Association Rules

Noah Giebink and Sebastian Deimen

March 3, 2020

Discretize variables

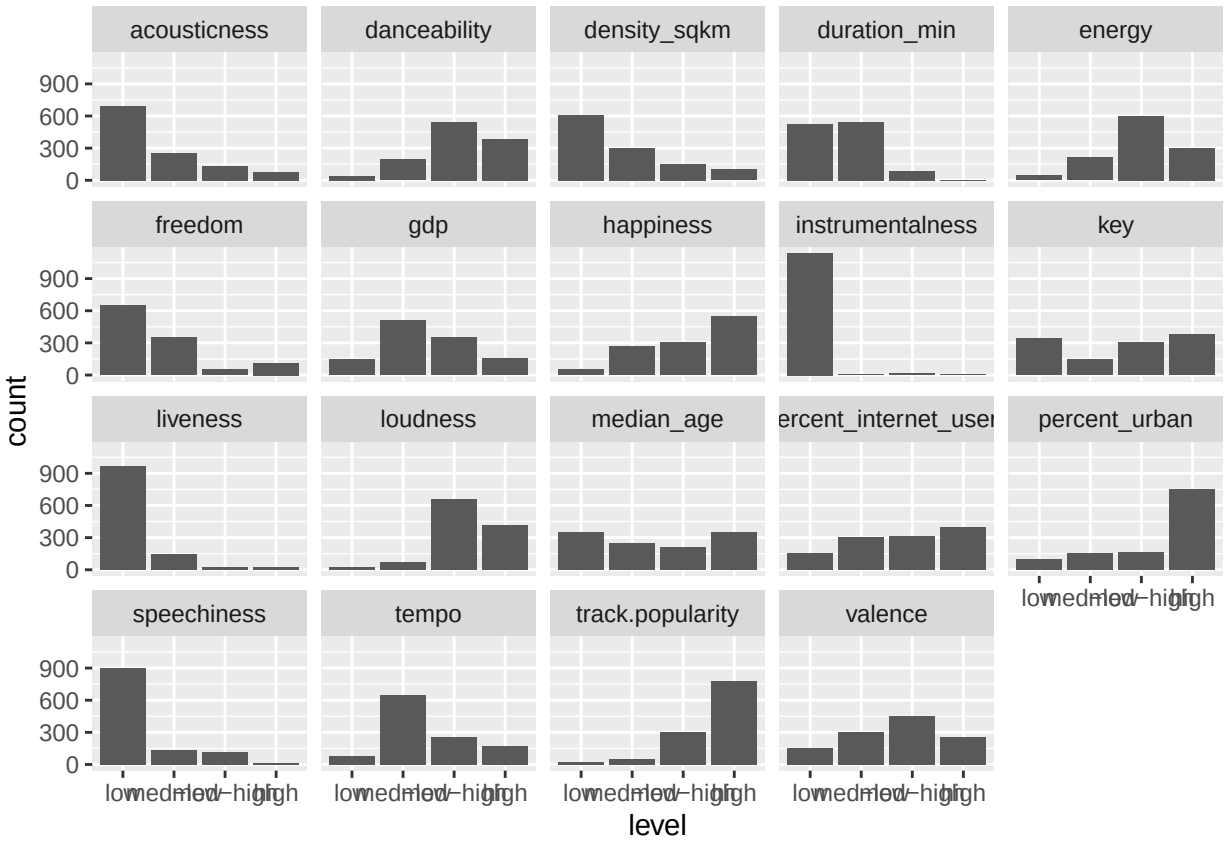
track information variables: track.name, track.popularity, *audio metrics:* danceability, energy, key, loudness, speechiness, acousticness, instrumentalness, liveness, valence, tempo, duration_min, *sociopolitical variables:* country, happiness, median_age, percent_urban, percent_internet_users, density_sqkm, freedom, gdp

```
# select subset of variables relevant to question
spot <- spot %>% select(track.name, track.popularity,
                        danceability, energy, key, loudness,
                        speechiness, acousticness, instrumentalness,
                        liveness, valence, tempo, duration_min,
                        country, happiness, median_age, percent_urban,
                        percent_internet_users, density_sqkm,
                        freedom, gdp)

# Discretize variables
  # which need it? everything except name, country
chr_df <- select(spot, track.name, country)
chr_df$track.name <-factor(chr_df$track.name)
chr_df$country <- factor(chr_df$country)
dbl_df <- select(spot, -track.name, -country)

  # function to discretize variables
disc <- function(x){
  cut(x, breaks = 4,
      labels = c('low', 'med-low', 'med-high', 'high'))}
  # apply disc fun to all dbl vars
dbl_df <- mutate_all(dbl_df, funs(disc))
  # bind data frame back together by cols
spot <- cbind(chr_df, dbl_df)

# plot distribution of levels for each variable
dbl_long <- pivot_longer(dbl_df, cols = colnames(dbl_df),
                         names_to = 'variable', values_to = 'level')
ggplot(dbl_long, aes(level))+
  geom_bar()+
  facet_wrap(~variable)
```



Most variables have a decent spread of values after discretization, except for instrumentalness, liveness, and speechiness. Since we think this is due to their irrelevance to the top 50 tracks, we chose to omit these variables from association rule mining.

