Movie Lens Data Cleaning and Merging Demo

CZ

December 5, 2016

Project Overview

Assignment

Using the Movie Lens data, we were instructed to clean and merge the data sets in order to:

- Construct, clean, and label a data frame including the following variables:
  - Demographic Info: user id, gender, age
  - Movie Info: movie id, rating, time stamp, title, genre, year of movie
  - Location Info: zip code, state, longitude, latitude
- Generate a summary table of average reviews by gender for each gender. Visualize the results.
- Create a heat map of the entire country of the count of reviews.
- Create a csv file of the new data frame.

Data Intro

This exercise was designed to utilize publicly available Movie Lens data sets provided generously by the University of Minnesota. In total, four data sets were used. They are summarized in the table below.

Data set	Description	Key Variables
u.data	The full u data set, 100,000 ratings by 943 users on 1682 items. Each user has rated at least 20 movies. Users and items are numbered consecutively from 1. The data is randomly ordered.	user id, item id, rating, time stamp
u.item	Information on the items (movies).	movie id, movie title, release date, video release date, IMDb URL, 19 genres
u.user	Demographic information about the users	user id, age, gender, occupation, zip code
zips.csv	Information on location.	zip code, state,

longitude, latitude

Document Sections

The following document will proceed as follows:

1. Set up
2. Cleaning the data
3. Merging the data frames
4. Summary Statistics
5. Heat Map
6. Export to CSV

Set Up

Set up includes installing and reading in the necessary packages and libraries. It also includes reading in the four data sets listed above. Packages used include:

