Group: Shashank Bharatlal Khede (A20386301)

Navtej Singh Chawla (A20384083)

Prashant Patil (A20356927)

# Title: People survey to analyze preferences, phobias and interests

(Note, if you want to use the same project for ITMD 527, you must mention it as above)

Each group must submit ONLY one copy by a single team member!

**Type of Your Projects:** Data Analysis/Mining

1. **Introduction**

People survey is an interesting collection of data where people were asked questions about their music preferences, movie preferences, their hobbies, phobias, views on life and personal traits, spending habits to analyze that is there any group of people with same interests or likings. People answer these questions on a range of 1 to 5 based on their likings. This survey was done electronically as well as on paper. This survey can be used in recommender systems where a person can be recommended movie or music based on his/her interests. The data file for this survey has around 1010 data with 150 attributes. This survey is anonymous and the person name is not taken so that we can answer multiple questions and take up this as a research project.

1. **Data Sets**

This data set is collected from [www.kaggle.com](file:///E:\downloads\www.kaggle.com). The dataset contains 1010 data where each row has around 150 columns of data. There are multiple variables namely Music preferences, Movie preference, Hobbies & Interests, Phobias, Health habits, Personal traits, opinions and views on life, the spending habits and other demographics. Each of these variables have multiple options like movie preference has further options which are horror, thriller, comedy, romantic, sci-fi, war, animated or action. Each row represents a person’s response to question on the range of 1-5 where 1 is strongly disagree/dislike and 5 is strongly agree/like. The dataset is quite large with the combination of numerical and categorical data along with missing data which makes this data set a challenging one to work on. This is basically a survey data conducted online and on paper both where data helps in analyzing that according to a person’s interest or preference what kind of music he/she likes or what phobia he/she has or do left-handed people have different kind of interests than that of right-handed people. This kind of datasets can be used to answer different questions on personal behaviors, phobias, interests and spending habits. The knowledge out of this data can be implemented in the recommender systems where people will be recommended items based on their interests.

1. **Research Problems**

Below are the research problems that we are going to solve taking this dataset:

1. We don’t have any idea about people who make up same cluster according to this survey so, by the kind of music they listen or movies they see, can we cluster people with same area of interest?
2. Are there any difference in the interest of people according to their gender?
3. Is there a possibility that we can find those people from the survey who cheated and answered questions randomly?
4. Can we find any significant details according to phobia data that what are the phobias that male and female fear and are there any pattern in them? Or is there any significant difference in people having phobia who live in village/city or according to their gender?
5. Can we analyze and find any patterns in missing values? How can we handle missing values to get effective output?
6. What are the attributes of people who save more money than that of people who save lesser money?
7. **Potential Solutions**

Following are the potential solution which we figured out for our problem statements:

1. We will be using k-mean clustering to cluster people with same movie preferences according to music they listen. After doing this we will be able to say people in which cluster watch which genre of movies.
2. Using hypothesis testing for this problem where we can say null hypothesis as there are no differences between their interests and taking alternative hypothesis as there are differences between their interests which depends on they are left or right handed.
3. We decided to use outlier detection method to decide if the point or data is still far away from other user’s data which can show us that this data is just too different from other data.
4. We will use Hypothesis testing to solve this problem where we can say that women fear certain phobia more than that of men to analyze that what is the outcome. Also, we can have null hypothesis as people living in village have chances of having phobia more than that of people living in the city and alternate hypothesis as people living in village have less chances of having phobia than that of people living in city and the same check for the gender as well.
5. This is one of the earliest data mining steps where we fix the missing values because this noisy data may lead to error in accuracy. So, we analyze that are there any patterns in the missing values and how can we fill the gap with some values for proper analysis.
6. We will use Principal component analysis technique to analyze the interests of people according to gender and then decide that what are the areas of interests of male and females.
7. For this problem, we will first check the correlation between all variables and then by PCA algorithm we will get top 3 attributes of people who save more and top 3 attributes of people who save less. In this way, we can identify attributes of people why they couldn’t save more.
8. **Evaluations**

We have multiple problems listed for the dataset which we choose for this project. The evaluation criteria for each of the solution is decided based on solution which gives good results and are appropriately fit for a problem. For example, we have missing values for nominal as well as numeric fields and we need to find a way to fill in the missing values and we have multiple solutions to do that. But we choose to fill in the average value so that we don’t have much deviation in the data.

Similarly, we will be using the clustering to cluster people who like to watch same genre of movies. We are using clustering algorithm to bifurcate data according to different areas. We will get final clusters with people who like watching same kind of movie in same cluster. In the same way, we can cluster people with same music interests.

For Hypothesis testing, we see that the data is large enough where we can use z-test to check if the null hypothesis or alternative hypothesis holds true.

We will be using the clustering methods which gives best results and are good to do analyses of the data sets. We use the k-means clustering method to cluster the user data so that we can group by music type or movie genre. Clustering gives us different groups where each group contains same type of data. We have hierarchical clustering method also which we are not using as of now because it is used when we want to see the classification or categorization of data.

We will be using PCA to interests of people according to gender. We will be able to differentiate interests of male and females. Also, we will be able to find the attributes of people who save more money and attributes of people who save less money. The spending habits tells us where they spent more money.

1. **Expected Outcomes**

We can expect below outcomes from our project after successful completion:

1. We will be able to list out clusters having people with same kind of interest.
2. We will be able to distinguish between the interests of right-handed and left-handed people.
3. We will be able to predict user data based on his other data. Prediction can be made from the user data that if a person likes romantic songs, he may like love fiction kind of movies.
4. People who cheated while surveying can be listed out.
5. Pattern or significance in the phobia phenomena can be jotted down and we will be able to differentiate between male and female data along with the place they live.
6. After analyzing the missing value pattern, we will be able to fill in the values with some constant value to reduce noisy data.
7. We will be able to say what are the difference in areas of interest in male and female.
8. We will be able to list top three attributes of people who save more and top 3 attributes of people who save less.