An Explorotory Analysis of Young People

Ryan Pugh. Aspen Bass, Zongzhe Lin, Joseph Saah

October 17, 2017

## Introduction

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

The data file (responses.csv) consists of 1010 rows and 150 columns (139 integer and 11 categorical). For convenience, the original variable names were shortened in the data file. See the columns.csv file if you want to match the data with the original names. The data contain missing values. 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. All participants were of Slovakian nationality, aged between 15-30.

The variables can be split into the following groups: *Music preferences (19 items)* Movie preferences (12 items) *Hobbies & interests (32 items)* Phobias (10 items) *Health habits (3 items)* Personality traits, views on life, & opinions (57 items) *Spending habits (7 items)* Demographics (10 items)

The groups relevant to our analysis are Music preferences, Movie preferences, and Demographics. This analysis is designed to show the relationship between media preferences and hand dominance, and if there are any correlations in the preferences for different media types. More specifically, the data analysis will answer the following questions:

1. Are there differences in music preferences between left handed people and right handed people?
2. Are there any difference in movie preference between men and women?

...

## 1. Data Import and Cleaning

youthdata2<-select(youthdata, Music:Action, Age:`Left - right handed`, -`Only child`)  
youthdata2<-gather(data = youthdata2, key = "MusicGenre", value = "MusicGenreScore", Dance:Opera)  
youthdata2<-gather(data = youthdata2, key = "MovieGenre", value = "MovieGenreScore", Horror:Action)  
youthdata2<-select(youthdata2, MusicScore=Music, MusicGenre, MusicGenreScore, MoviesScore=Movies, MovieGenre, MovieGenreScore, Age, Height, Weight, Gender, DominateHand=`Left - right handed`)  
youthdata2

## # A tibble: 188,870 x 11  
## MusicScore MusicGenre MusicGenreScore MoviesScore MovieGenre  
## <int> <chr> <int> <int> <chr>  
## 1 5 Dance 2 5 Horror  
## 2 4 Dance 2 5 Horror  
## 3 5 Dance 2 5 Horror  
## 4 5 Dance 2 5 Horror  
## 5 5 Dance 4 5 Horror  
## 6 5 Dance 2 5 Horror  
## 7 5 Dance 5 4 Horror  
## 8 5 Dance 3 5 Horror  
## 9 5 Dance 3 5 Horror  
## 10 5 Dance 2 5 Horror  
## # ... with 188,860 more rows, and 6 more variables: MovieGenreScore <int>,  
## # Age <int>, Height <int>, Weight <int>, Gender <chr>,  
## # DominateHand <chr>

We downloaded the data CSV and imported it into a table. The dataset had a lot of irrelevant variables to our analysis, so first we selected the relevant groups. To tidy the data, we took all the genre obserevations for Music and Movies and gathered them into the MusicGenre and MovieGenre variables respectively, with each having their values being outputted to a GenreScore. Lastly, we renamed some variables to make them more code friendly.

In are tidy dataset there are 188870 rows, and 11 columns. 7 are numerical, 4 are categorical.

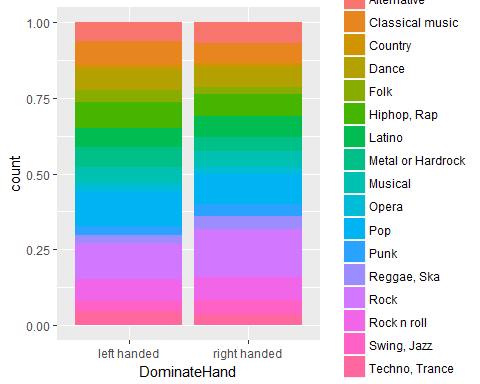
## 2. Data Analysis

Each question has its own data transformation done within the visualization code block. General transformation was completed in the Importing and Cleaning phase. Relevant code can be viewed above.

### Question 1: Are there differences in music preferences between left handed people and right handed people?

#### 1a: Genre Preference comparison for Dominate Hand

filter(youthdata2, MusicGenreScore == 5, !is.na(DominateHand)) %>%   
 ggplot() +  
 geom\_bar(mapping = aes(x = DominateHand, fill = MusicGenre), position = "fill")



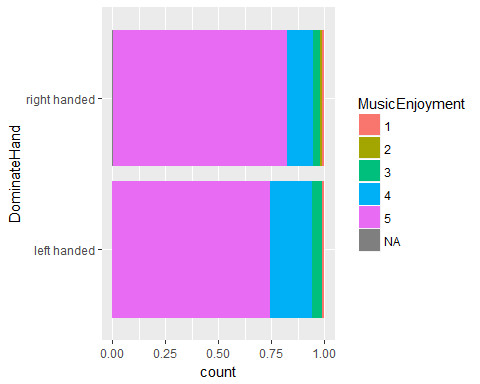
For this graph, we wanted to filter all the music scores (how high people rated the music) to 5. This is the highest score, so we assume that the people who scored a genre with a 5 consider that their favorite genre.

