Indiana University Bloomington

CSCI B 565

Data Mining

Young People Survey:

Exploring the Preferences, Interests, Habits, opinions and fears of young people

---

Authors:
Meghana Kantharaj
Arpitha N Kashyap

Professor:
Dr. Christopher Raphael

April 27th, 2017

ABSTRACT

The data set we have chosen is Young people survey dataset from Kaggle. We are concentrating on hobbies and interests' data present in the data set.

The goal of this project is to predict the gender based on hobbies and interests of people.

We use predictive modelling, i.e. we try to build a model that will help us predict the gender based on hobbies and interests. We first build a model and then classify the data. We run an algorithm to perform dimensionality reduction. This will help us describe a large number of human interests by a smaller number of latent concepts. This will help to provide a higher accuracy to improve the prediction model.

INTRODUCTION

In 2013, FSEV UK conducted a survey. They asked the students of the Statistics class to invite their friends to participate in this survey.

- The data file consists of 1010 rows and 150 columns (139-integer and 11-categorical).
- For convenience, the original variable names were shortened in the data file.
- The data consisted of missing values.
- The survey was presented in two forms-electronic and written.
- The original questionnaire was in Slovak language and was later translated into English.
- The participants in this survey were all of Slovakian nationality.
- All participants were aged between 15-30.

The variables can be split into the following groups:

- Music preferences
- Movie preferences
- Hobbies & interests
- Phobias
- Health habits
- Personality traits, views on life, & opinions
- Spending habits
- Demographics

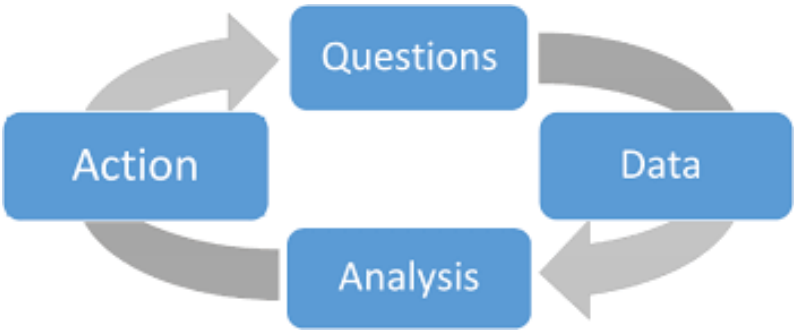
The questionnaire under hobbies and interests included the following:

A rating scale from 1-5 is used to rate each hobby/interest.

HOBBIES & INTERESTS

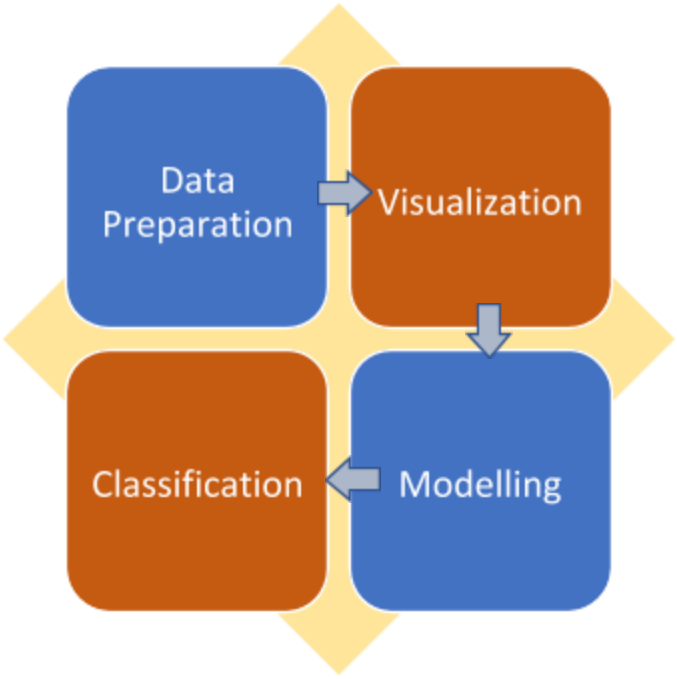
1. **History**: Not interested 1-2-3-4-5 Very interested (integer)
2. **Psychology**: Not interested 1-2-3-4-5 Very interested (integer)
3. **Politics**: Not interested 1-2-3-4-5 Very interested (integer)
4. **Mathematics**: Not interested 1-2-3-4-5 Very interested (integer)
5. **Physics**: Not interested 1-2-3-4-5 Very interested (integer)
6. **Internet**: Not interested 1-2-3-4-5 Very interested (integer)
7. **PC Software, Hardware**: Not interested 1-2-3-4-5 Very interested (integer)
8. **Economy, Management**: Not interested 1-2-3-4-5 Very interested (integer)
9. **Biology**: Not interested 1-2-3-4-5 Very interested (integer)
10. **Chemistry**: Not interested 1-2-3-4-5 Very interested (integer)
11. **Poetry reading**: Not interested 1-2-3-4-5 Very interested (integer)

