**How do you know who likes Alcohol? A Factor Analysis Study on Survey Data**

Team D10: Kumara Adavelly, Sijia Wu, Jing Zhou, Nowed Patwary

MKT 591 Spring 2020   
MSBA - Cohort D

**Executive Summary**

Survey datasets are often useful for doing many kinds of exploratory data analysis (EDA) in marketing, as they can quickly collect lots of data on a short document. The dataset we used had 150 different columns or attributes that people were asked to rate on a Likert Scale. A good way to deal with this much data is to use Factor Analysis to reduce the dimensionality and make it easier to understand that data. Once we have created the new factors, we used them in a logit regression to see how well they predict someone’s likelihood of drinking alcohol. We did this 3 times, extracting more important features based on significance until we reached 20 factors from the survey reduced to 7 factors based on significance. Our final prediction score for the logit function was 0.775, meaning we could predict someone’s liking to drink alcohol based on these factors with a 77.5%.

**Introduction/background**

In 2013, students of the Statistics class at Fakulta sociálnych a ekonomických vied (FSEV) UK were asked to invite their friends to participate in a survey which explores the preferences, interests, habits, opinions, and fears of young people. The survey was presented to participants in both electronic and written form. The original questionnaire was in Slovak language and was later translated into English. A total of 1010 participants responded to the survey and all of them were of Slovakian nationality, aged between 15-30. We found this data in Kaggle. Although there are some missing values, the data itself is well structured. As this is a large dataset, we have decided to focus on the variables of our interests. We would like to explore the relationship between alcohol preference and a wide range of factors related to music, movies, socializing, hobbies and interests. The objective of this project is to find factors that relate well to each other and can be used to build a model for predicting alcohol preference with logit model. We will try to more accurately predict someone's propensity for drinking alcohol based on the other attributes in the survey, utilizing factor analysis.

By utilizing this method, we can find strong correlations between distinct factors, and condense them into groups of factors that reduce dimensionality and work together to more accurately make predictions. These are also called latent factors, in that they cannot be directly measured, only inferred through other factors that seem to work together. Once we do the Factor Analysis, the latent factors are what group them together and allow us to make simpler and robust models, like the logit function. We believe that the findings would help us to better understand consumer behaviours and provide some useful insights for decision making in marketing.  **Data and methodology description**

**1. Data description**

The original dataset has two csv file-response and columns, the response file consists of 1010 rows and 150 columns (139 numeric and 11 categorical). For our analysis, we picked columns of 74 independent variables, including music preferences, movie preference, hobbies & interests, phobias and healthy eating, also picked Alcohol column as dependent variable, so our own dataset has 75 columns,797 rows after cleaning the data. For convenience, the original variable names were shortened in the data file, so the columns file helps us to match the data with the original names.

The questionnaire collected the answers based on the Likert Scale, a widely used scale in survey research. The 1-2-3-4-5, 5 integers, which presents different degrees of agreement for the question for each column, for example: Country: Don't enjoy at all 1-2-3-4-5 Enjoy very much. And Alcohol is a categorical variable, it is measured by “Never - Social drinker - Drink a lot”. We binned “Never “and “Social Drinker” responses as one group and “drinks a lot” as another group represented by 1 in the column.