Our remaining variables are the following:

```
# The remaining dataset
spot <- select(spot, -instrumentalness, -liveness, -speechiness)
variable.names(spot)
```

```
##  [1] "track.name"             "country"
##  [3] "track.popularity"       "danceability"
##  [5] "energy"                 "key"
##  [7] "loudness"               "acousticness"
##  [9] "valence"                "tempo"
## [11] "duration_min"           "happiness"
## [13] "median_age"             "percent_urban"
## [15] "percent_internet_users" "density_sqkm"
## [17] "freedom"                "gdp"
```

Make transactional database

Inspect

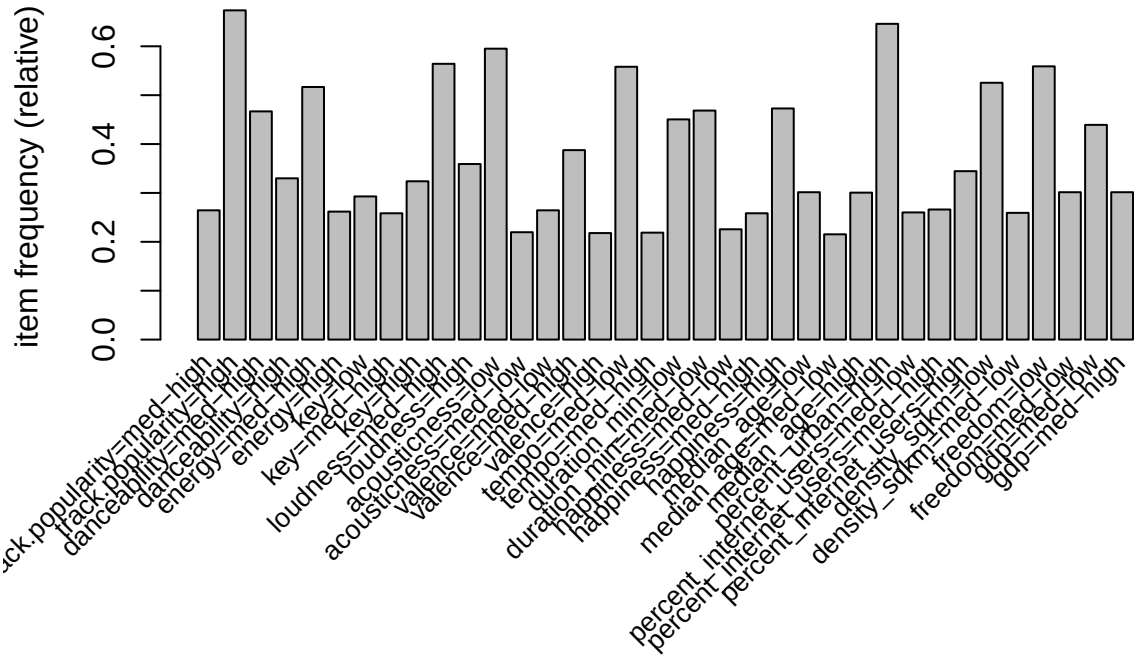
```
# make transactional dataset
spot <- as(spot, 'transactions')
inspect(spot[1])
```

```
##     items                            transactionID
```

```
## [1] {track.name=Tusa,
##      country=Dominican Republic,
##      track.popularity=high,
##      danceability=high,
##      energy=med-high,
##      key=low,
##      loudness=high,
##      acousticness=med-low,
##      valence=med-high,
##      tempo=med-low,
##      duration_min=med-low,
##      happiness=med-low,
##      median_age=low,
##      percent_urban=high,
##      percent_internet_users=med-high,
##      density_sqkm=med-low,
##      freedom=med-low,
##      gdp=med-high}                                  1
```

Plot

```
itemFrequencyPlot(spot, support = 0.2, cex.names = 0.8)
```



Notably, none of the top 50 tracks are above the relative frequency threshold minimum, despite some tracks being nearly ubiquitously popular.

Mine and Inspect Frequent Itemsets