12. **Geography**: Not interested 1-2-3-4-5 Very interested (integer)
13. **Foreign languages**: Not interested 1-2-3-4-5 Very interested (integer)
14. **Medicine**: Not interested 1-2-3-4-5 Very interested (integer)
15. **Law**: Not interested 1-2-3-4-5 Very interested (integer)
16. **Cars**: Not interested 1-2-3-4-5 Very interested (integer)
17. **Art**: Not interested 1-2-3-4-5 Very interested (integer)
18. **Religion**: Not interested 1-2-3-4-5 Very interested (integer)
19. **Outdoor activities**: Not interested 1-2-3-4-5 Very interested (integer)
20. **Dancing**: Not interested 1-2-3-4-5 Very interested (integer)
21. **Playing musical instruments**: Not interested 1-2-3-4-5 Very interested (integer)
22. **Poetry writing**: Not interested 1-2-3-4-5 Very interested (integer)
23. **Sport and leisure activities**: Not interested 1-2-3-4-5 Very interested (integer)
24. **Sport at competitive level**: Not interested 1-2-3-4-5 Very interested (integer)
25. **Gardening**: Not interested 1-2-3-4-5 Very interested (integer)
26. **Celebrity lifestyle**: Not interested 1-2-3-4-5 Very interested (integer)
27. **Shopping**: Not interested 1-2-3-4-5 Very interested (integer)
28. **Science and technology**: Not interested 1-2-3-4-5 Very interested (integer)
29. **Theatre**: Not interested 1-2-3-4-5 Very interested (integer)
30. **Socializing**: Not interested 1-2-3-4-5 Very interested (integer)
31. **Adrenaline sports**: Not interested 1-2-3-4-5 Very interested (integer)
32. **Pets**: Not interested 1-2-3-4-5 Very interested (integer)



Model

Steps involved:



Data

The file with the data is called responses.csv

The dimensions are
> dim(responses)
1010 150

The variable hni corresponds to the hobbies and interests data.

>hni = responses[, 32:63]
This data is present in columns 32 to 63 in dataset
Gender is column 145.

We divide the data into training and testing set.

>dim(train)
707 33
>dim(test)
304 33

1.Data Preprocessing / Preparation of the data:

