Visualizing Movie Ratings

Data Computing

March 17, 2016

A set of 100,000 ratings of movies by individuals was collected in the late 1990s by the *grouplens* research team at the University of Minnesota. They provide the data directly at http://grouplens.org/datasets/movielens/100k/. These data were reformatted by DTK and can be downloaded to your own computer with this statement:

```
download.file("http://tiny.cc/dcf/MovieLens.rda", destfile = "MovieLens.rda")
```

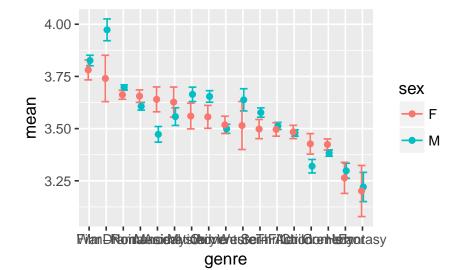
Use load() to read in the data to your R session. MovieLens.rda contains three data tables:

- Ratings has the individual movie ratings and the time at which they were entered. It also includes an ID variable for both the user and the movie.
- Movies provides the name of the movie and information about genres.
- Users gives basic information about the person who made the rating.

Your task: Construct each of these graphics.

One: Showing the appeal of different genres to the different sexes

```
## Joining, by = "movie_id"
## Joining, by = "user_id"
## Joining, by = "genre"
```



Which genres are connected?

```
Look at correlation between genres?
```

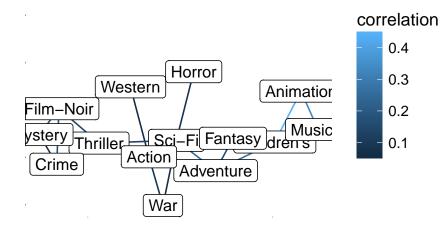
```
Genres <- Movies[,6:23]</pre>
tmp <- cor(Genres) %>% as.data.frame(stringsAsFactors = FALSE)
tmp$genre <- row.names(tmp)</pre>
Genre_pairs <-
  tmp %>%
  gather(key = genre2, value = correlation, -genre) %>%
  filter(genre != genre2) %>%
  filter(genre > genre2) %>%
  group_by(genre) %>%
# filter(rank(desc(correlation)) <= 3) %>%
  filter(correlation > 0.05)
Genre_pairs %>%
  ggplot(aes(x = genre2, y = correlation)) +
  geom_point(aes(color = genre2)) +
  facet_wrap( ~ genre)
                                                        Action
         Adventure Children's
                              Fantasy
                                       Film-Noir
                                                        Adventure
                                                        Animation
                                                        Children's
  correlation
          Musical
                    Mystery
                             Romance
                                         Sci-Fi
                                                        Comedy
                                                        Crime
                                       ......
                                                        Fantasy
          Thriller
                      War
                              Western
                                                        Film-Noir
                                                        Horror
                   manna tamana
          .....
                                                        Mystery
                         genre2
                                                        Sci-Fi
  As a network
```

```
library(igraph)
```

```
##
## Attaching package: 'igraph'
## The following objects are masked from 'package:tidyr':
##
##
       %>%, crossing
## The following objects are masked from 'package:dplyr':
```

```
##
       %>%, as_data_frame, groups, union
##
## The following objects are masked from 'package:stats':
##
##
       decompose, spectrum
## The following object is masked from 'package:base':
##
##
       union
Vertices <-
 Genre_pairs %>%
  edgesToVertices(from = genre, to = genre2)
Edges <-
 Vertices %>%
  edgesForPlotting(ID = ID, x, y, Edges = Genre_pairs, from = genre, to = genre2)
Vertices %>%
 ggplot(aes(x = x, y = y)) + geom_point()+
  geom_segment(data = Edges,
               aes(x = x, y = y, xend = xend, yend = yend,
                   color = correlation)) +
 theme_map() +
    geom_label(aes(label = ID), fill = "white")
## Warning: 'panel.margin' is deprecated. Please
## use 'panel.spacing' property instead
```

Comedy hance



Who are the reviewers?

```
Users %>%
  ggplot(aes(x = age)) +
```

```
geom_density(aes(fill = occupation),
               color = NA, alpha = .7, position = "fill") +
  facet_wrap( ~ sex)
                                    administrator
                                                       marketing
                       M
                                     artist
                                                       none
      1.00
                                     doctor
                                                       other
      0.75 -
                                     educator
                                                       programmer
  density
                                     engineer
                                                       retired
      0.50
                                     entertainment
                                                       salesman
                                     executive
                                                       scientist
      0.25 -
                                    healthcare
                                                       student
      0.00
                                     homemaker
                                                       technician
                    204060
            204060
                                     lawyer
                                                       writer
                 age
                                    librarian
Users %>%
  ggplot(aes(x = age)) +
  geom_density(aes(fill = sex),
               color = NA, alpha = .7, position = "stack") +
  facet_wrap( ~ occupation)
           ministrat
                      artist
                               doctor
                                       educator
                                               engineer
      tertainme executive lealthcare
                                       omemak
                                                 lawyer
      sex
  density
           librarian
                               none
                    narketing
                                        other
                                                ogramm
                                                                F
                                                                M
            retired
                    salesmar
                              scientist
                                       student
                                                echniciai
      204060
                              204060
                                       204060
                                                 204060
             writer
      204060
                               age
Users %>%
  group_by(occupation) %>%
  tally() %>%
```

arrange(desc(n))

```
## # A tibble: 21 × 2
##
         occupation
                         n
##
              <chr> <int>
## 1
            student
                       196
## 2
              other
                       105
## 3
           educator
                        95
      administrator
                        79
## 4
## 5
           engineer
                        67
## 6
         programmer
                        66
## 7
          librarian
                        51
## 8
             writer
                        45
## 9
          executive
                        32
## 10
          scientist
                        31
## # ... with 11 more rows
Ratings as people age
All %>%
  filter( genre != "unknown") %>%
  ggplot(aes(x = age, color = sex, y = rating)) +
  geom_smooth() +
  facet_wrap( ~ genre)
## 'geom_smooth()' using method = 'gam'
          Action
                  Adventure Animation Children's Comedy
                                      Fantasy
                                               Film-Noir
          Crime
                  ocumenta
                             Drama
                                                            sex
                                                                F
          Horror
                            Mystery
                                      Romance
                                                 Sci-Fi
                   Musical
                                                                M
                                      204060
                                                204060
         Thriller
                     War
                            Western
         204060
                            204060
                   204060
                              age
All %>%
  ggplot(aes(x = age, color = sex, y = rating)) +
```

geom_smooth()

'geom_smooth()' using method = 'gam'