- plyr
- dplyr
- tidyr
- stringer
- ggmap
- ggplot2

Summaries of the four data sets are below. As you can see, only the last data set, zips_csv has variable labels. We will add them in the next section, Data Cleaning.

```
#Dataset 1: U.data contains user id, movie id, rating, timestamp
summary(u_data)

##        V1              V2               V3              V4
##  Min.   :  1.0   Min.   :   1.0   Min.   :1.00   Min.   :874724710
##  1st Qu.:254.0   1st Qu.: 175.0   1st Qu.:3.00   1st Qu.:879448710
##  Median :447.0   Median : 322.0   Median :4.00   Median :882826944
##  Mean   :462.5   Mean   : 425.5   Mean   :3.53   Mean   :883528851
##  3rd Qu.:682.0   3rd Qu.: 631.0   3rd Qu.:4.00   3rd Qu.:888259984
##  Max.   :943.0   Max.   :1682.0   Max.   :5.00   Max.   :893286638

#Dataset 2: U.item contains movie id, movie title, release data, genre
summary(u_item)

##        V1                                      V2                 V3
##  Min.   :   1.0   Body Snatchers (1993)         :   2   01-Jan-1995:215
##  1st Qu.: 421.2   Butcher Boy, The (1998)       :   2   01-Jan-1994:213
##  Median : 841.5   Chairman of the Board (1998)  :   2   01-Jan-1993:126
##  Mean   : 841.5   Chasing Amy (1997)            :   2   01-Jan-1997: 98
##  3rd Qu.:1261.8   Deceiver (1997)               :   2   01-Jan-1992: 37
##  Max.   :1682.0   Designated Mourner, The (1997):   2   01-Jan-1996: 26
##                   (Other)                       :1670   (Other)    :967
##     V4
##  Mode:logical
##  NA's:1682
##
##
##
##
##
##
V5
##
:   3
##  http://us.imdb.com/M/title-exact?Body%20Snatchers%20(1993)
:   2
##  http://us.imdb.com/M/title-exact?Chasing+Amy+(1997)
:   2
##  http://us.imdb.com/M/title-
exact?Designated%20Mourner%2C%20The%20%281997%29:   2
##  http://us.imdb.com/M/title-exact?Fly%20Away%20Home%20(1996)
:   2
##  http://us.imdb.com/M/title-exact?Hugo+Pool+(1997)
:   2
##  (Other)
:1669
##        V6                 V7               V8                V9
##  Min.   :0.000000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.000000   Median :0.0000   Median :0.00000   Median :0.00000
##  Mean   :0.001189   Mean   :0.1492   Mean   :0.08026   Mean   :0.02497
##  3rd Qu.:0.000000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.000000   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000
##
##       V10              V11              V12              V13
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000
##  Median :0.00000   Median :0.0000   Median :0.0000   Median :0.00000
##  Mean   :0.07253   Mean   :0.3002   Mean   :0.0648   Mean   :0.02973
##  3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000
##
##       V14            V15               V16               V17
##  Min.   :0.000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.000   Median :0.00000   Median :0.00000   Median :0.0000
##  Mean   :0.431   Mean   :0.01308   Mean   :0.01427   Mean   :0.0547  
##  3rd Qu.:1.000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.0000  
##  Max.   :1.000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000  
##                                                                      
##       V18               V19               V20              V21         
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.00000  
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.00000  
##  Median :0.00000   Median :0.00000   Median :0.0000   Median :0.00000  
##  Mean   :0.03329   Mean   :0.03627   Mean   :0.1468   Mean   :0.06005  
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.00000  
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :1.00000  
##                                                                        
##       V22              V23               V24         
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.00000  
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000  
##  Median :0.0000   Median :0.00000   Median :0.00000  
##  Mean   :0.1492   Mean   :0.04221   Mean   :0.01605  
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000  
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.00000  
## 
```

```
#Dataset 3: U.user contains user id, age, gender, zip code
summary(u_user)
```

```
##         V1              V2          V3                   V4             V5     
##  Min.   :  1.0   Min.   : 7.00   F:273   student      :196   55414  :  9  
##  1st Qu.:236.5   1st Qu.:25.00   M:670   other        :105   55105  :  6  
##  Median :472.0   Median :31.00           educator     : 95   10003  :  5  
##  Mean   :472.0   Mean   :34.05           administrator: 79   20009  :  5  
##  3rd Qu.:707.5   3rd Qu.:43.00           engineer     : 67   55337  :  5  
##  Max.   :943.0   Max.   :73.00           programmer   : 66   27514  :  4  
##                                          (Other)      :335   (Other):909
```

```
#Dataset 3: U.user contains user id, age, gender, zip code
summary(zips_csv)
```

```
##     zip.code     X..state.abbreviation.        X..latitude.  
##  006HH  :     1    "TX"  : 1936           " 29.896156":    2  
##  006XX  :     1    "PA"  : 1776           " 31.134863":    2  
##  007HH  :     1    "CA"  : 1757           " 32.186824":    2  
##  007XX  :     1    "NY"  : 1675           " 33.747017":    2  
##  009HH  :     1    "IL"  : 1375           " 33.766057":    2  
##  010HH  :     1    "OH"  : 1189           " 33.824496":    2  
##  (Other):33172   (Other):23470          (Other)      :33166  
##         X..longitude.              X..city.                X..state.    
##   "-101.19002":    2    ""            : 1161    "Texas"       : 1936  
##   "-105.01123":    2    "Houston"     :  102    "Pennsylvania": 1776  
##   "-105.08431":    2    "New York"    :   64    "California"  : 1757  
##   "-105.10036":    2    "Los Angeles" :   56    "New York"    : 1675  
##   "-109.54223":    2    "Philadelphia":   53    "Illinois"    : 1375  
##   "-110.98801":    2    "Dallas"      :   52    "Ohio"        : 1189  
##  (Other)       :33166   (Other)       :31690   (Other)        :23470  
```

Cleaning the Data

In order to prepare the data sets for merging, a series of data cleaning tasks needed to be carried out.

