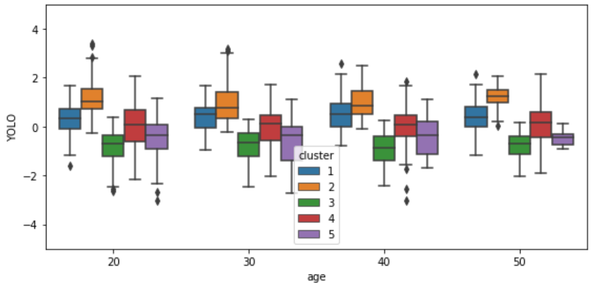
< Appendix 6 – Inertia Plot to find the optimal number of clusters for K-Mean Clustering >



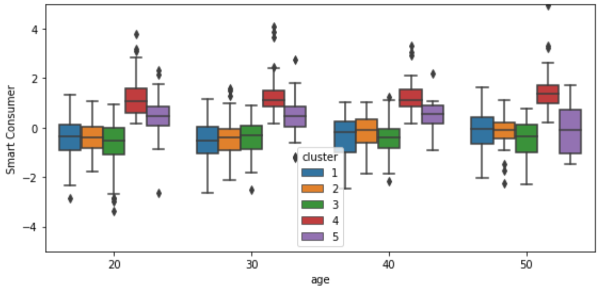
- Conservative vs Age -



- Follower vs Age -

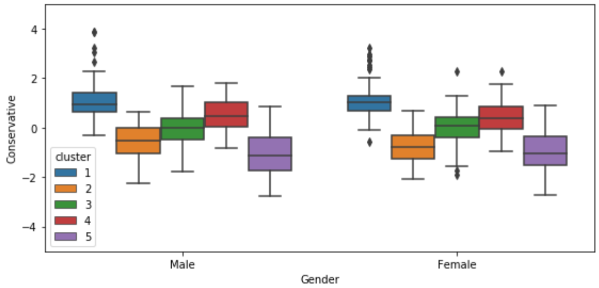


- YOLO vs Age -

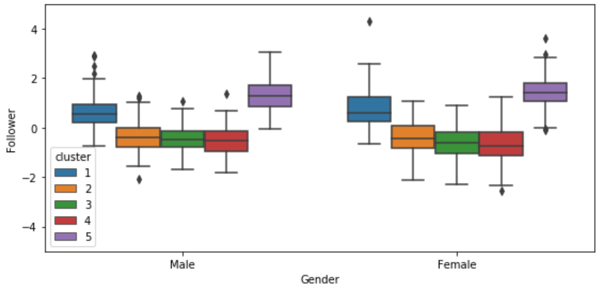


- Smart Consumer vs Age -

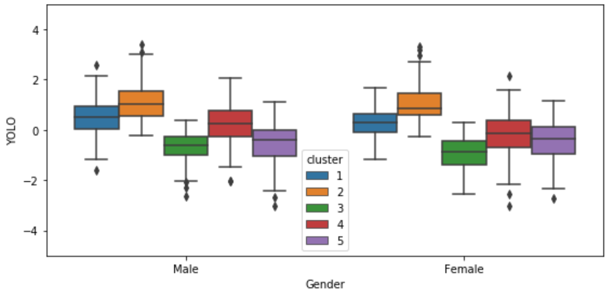
< Appendix 7 – Box Plot (PCA groups vs Age) >