**2. Methodology description**

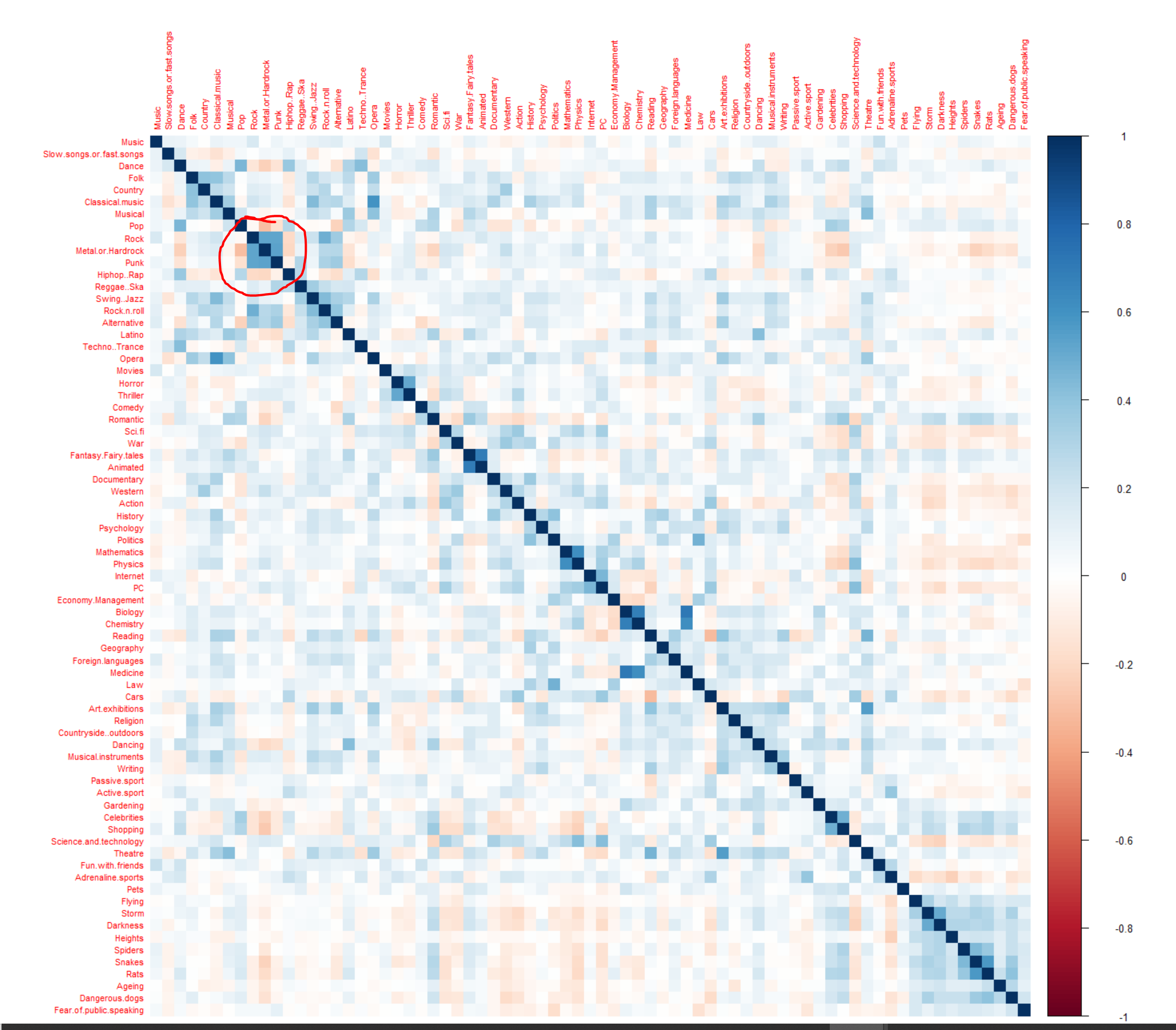
For the methodology part, we used three main methods to start our report: coefficient correlation, dimension reduction and logic regression. Firstly, we preprocessed the data with missing value and performed a correlation matrix plot of independent variables to check if there are enough levels of correlation present between variables to start factor analysis. Then we started the dimensionality reduction to reduce all independent variables to fewer, more manageable and interpretable factors. Using the factors in the logit regression model, we were able to generate results within the same ballpark of the original model in terms of Accuracy, while benefiting from dimensionality reduction.

|  |  |  |
| --- | --- | --- |
| **Model 1**  We first performed factor analysis on 74 columns and used the factors derived from the analysis in the logit regression model to predict whether a person is alcoholic or not. | **Model 2**  In model 2 we only kept the significant factors from the factor analysis and then performed a logit regression to predict whether a person is alcoholic or not. | **Model 3 (Original model)**  General model with all the 74 variables used in a logit regression to predict whether a person is alcoholic or not. |

Our intention was to take advantage of the correlation matrix, and factor analysis concepts to reduce the dimensionality of the original dataset yet keep the explainability and performance of the logit model.

**Data Analysis and Interpretation**

After factor analysis, we were left with 20 factors , with each factor explaining the variability of one or several variables in the dataset. For instance, Based on the loading values, Factor 1 primarily explains the variability of Biology, Chemistry, and Medicine columns. These columns measure a person's interest in those respective subjects. The 20 factors together explained 46% of the variance from the original dataset, which can be considered a decent result.