1. Variable labels were manually added.
    a. Note: I assumed that the variable "item_id" from u.data was the same as "movie_id" in u.item.
2. The date variable from u_item was split into day, month, year so we could access the movie release year.
3. Latitude and Longitude in zips_csv were converted to numeric variables from factor.
    a. Note: In additional to using as.numeric, I also needed to remove the quotes around the numbers.
    b. Note: An area of concern is that there are many missing lat and long values. This will interfere with the accuracy of the heat map later.

Merging Data Sets

After cleaning the data, I simply used the merge function to create one aggregate data frame, merge_all.

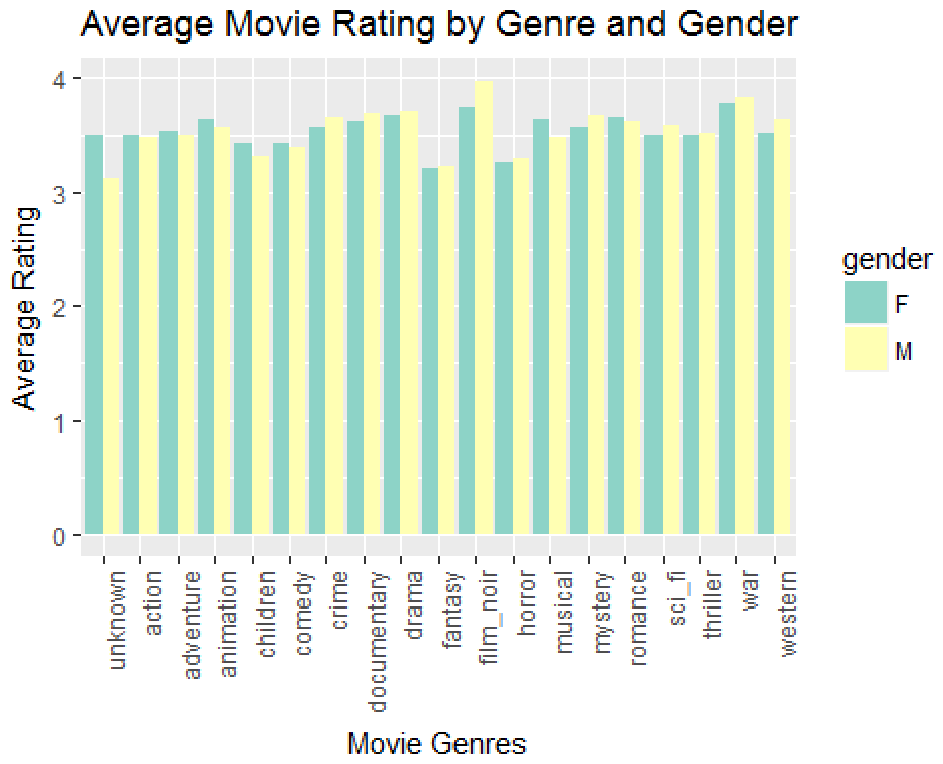
Summary Statistics

We were tasked with creating a summary table to display the average review by genre for each gender. This section turned out to be a lot trickier than I initially realized. Complication: A simple aggregate function does not work because the movie genre categories are not mutually exclusive. Because movies are labeled under multiple genres, the aggregate function splits up the genres into too many groups. Whereas I want an average of all movies under a certain genre say, drama, the aggregate function was returning different values for movies that were just drama, drama and action, and drama, action, and film noir, for example.   Solution: In order to address this issue, I created separate columns for each genre of the ratings assigned to that movie. If there was a 1 in the genre row, the associated rating for that movie would be carried over to the new genre rating column. For instance, if we were only looking at a section of the data frame (rating, action, and drama), we would create the two new columns on the right hand side.

