Examining the Linkage Between Different Traits and Preferences in the Young People Survey

Emily Wang, Xi Chen, Tuan Huynh, Jiazhen Cui
INFO 633 – Information Visualization
College of Computing and Information
Drexel University
Professor Chaomei Chen
December 11, 2021

***Abstract*—The goal of this study was to uncover behavior patterns and correlations between different traits and preferences among 1,010 young people (aged 15-30) in Bratislava, Slovakia. We used several computer-based technologies including Python, Orange, and Gephi to create various visualization to examine the linkage between different traits and preferences. Through dimensionality reduction by t-SNE and PCA along with hierarchical clustering, we found differences in music preferences in males and females. Specifically, we found that the female participants tend to enjoy music more than male participants. We also created network visualizations based on the survey responses to personality traits and music preferences and found the connection between different traits. Through community detection in the networks, we found five clusters each in the personality traits network and music preferences network, which coincides with results from previous studies.**

***Keywords*—information visualization, network visualization, dimensionality reduction, clustering, survey, human behavior, Orange, Gephi**

I. Introduction

Every generation has unique interests which can be used to identify trends. These trends can provide vital information for various aspects of life. Age cohort give researchers a tool to analyze changes in views over time [1]. If we look at music preferences, for example, Baby Boomers will probably not enjoy listening to the same type of music as Millennials. By analyzing the results of this survey, we can provide a concrete visual representation of common traits among young people. This information can ultimately be used by many industries including marketing, e-commerce, and education.

II. Data

The chosen dataset is provided by Kaggle [2]. In 2013, students of the Statistics class at FSEV UK, located in Bratislava, Slovakia, were asked to invite their friends to participate in this survey. All participants were of Slovakian nationality, aged between 15 and 30. This survey presents various questions' results. The research questions consist of 139 numerical rating questions and 11 categorical questions. Those questions can be split into eight different groups such as Music Preferences (19 features), Movie Preference (12 features), Hobbies & Interests (32 features), Phobias (10 features), Healthy Habits (3 items), Personality Traits, Views on Life, & Opinions (57 features), Spending Habits (7 features) and Demographics (10 features). The features further separate each category into specific genres or subjects. Students are then asked to rank their interest level in each of these features. Based on the variables, this CSV data file contains 1,010 rows and 150 columns. The survey was presented to the participants in both electronic and written form. Also, the original survey was in Slovakian, and it was later translated into English.

Our mission in analyzing this dataset is to uncover interesting behavior patterns in young adults using an unsupervised learning approach through clustering and show the correlations between different interests using network graphs.

The authors are graduate students at Drexel University

Below are the motivating questions:

1. Do people make up any clusters of similar behavior? How many clusters will be found?
2. Do women have different interests than men?
3. Are there connections between different preferences?

Based on past studies, we expect music to cluster into five groups [3] and personality will cluster into five groups [4].

III. Methods

To create appropriate visualizations of the Young People Survey, several technologies need to be integrated. The data is obtained from Kaggle, which contains responses to all the questions on the survey.

To answer the first and second motivating questions, we need to create clusters based on the data. Since our dataset has high dimensionality (150 columns), the plan therefore is to visually explore the dataset in some lower dimensional space with dimension reduction. We explore two methods of representing data in a lower-dimensional space, with the first being using t-SNE. T-distributed stochastic neighbor embedding (t-SNE) is a useful statistical method for visualizing data with dimension reduction [5]. The second method is to utilize principal component analysis (PCA), which computes the principal components onto which each data point is projected to obtain lower-dimensional data while preserving as much of the data's variation as possible [6]. The clusters are produced using agglomerative (hierarchical) clustering, which forms clusters by calculating the distance between pairs of observations.

By using clustering for labeling in combination with dimensionality reduction for visualization, the data can be represented by a scatter plot. On this plot, highly similar topics will be close together while dissimilar topics will be further apart. We can then examine the clusters produced to find meanings in them.

After the analysis of the whole dataset, we split this data into different groups based on various features. In this project, we chose two group features: 1. music preferences; and 2. personality traits, views on life, & opinions. We aim to find the connections between these responses through graph visualization.