After filtering the score, we wanted to see if there was a relationship between dominate hand and favorite music. To view this more neatly, we put the dominate hand variable on the x axis and the number of people in the survey on the y axis. Next, we organized the the music genre by color so we could see the split between how many left-handers liked a certain genre of music vs. how many right-handers liked a certain genre of music.

According to the graph, a larger number of left-handers preferred Pop over right-handers, but a larger number of right-handers preferred Rock over left-handers. Overall, there is not much of a discrepensy between dominate hand and preferred music. People of all types of hands like all the same types of music.

#### 1b: Dominate hand and general score of music enjoyment

filter(youthdata2, !is.na(DominateHand)) %>%  
mutate(MusicEnjoyment = cut(MusicScore, c(-Inf,1,2,3,4,Inf), labels=c("1","2","3","4","5"))) %>%   
 ggplot() +  
 geom\_bar(mapping = aes(x = DominateHand, fill = MusicEnjoyment), position="fill")+  
 coord\_flip()

 ...

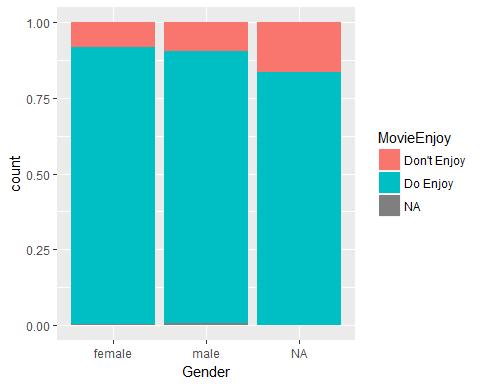
For this graph, we wanted to filter out all the missing Dominate Hand values, and converted the MusicScore data (how high people rated the music) into categorical data. This enables us to show proportions in the bar graph.

There is not much of a difference between music preferences of left handed and right handed people. Right handed people and left handed people seem to share similar music preference. However, in this dataset right handed people are the majority of the dataset.

### Question 2: Are there any difference in movie preference between men and women?

#### 2a: Movie Enjoyment by Gender

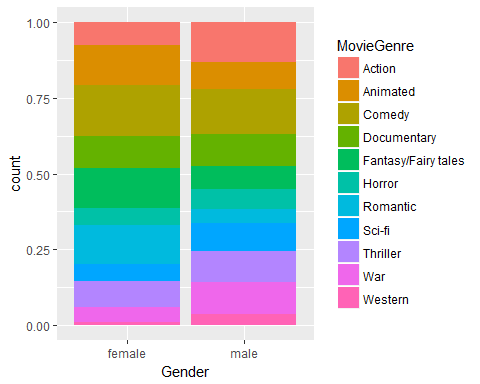
mutate(youthdata2,   
 MovieEnjoy = cut(MoviesScore, c(-Inf,3,Inf), labels=c("Don't Enjoy","Do Enjoy"))) %>%   
 ggplot() +  
 geom\_bar(mapping = aes(x = Gender, fill = MovieEnjoy), position="fill")



We transformed the dataset to contain a new categorical variable "MovieEnjoy", where we set the standard score of "enjoy watching movie" to be at least a 4, and created a bar chart displaying the proportion. The bar chart demonstrates that women are slightly more likely to enjoy watching movies in general.

#### 2b: Genre Preference by Gender

filter(youthdata2, MovieGenreScore > 3, !is.na(Gender)) %>%   
 ggplot() +  
 geom\_bar(mapping = aes(x = Gender, fill = MovieGenre), position = "fill")



Here we also transformed the dataset to set the standard score of "enjoy watching movie" to be at least a 4, and created a bar chart displaying the proportion, using the fill function to represent different genres by color. Female's preference are focused more on comedy, fantasy and romantic, while male's preference are focused more on War, Sci-fi, and Action.

## 3. Summary

Reflect on the exploratory analysis:

*(i) Summary*

Neither dominate hand nor gender have a strong influence on general media preference. It was interesting that righthanders were more inclined to rock and that left handers were more inclined to pop, but the difference may not be significant. It was also interesting that genre proportions for gender seem to reflect some of those genre's stereotypes, such as a greater proportion of men liking action and war and a greater proportion of women liking romance and fantasy. However, tests were not conducted to see if the difference was significant.

*(ii) Analysis Limitations*

The survey was a convenience sample liable to voluntary response bias, though it is of large size. A rating from 1 to 5 doesn't give a very good spread.

*(iii) Future Analysis*

It would be useful to analyze if Movie Score and Music Score have any correlation. Analysis could be done to find associations between movie genre preference and music genre preference.