Rating	Action	Drama	Action Rating	Drama Rating
2	1	0	2	0
3	0	1	0	3
3	1	1	3	3
4	0	0	0	0

After using a for loop to create these new rating variables, I was then able to use the aggregate function to create the table below. Following that is a bar graph of the same table.

```
##   Group.1 unknown_r action_r adventure_r animation_r children_r comedy_r
## 1       F     3.500 3.484013    3.517988    3.627136   3.426971 3.424021
## 2       M     3.125 3.479228    3.499246    3.557471   3.320000 3.382972
##    crime_r documentary_r  drama_r fantasy_r film_noir_r horror_r musical_r
## 1 3.556299      3.614973 3.662246  3.201102    3.740260 3.263993  3.640083
## 2 3.654049      3.691769 3.696957  3.220425    3.973294 3.298058  3.472665
##   mystery_r romance_r sci_fi_r thriller_r    war_r western_r
## 1  3.560122  3.655685 3.497908   3.496068 3.781179  3.514825
## 2  3.664208  3.607072 3.577072   3.512927 3.826328  3.637896
```

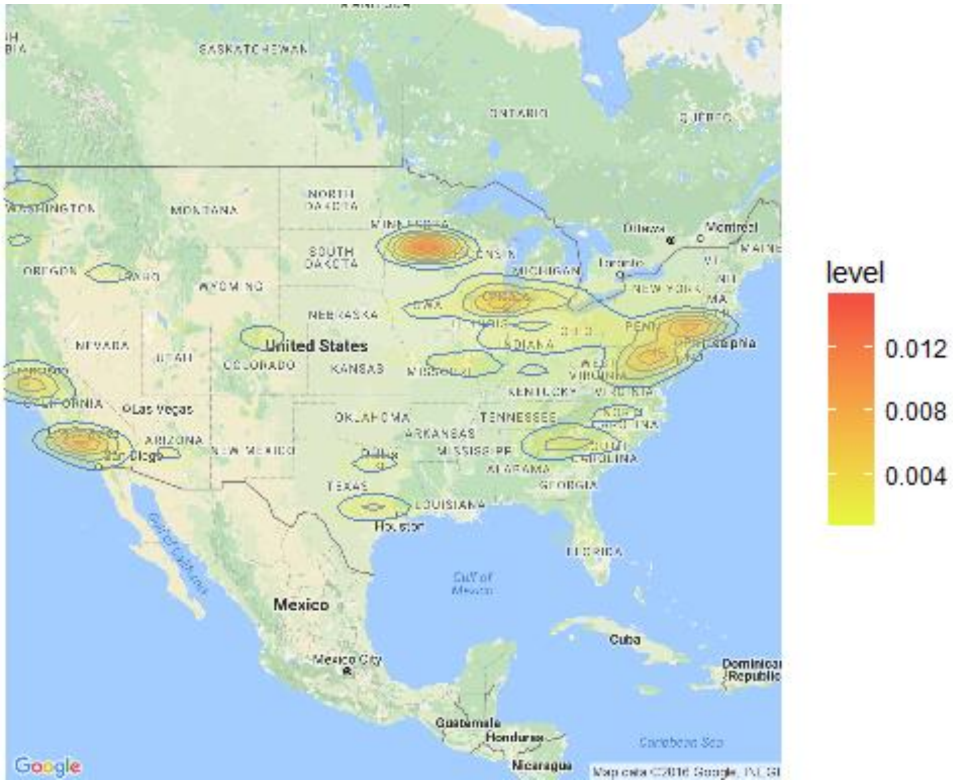


Heat Map

The last visualization was a heat map of the number of ratings per regions. See below for the results.

```
## Map from URL :
http://maps.googleapis.com/maps/api/staticmap?center=united+states&zoom=4&siz
e=640x640&scale=2&maptype=roadmap&language=en-EN&sensor=false

## Information from URL :
http://maps.googleapis.com/maps/api/geocode/json?address=united%20states&sens
or=false
```



Export to CSV

This was the final part of the assignment. It was completed with a simple line of code:

```
write.csv(merge_all, file = "Movie Lens_Merged_DF.csv")
```

That's it! Thanks for reading.