The data is pre-processed to create the node and edge files for the network graph. A correlation matrix is built using the pairwise Pearson correlation scores calculated from the ordinal responses to the survey questions. The correlation scores characterize the similarities or distances between the survey questions' Responses. These correlation scores are used as the edge weights, whereas the questions are the nodes. The nodes and edges are then visualized. An edge weight threshold is used to remove edges that are not significant. A graph of important edges provides a compact representation of the entire complex dataset [7]. To obtain easily interpretable networks, different layouts are then compared to yield the best result.

We will then detect communities in the networks to examine the clusters again. Community detection is a common operation in network analysis, and it consists of grouping nodes based on the graph topology. Communities, or clusters, are usually groups of nodes having a higher probability of being connected to each other than to members of other groups [8]. We will apply an algorithm to detect communities in the nodes and color them accordingly.

IV. Tools

Python is used to facilitate the pre-processing of the dataset. Python is an open-source programming language that has an extensive collection of libraries for a variety of tasks. Primarily, the Python pandas library was used to format and split up the data files based on the question category. The edge and node files for each question category are also created in Python. The node and edge files are then imported to Gephi to visualize networks. The clustering and network visualization tasks can both be completed in Python through other libraries, but we chose to explore other codeless data visualization tools due to their ease of use.

An exploration of the clustering is done in Orange to examine patterns in the entire dataset. Orange is an open-source data mining toolkit that features a visual programming "workflow" like presented in Fig 1. It provides a platform for powerful explorative data analysis and interactive data visualization. Each widget represents an operation on the data. Our workflow loads the data calculates the Euclidean distance completes hierarchical clustering based on the Euclidean distance., and then produces a scatter plot on the PCA axes showing the result of the clustering.

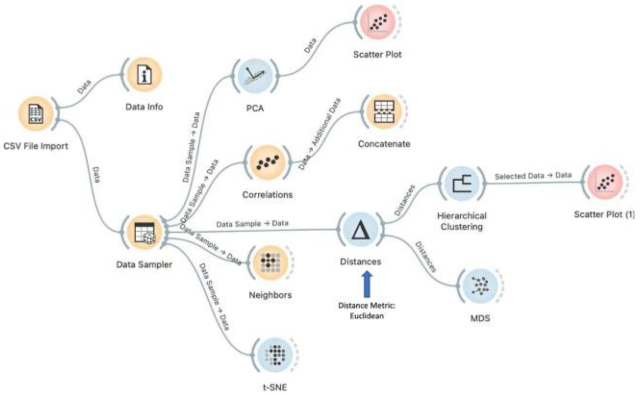

Fig. 1. Whole dataset Orange workflow.

Network visualization is created using Gephi. Gephi is a network visualization software used in various disciplines (social network analysis, biology, genomics...) [9]. One of its features is the ability to explore multiple layout options for the network, using layout algorithms such as ForceAtlas2, Frutcherman Reingold, Yifan Hu Proportional, just to name a few. It can also filter out nodes and edges of low importance. With the network visualization created in Gephi, it is possible to identify different linkage patterns and the formation of distinct clusters.

V. Results

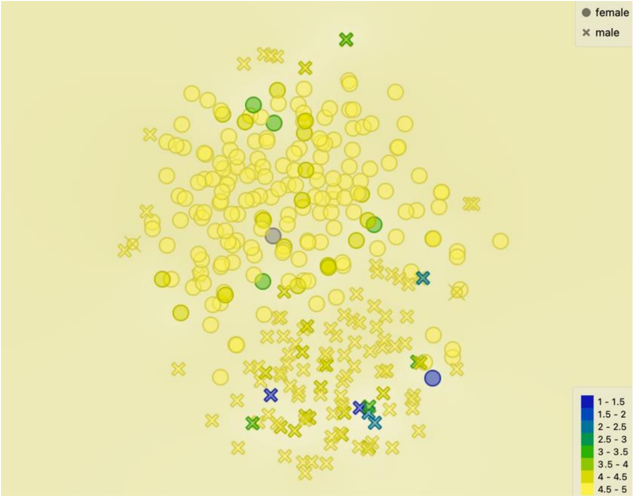

Fig. 2. Music prefrences t-SNE output.