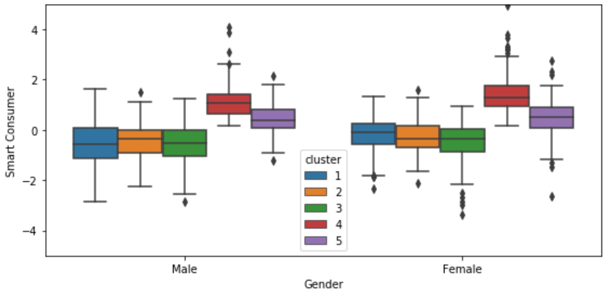
- Conservative vs Gender -



- Follower vs Age -

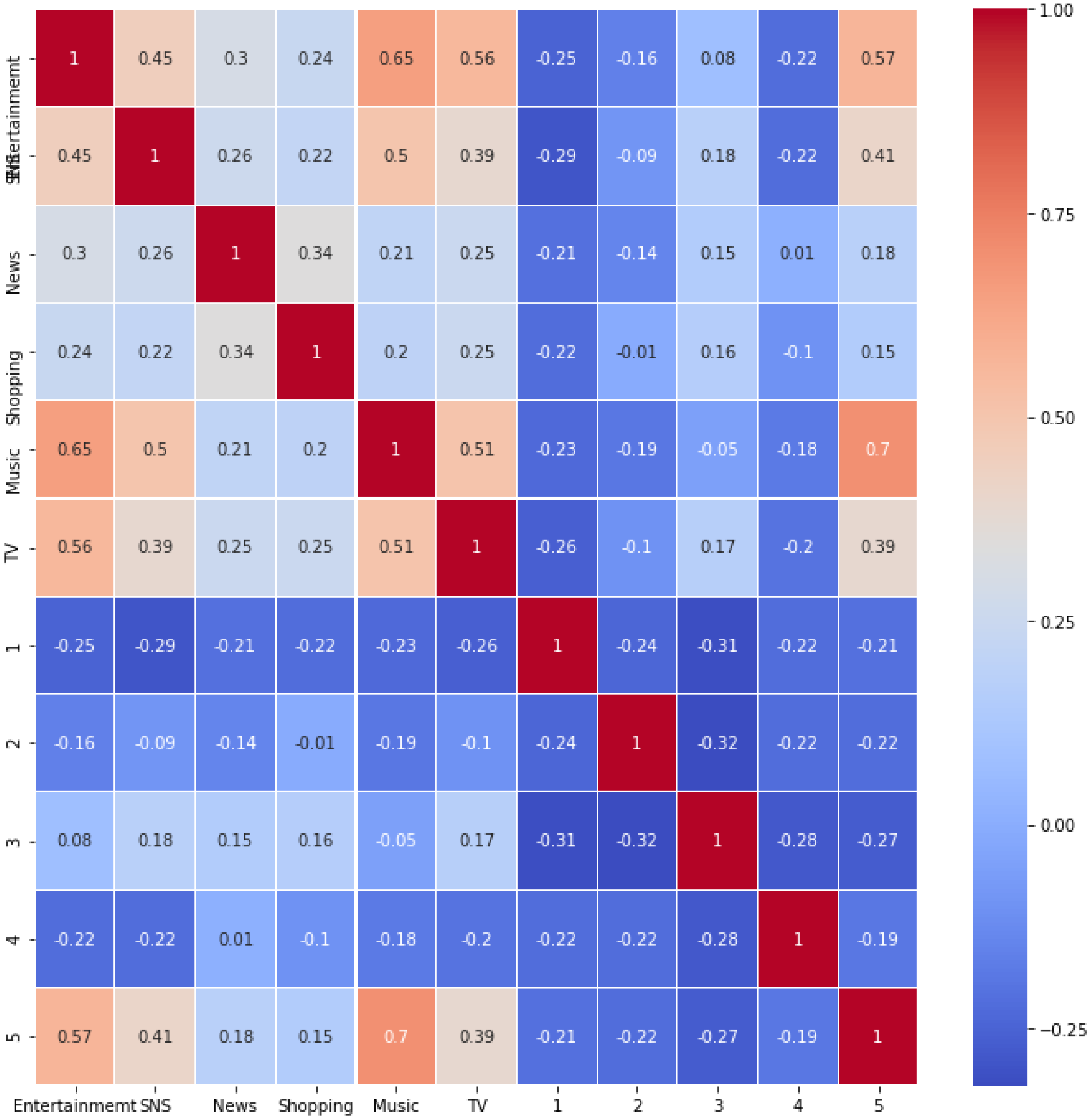


- YOLO vs Gender -



- Smart Consumer vs Gender -

< Appendix 8 – Box Plot (PCA groups vs Gender) >



< Appendix 9 – Heatmap (Correlation Purpose vs Cluster) >