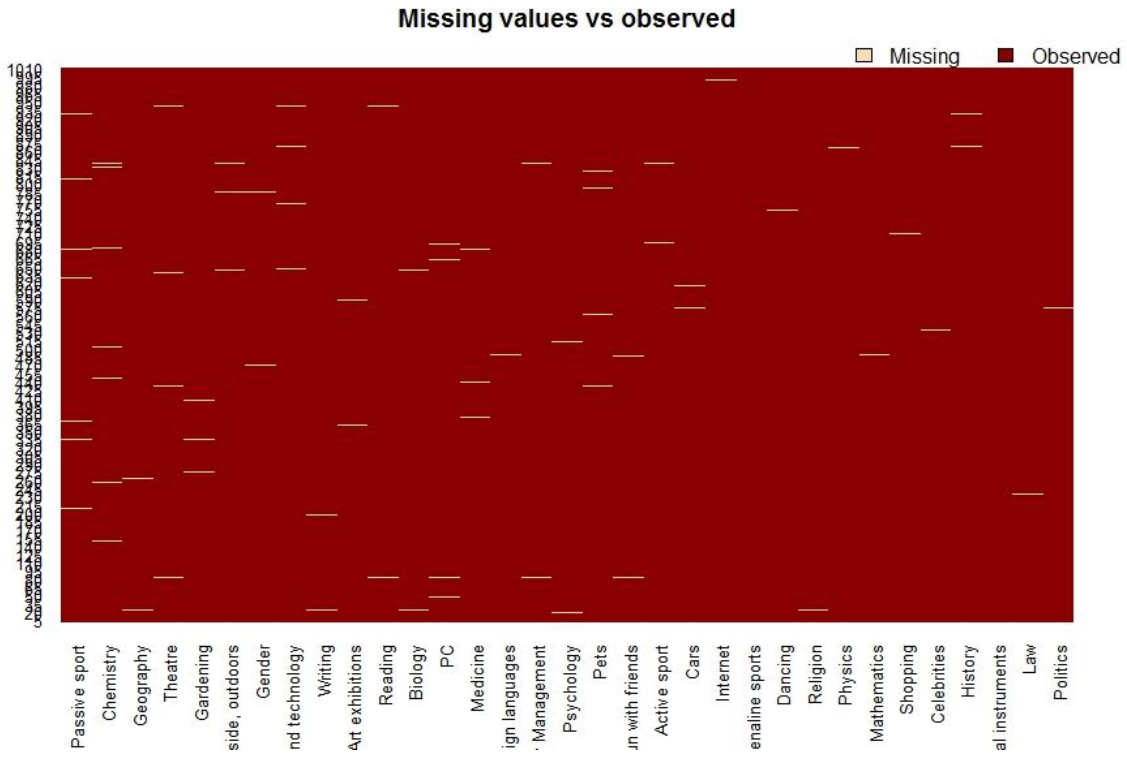
In the Responses dataset, we have considered the category of hobbies and interests subset, which has columns as mentioned above. These are of the type Ordinal. the variable to be predicted is Gender, which is nominal. We use Gender as a single discrete variable that we want to classify from the other discrete data.

On performing SVD on the hobbies and interests dataset, we get the following eigenvalues-

```
445.71591  66.42560  58.23084  55.25723  44.85127  42.44559  40.30508  38.61321  37.01056  34.80744
 34.23462  32.66807  32.15155  31.67657  29.65653  29.15989  28.84355  27.91973  27.31059  27.05449
 25.25131  24.67815  24.27622  23.07587  22.28519  21.95095  21.08346  19.97164  19.48698  18.85617
 17.88118  17.11456
```

From the above data, we see that all dimensions contribute to the predicted variable. Thus, dimensions cannot be reduced in the given dataset.

We first visualized missing data in the dataset, the visualization can be seen below-



The plot shows the approximate indices of missing data in their respective columns, where the indices are along y axis and the distinct columns are along x axis. For example, the politics column, represented on the right end of the plot has one missing

value as seen somewhere around 500 index.