One of the attributes we first wanted to analyze was Music. In Fig. 2 presented above, we chose to represent this data through a t-SNE visualization of Music based on gender difference. This figure represents the dimension reduction analysis of the data. The first interesting observation is that the clusters are clearly defined and segregated. With the exception of a few points outside the clusters, we are able to differentiate men and women into two distinct groups. Since about 95% of this visualization is yellow, we can also conclude that about 95% of people, both male, and female, enjoy listening to music. Then using the color scheme, which is used to portray music ratings, we are able to draw an additional conclusion that females enjoy listening to music more than males. This is because only a singular blue circle, compared to a darker green and blue “x’s”.

For Fig. 3, Multidimensional scaling (MDS) is a multivariate data analysis technique that displays “distance” data structure in low dimensional space. MDS is also suitable for sociology, quantitative psychology, marketing, and other statistician certificate analysis methods. Because our dataset comes from the Social Survey, we concluded that it is important to show the representation of MDS output. Compared with MDS, PCA also is a good way to reduce dimensions for data. In PCA, the methods we use for dimensionality reduction are to find the directions with orthogonal and maximum variance, because variance can save the information in the original feature space to the greatest extent.

As we can see, we did Hierarchical Clustering, which can combine the two most similar data points by calculating the similarity between the two types of data points and iterating this process repeatedly. Simply speaking, hierarchical clustering determines the similarity between data points of each category and all data points by calculating the distance between them. The smaller the distance, the higher the similarity. After doing Hierarchical Clustering, we chose to show our results in a Scatter Plot. From Fig. 4, it is split into 5 clusters. It looks similar to our t-SNE result which can show more information such as how many different clusters. We picked PC1 as Axis x and PC2 as Axis y. For choosing our attributes, we used Cluster to define the different colors and Gender as shapes. As we can see, approximately 95% red color and 100% blue color are female, and most of green and orange and yellow are male. For the blue color and green color, the distinction is clearer to see because the clusters are separated well. Almost 50% of red color and blue colors are overlapping. The most interesting thing is blue color is all-female without including any male.

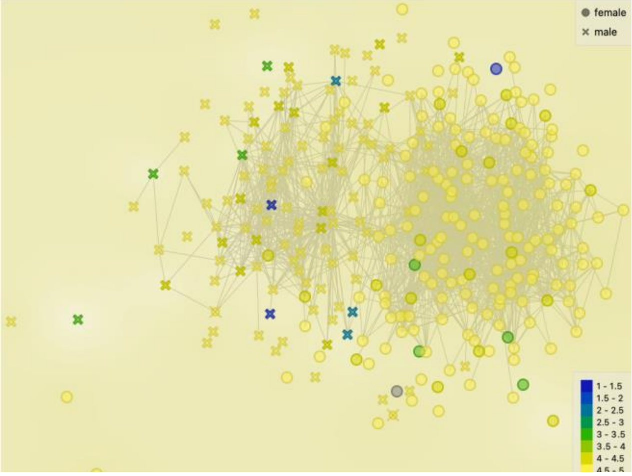

Fig. 3. Whole dataset MDS output.

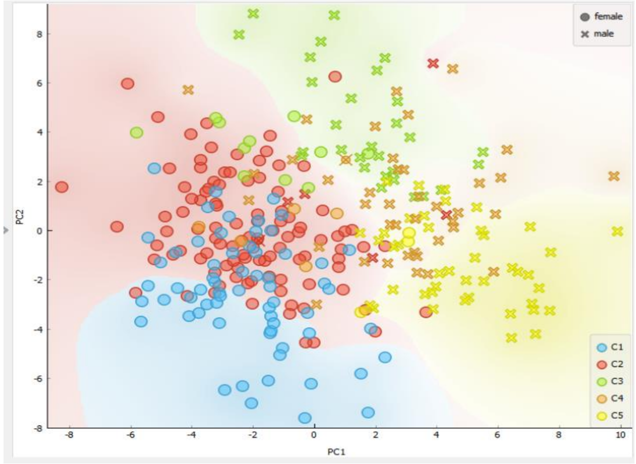

Fig. 4. Whole dataset hierarchical clustering output.