Significant Factors and the variables they explain the most are listed below:

Factor 1: Folk, Classical

Factor 3: Biology, Chemistry, Medicine

Factor 4: Metal, Hardrock, Punk

Factor 6: Law, Politics

Factor 7: Metal, Hardrock, Punk

Factor 12: Rock and Roll

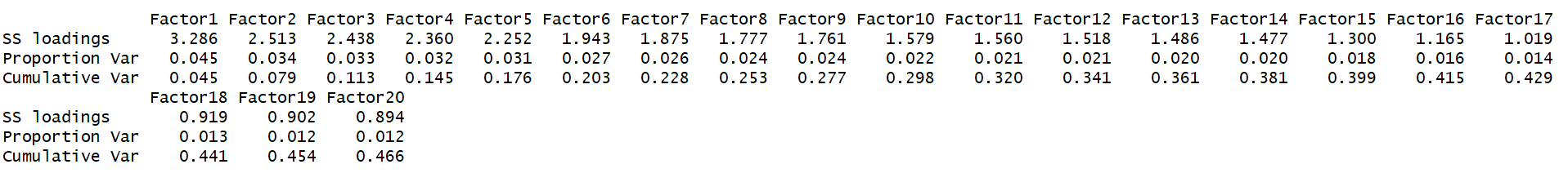
Factor 13: Thriller

Factor 16: mixed

Factor 17: Reggae aka ska

Figure 1.0 Correlation Plot of all 73 factors

Based on the results, some of the factors capture the variability of columns with a high correlation. For instance, Factor 7 best explains the variability in the column’s Punk, Hardrock, & Metal, which are highly correlated in the correlation matrix. The variance explained by the factors also follows the order listed above with Factor 1 explaining 4.5% of the variability and Factor 20 explaining 1.2% of the variability of the original model, as can be seen in the figure below.



|  |  |  |  |
| --- | --- | --- | --- |
|  | Model 1(All factors) | **Model 2 (Only Significant factors)** | Model 3 (Original Model) |
| AIC score | 812.86 | **820.86** | 849.05 |
| Accuracy | 0.772 | **0.774** | 0.782 |

Figure 2.0: Model Accuracy and AIC scores

**Actionable Marketing**

As shown in Figure 3.0 above, the model maintained a steady performance through it’s 3 iterations, indicating to us that we created a fairly robust model. This also means that the Factor Analysis worked in choosing good factors to bring together to reduce dimensionality. We also see this in the AIC scores which are not that different from one another. Although Model 1 seems to have the best AIC score, it lacks accuracy compared to other models, On the other hand, Model 3 utilizes all the 74 attributes to get slightly better accuracy. We considered Model 2 as our base model, given the dimensional reduction obtained and retainment of accuracy compared to the original model.

**Model 2: logit Regression output and interpretation**

Significant factors and their effect on the log odds ratio

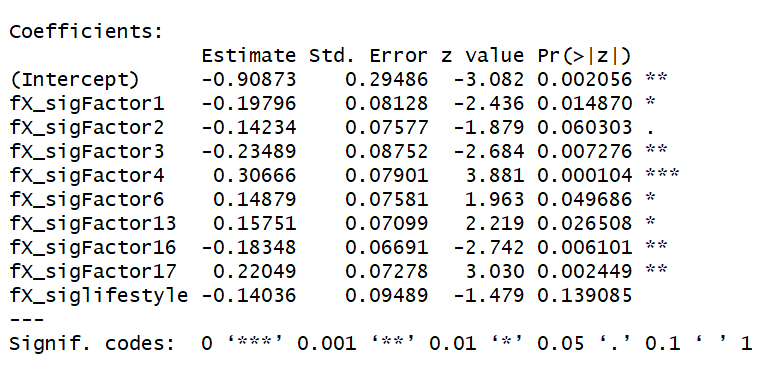
**Positive effect: Factors 4,6,13,16.** These factors together best explain the variability for Metal, Hardrock, Punk, Law, Politics, and Thriller columns.

Therefore, based on the output, an alcoholic beverage manufacturer should be targeting groups who enjoy Metal, Hardrock, Punk music genres, who are interested in law and politics, and love thriller movies.

**Negative effect: Factors 1,3,16.** These factors together best explain the variability for Folk, Classical, Biology, Chemistry, Medicine, and Thriller movies

The negative effect factors highlight that people in the field of Biology, Chemistry, or Medicine don’t like alcohol as much and the same goes with people who like Folk, and Classical music genres. This result can allow marketers to not target these groups.

We did try adding an ‘Eating healthy” after doing the Factor Analysis, to see if we could further improve the model. Our theory was that maybe people who eat healthy don’t drink, but alas, it was not significant, and so we did not keep it.

****

So we wanted to find out if we can trim the number of attributes down and maintain similar scores. We believe we achieved that with Model 2 was based on the logit regression for Model 1.

|  |  |  |
| --- | --- | --- |
| Figure 3.1 (Model 1) | **Figure 3.2 (Model 2)** | Figure 3.2 (Model 3) |

Both of our second and third models have lower AIC scores than the original model Therefore we can say that using our method we were able to create and improve a model. The implications of this model is that we have some unique factors that can be used to identify people who like to drink alcohol. This has many implications for marketing, advertising; if we were affiliated with a firm or company we could recommend ways of marketing alcohol based on these factors.