On querying, we see that the missing value is at index 438.

```
> which(is.na(data[,3]))
[1] 438
```

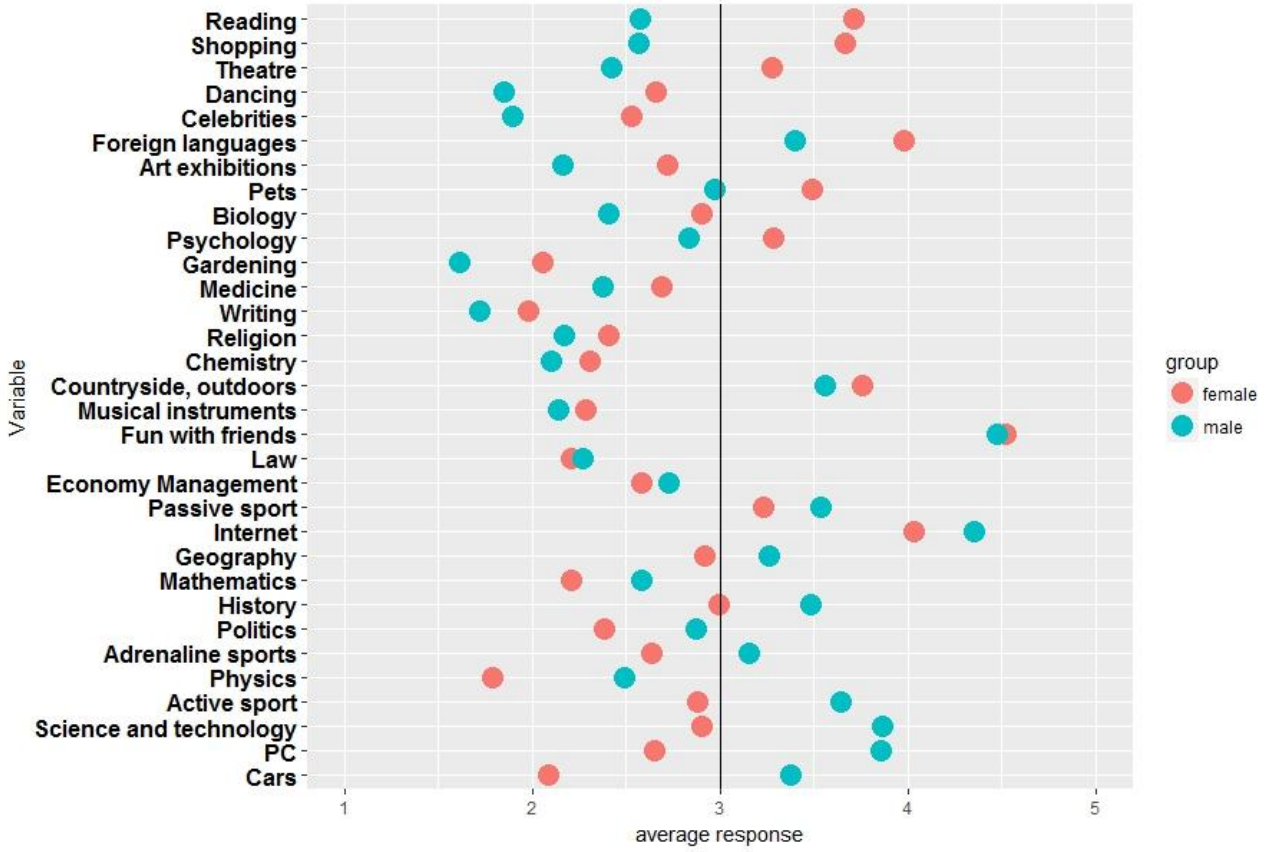
The following missing count was obtained for the columns. There were not too many missing values.

History	Psychology	Politics	Mathematics
2	5	1	3
Physics	Internet	PC	Economy Management
3	4	6	5
Biology	Chemistry	Reading	Geography
6	10	6	9
Foreign languages	Medicine	Law	Cars
5	5	1	4
Art exhibitions	Religion	Countryside, outdoors	Dancing
6	3	7	3
Musical instruments	Writing	Passive sport	Active sport
1	6	15	4
Gardening	Celebrities	Shopping	Science and technology
7	2	2	6
Theatre	Fun with friends	Adrenaline sports	Pets
8	4	3	4
Gender			
6			

We replace the missing data with the mean of the columns for further computation and model fitting.