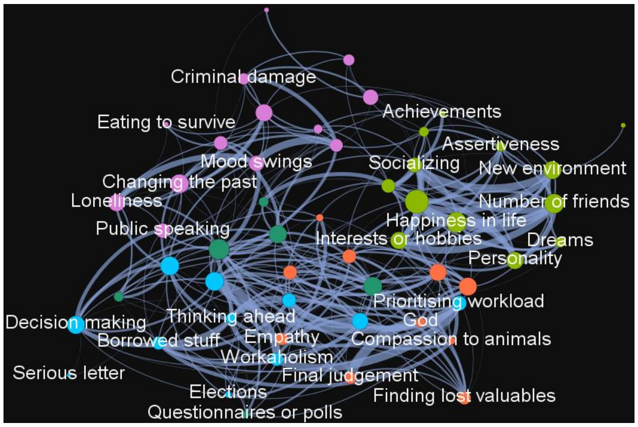

Fig. 5. Personality traits, views on life, & opinions Gephi output.

For Fig. 5, after data cleaning and preprocessing, this dataset contains 53 features. We controlled the threshold argument to remove edges that were not significant. The graph visualization network gives us some useful information. The first is that larger nodes are ranked at higher importance by the young people who participated in this survey. We can also conclude that these larger nodes have more features that are related to each other, meaning they have higher degrees. The colors of the nodes show the modularity class, or cluster, that they belong to. Through these connections, we can see that young people who have a higher "Number of friends" also have a higher number of "Interests or hobbies". At the same time, we can also gather that young people who have higher "Interests or hobbies" will also a higher "Number of friends" and higher "Energy levels".

Five modularity classes are found in the personality traits, which agrees with the theoretical predictions. Based on the node labels, we can map each of the modularity classes to one of the big five personality traits: light green – extroversion, purple – neuroticism, blue – conscientiousness, orange – agreeableness, and dark green – openess [4].

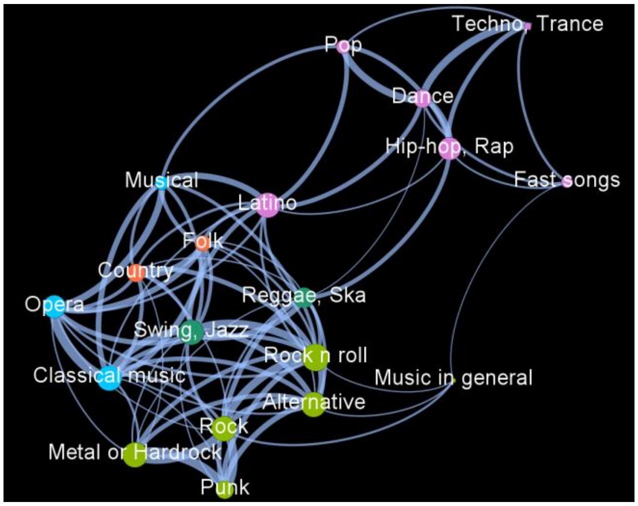

Fig. 6. Music preferences Gephi output.

Similar to Fig 5, Fig. 6 shows a graph network to analyze the music preferences. The node size represents the degree of the node and the edge thickness represents the strength of the connection. Through the visualization, we can see that the music preferences can be separated into two visual clusters, one centered around “Swing, Jazz” music and the other around “Hip-hop, rap”. The results in Fig 6 show strong links between certain music genres, including between opera and classical music, between punk and rock music, and between dance and pop music. Also notable are the connections between Latino music and musical. Latino music serves as a structural hole in the network, connecting nodes from different modularity classes. There are a couple of other structural holes in the network connecting these clusters, including the interests in music in general, reggae/ska music, and musicals.

The strong links are mostly found within modularity classes, which are represented by the node colors. The five modularity classes found also coincide with theory, which suggests that there exists a latent five-factor structure underlying music preferences [3]. Based on the modularity classes shown in Fig 6, we have interpreted and labeled these factors as orange – ruralness, pink – urbanness, blue – sophistication, dark green – smoothness, and light green – intensity.