2.Visualization:

The data was visualized with a scatter plot where the averages of interest ratings of each topic of men and women were plotted for a visual comparison. As seen below, a majority of columns are visually distinguishable averages for men and women, with minimal overlaps. Women seem more interested in artistic fields such as theatre, dancing, art whereas men seem more interested in sciences and math.



But the averages were not distinct for fields such as sports and entertainment.
These differences in opinion help us predict the gender of the person based on their ratings of the following fields of interests.

3.Model Training:

We use the logistic regression model to train the data. Logistic regression is a predictive analysis. It is used to describe data and explain the relationship between one dependent binary variable and one or more nominal, interval, ratio-level or ordinal independent variables (in this case it is ordinal).

Although binomial distribution is most commonly used in logistic regression, we use the Gaussian distribution as it suits our data well.

Major assumptions of logistic regression on our data:

1. The outcome is discrete. The dependent variable(Gender) is dichotomous in nature.
2. There are no outliers in the data. This can be assessed by converting the continuous predictors to standardized. This can also be done by remove values below -3.29 or greater than 3.29 for z.
3. There is no high intercorrelations among any of the predictors.
4. The dependent variable is a stochastic event.
5. There is a linear relationship between any continuous independent variables and the logit transformation of the dependent variable.
6. There is independence of observations and the dependent variable has mutually exclusive and exhaustive categories.
7. The average summarizes the population fairly well.

Summarizing the fit and interpreting what the model is telling us:

First of all, there are some variables that are not statistically significant. This can be seen from the significance codes mentioned below. As for the statistically significant variables, cars has the lowest p-value suggesting a strong association of this variable with predicting the gender.

```
Call:
glm(formula = train$Gender ~ ., family = gaussian(), data = train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.00310  -0.22065   0.01836   0.22285   0.87955

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               0.5666165  0.1125178   5.036 6.12e-07 ***
History                  -0.0465465  0.0119692  -3.889 0.000111 ***
Psychology                0.0355105  0.0113235   3.136 0.001787 **
Politics                 -0.0279524  0.0122071  -2.290 0.022338 *
Mathematics               0.0125990  0.0119550   1.054 0.292323
Physics                  -0.0326900  0.0143617  -2.276 0.023147 *
Internet                  0.0029889  0.0162146   0.184 0.853806
PC                       -0.0618342  0.0124650  -4.961 8.90e-07 ***
`Economy Management`     -0.0003657  0.0109451  -0.033 0.973358
Biology                   0.0029365  0.0146295   0.201 0.840975
Chemistry                 0.0019701  0.0137276   0.144 0.885928
Reading                   0.0547764  0.0108476   5.050 5.70e-07 ***
Geography                -0.0045629  0.0109286  -0.418 0.676435
`Foreign languages`       0.0473979  0.0128084   3.701 0.000233 ***
Medicine                  0.0130359  0.0134677   0.968 0.333423
Law                       0.0192710  0.0120767   1.596 0.111021
Cars                     -0.0626775  0.0104792  -5.981 3.60e-09 ***
`Art exhibitions`         0.0040491  0.0120798   0.335 0.737581
Religion                 -0.0124328  0.0103553  -1.201 0.230321
`Countryside, outdoors`   0.0080570  0.0123848   0.651 0.515557
Dancing                   0.0480064  0.0100931   4.756 2.41e-06 ***
`Musical instruments`     0.0009578  0.0094133   0.102 0.918990
Writing                  -0.0345344  0.0117248  -2.945 0.003336 **
`Passive sport`          -0.0110686  0.0092835  -1.192 0.233567
`Active sport`           -0.0469069  0.0094879  -4.944 9.68e-07 ***
Gardening                 0.0237933  0.0121377   1.960 0.050375 .
Celebrities               0.0377814  0.0117054   3.228 0.001308 **
Shopping                  0.0676093  0.0124024   5.451 7.02e-08 ***
`Science and technology` -0.0464817  0.0117363  -3.961 8.27e-05 ***
Theatre                   0.0386573  0.0119749   3.228 0.001306 **
`Fun with friends`       -0.0356941  0.0177216  -2.014 0.044390 *
`Adrenaline sports`       0.0028736  0.0101947   0.282 0.778129
Pets                      0.0208243  0.0084241   2.472 0.013682 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.1039025)

    Null deviance: 168.03  on 706  degrees of freedom
Residual deviance:  70.03  on 674  degrees of freedom
AIC: 439.72

Number of Fisher Scoring iterations: 2
```

On performing ANOVA test, we get the following results-

History	1	4.5503	705	163.484
Psychology	1	7.4148	704	156.069
Politics	1	3.3969	703	152.672
Mathematics	1	5.1495	702	147.522
Physics	1	10.5999	701	136.923
Internet	1	1.5241	700	135.398
PC	1	16.4926	699	118.906
`Economy Management`	1	0.0085	698	118.897
Biology	1	3.7523	697	115.145
Chemistry	1	0.0595	696	115.086
Reading	1	13.5044	695	101.581
Geography	1	0.0905	694	101.491
`Foreign languages`	1	2.9119	693	98.579
Medicine	1	0.0510	692	98.528
Law	1	0.5613	691	97.966
Cars	1	5.5761	690	92.390
`Art exhibitions`	1	0.6519	689	91.738
Religion	1	0.0463	688	91.692
`Countryside, outdoors`	1	0.4634	687	91.229
Dancing	1	4.5848	686	86.644
`Musical instruments`	1	0.1555	685	86.488
Writing	1	0.8922	684	85.596
`Passive sport`	1	0.1696	683	85.427
`Active sport`	1	3.2218	682	82.205
Gardening	1	1.1730	681	81.032
Celebrities	1	4.2333	680	76.798
Shopping	1	3.4113	679	73.387
`Science and technology`	1	1.3782	678	72.009
Theatre	1	0.9631	677	71.046
`Fun with friends`	1	0.3634	676	70.682
`Adrenaline sports`	1	0.0171	675	70.665
Pets	1	0.6349	674	70.030

The difference between the null deviance and the residual deviance shows how our model is doing against the null model (a model with only the intercept). The wider this gap, the better it is. Analyzing the table we can see the drop in deviance when adding each variable one at a time.

Here, a large p-value indicates that the model without the variable explains more or less the same amount of variation. Ultimately we would like to see is a significant drop in deviance.

4. Classification

We tested the model by classifying new data which the model had not seen during the training stage. This data was stored in the variable test. The ratio used for training and testing was 70:30. Since we had around 1010 rows, around 300 rows were used for testing. On testing, we saw around 83% efficiency, which is acceptable since the dataset has a very small number of rows.

RESULTS

Once the classification is complete and we have fit the data, the accuracy is found to be 83.22%. This is a fairly good accuracy since we have considered 30% of the entire data for testing. Also, since the data distribution is normal, the gaussian function fit our model better than the other functions such as binomial, poisson, etc, which were slightly less efficient than the gaussian model, with the binomial yielding around 81% accuracy and the poisson yielding around 80% accuracy. This tells us that the model as well as the methods used are efficient and hence the prediction based on the given data would also be efficient.

CONCLUSION

The study or analysis we have done shows that women are more imaginative, creative or artistic compared to men. Hence most women incline towards more creative hobbies or interests like shopping, dancing, psychology field etc. while men are more inclined toward sports, mathematics, physics, cars and so on. We can predict the gender of a person based on their rating of the categories of common hobbies with an accuracy of around 83%.