VI. Discussion

The combination of different visualization tools and trial and error allowed the team to conduct a comprehensive investigation of the survey responses collected by FSEV UK and allowed for a thorough examination of the linkages between different personality traits and preferences. The “overview, zoom, filter, and details on demand” approach was utilized by the team to build a quality report as we first designed the visualization on the entire dataset, and then split the dataset up to investigate individual categories [10].

The overall goal was to identify and analyze the behavior patterns and correlations between various traits and preferences in young people. With the use of information visualizations, this goal was achieved by recognizing patterns and trends. The most interesting findings we uncovered included groupings of young individuals with similar tendencies, gender preferences for a variety of activities, and the correlation between these preferences.

To begin with, based on the results we have obtained, we detected that young people form 5 similar behavioral clusters for the motivating questions 1. In addition, in terms of the motivating question 2, the interest points of males and females are also different. The visualization in Fig. 2 indicates that males and females represent two different groups. In detail, the large yellow coverage indicates the preference for music among the young segments. The color scheme of the music ratings tells us the difference between them, women are more passionate about listening to music than men, as we can see some blue stands out compared to the visualization of women. Furthermore, as we can see in Figure 4, the red and 100% blue colors are female areas, while males occupy more areas in the green, orange, and yellow zones, which also demonstrates their differences in interests.

Last but not the least, between these different preferences, we detected a correlation between them. We capture some insights from Fig. 5 and Fig. 6. The nodes, connections, and preferences are visualized as a graph. The node size is used to represent the number of edges from the node, and the node colors orange, pink, blue, dark green, and light green are each used to designate the cluster they belong to. The edge thickness is the strength of the connection. Therefore, larger nodes and wider edge thicknesses have more interrelated features. For example, in Figure 5, more friends in life are associated with happiness in life, in other words, more friends in life will increase happiness in life, and vice versa. In short, through clustering, young people formed groups of similar behaviors and some differences existed between male and female music preferences. By creating a network visualization of personality traits and music preferences, the connections between the various traits were also recognized.

VII. Conclusion

As we had suspected, the results from these visualizations provided us with a great number of conclusions. These visualization techniques not only helped us draw conclusions, but have also aided in improving our understanding of the relationships between the features. There are times when analyzing information for the purpose of gaining value and insight can become dull and boring. However, these tools have helped us achieve this goal while being presented with aesthetic visualizations to enhance our learning experience.

References

[1] Pew Research Center, "The Whys and Hows of Generations Research," 2015.

[2] M. Sabo, "Young People Survey," 5 12 2016. [Online]. Available: https://www.kaggle.com/miroslavsabo/young-people-survey. [Accessed 8 12 2021].

[3] P. J. Rentfrow, L. R. Goldberg and D. J. Levitin, "The Structure of Musical Preferences: A Five-Factor Model," *Journal of Personality and Social Psychology,* vol. 100, no. 6, pp. 1139-1157, 2011.

[4] J. Morizot, "Construct Validity of Adolescents' Self-Reported Big Five Personality Traits: Importance of Conceptual Breadth and Initial Validation of a Short Measure," *APA Division 12: Assessment,* vol. 21, no. 5, pp. 580-606, 2014.

[5] J. F. Kruige, P. E. Rauber, R. M. Martins, A. Kerren, S. Kobourov and A. C. Telea, "Graph Layouts by t-SNE," *Computer Graphics Forum,* vol. 36, no. 3, pp. 283-284, 2017.

[6] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions: Mathematical, Physical, and Engineering Sciences,* vol. 374, no. 2065, pp. 1-16, 2016.

[7] A. a. Vathy-Fogarassy and J. Abonyi, Graph-based clustering and data visualization algorithms, London: Springer, 2013.

[8] S. Fortunato and D. Hric, "Community detection in networks: A user guide," *Physics reports,* vol. 659, pp. 1-44, 2016.

[9] M. Jacomy, T. Venturini, S. Heymann, M. Bastian and M. R. Muldoon, "ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software," *PloS one,* vol. 9, no. 6, pp. e98679-e98679, 2014.

[10] B. Shneiderman, "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations.," in *IEEE Symposium on Visual Languages*, Los Alamos, 1996.