**Analysis Code**

This document contains details of the **analysis done on the movie lens dataset with code and comments.** It also includes **the inference from the analysis.** Followed by the **code wtih comments.**

**Libraries Used**

library(ggplot2)

library(plyr)

library(RColorBrewer)

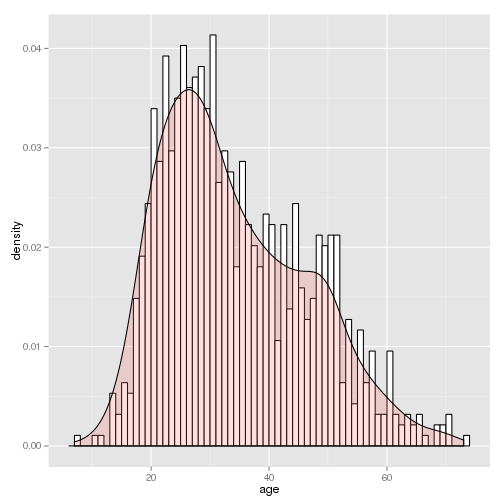
library(grid)

**#Investigating General Dataset**

**Age Plot**

**Code:**

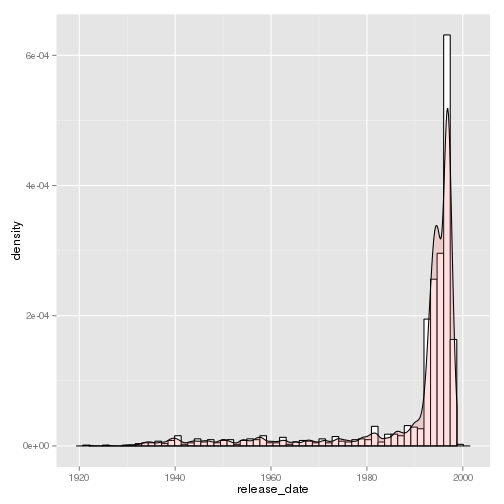
|  |
| --- |
| mlDat\_user <- ddply(mlDat, ~user\_id + age + gender + occupation, summarize, mean\_rating = mean(rating))  agePlot <- ggplot(mlDat\_user, aes(age)) + geom\_histogram(aes(y=..density..), binwidth=1, colour="black", fill="white")  agePlot <- agePlot + geom\_density(alpha=.2, fill="#FF6666")  print(agePlot)  ggsave(filename = "agePlot.jpg") |



**Observation:**

Users tend to be mostly in the late teens and mid thirties, though there seems to be another peak the occurs in the late forties.

**Movies Date Plot:**



**Observation:**

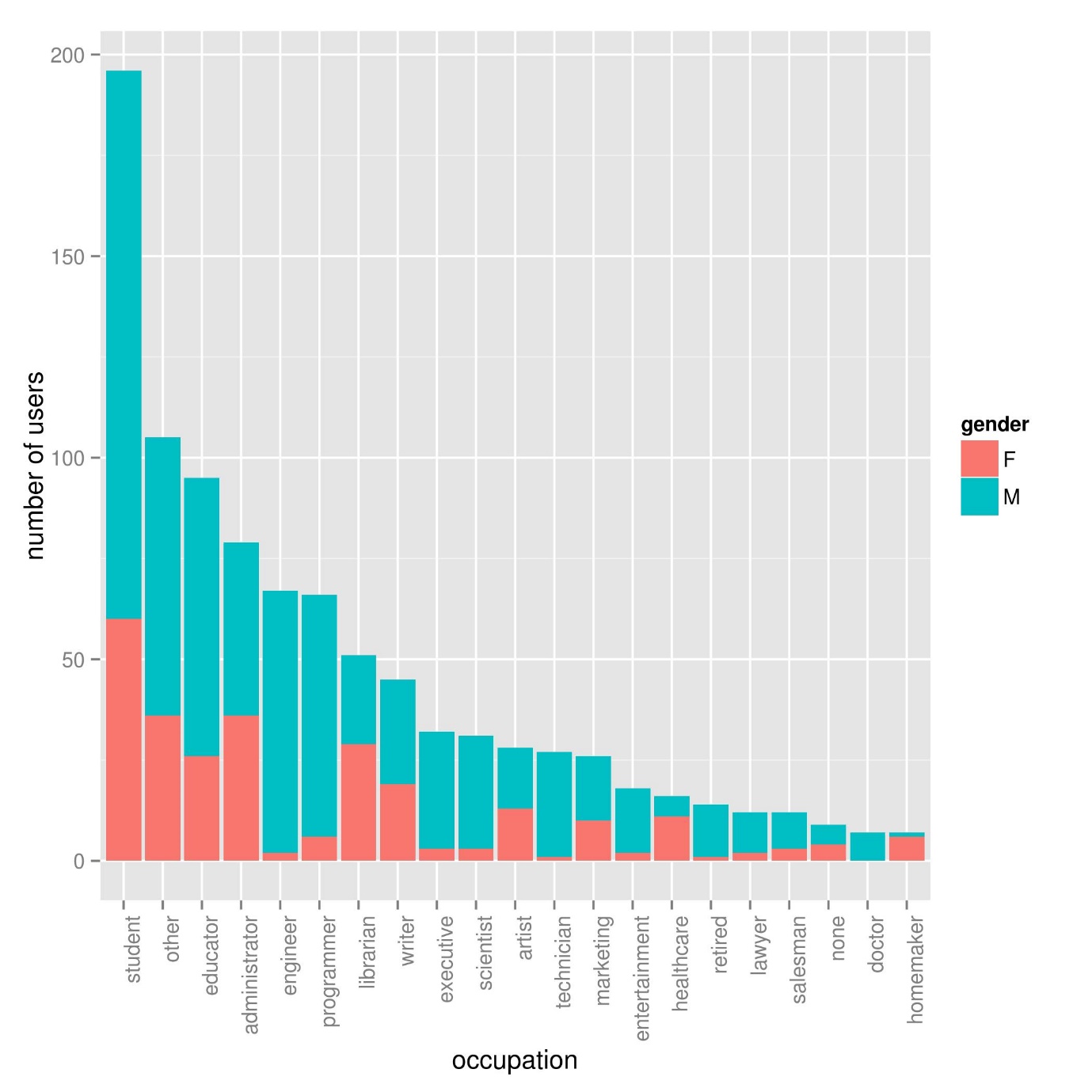
Most movies in the data tend to be from the 1990's. There is a pretty long tale meaning they have at least some moves from the past.

**Code Snippet:**

|  |
| --- |
| mlDat\_movie <- ddply(mlDat, ~movie\_title + release\_date + genre, summarize, mean\_rating = mean(rating))  datesPlot <- ggplot(mlDat\_movie, aes(release\_date)) + geom\_histogram(aes(y=..density..), binwidth=500, colour="black", fill="white")  #alter axis  datesPlot <- datesPlot + geom\_density(alpha=.2, fill="#FF6666")  print(datesPlot)  ggsave(filename = "datesPlot.jpg") |

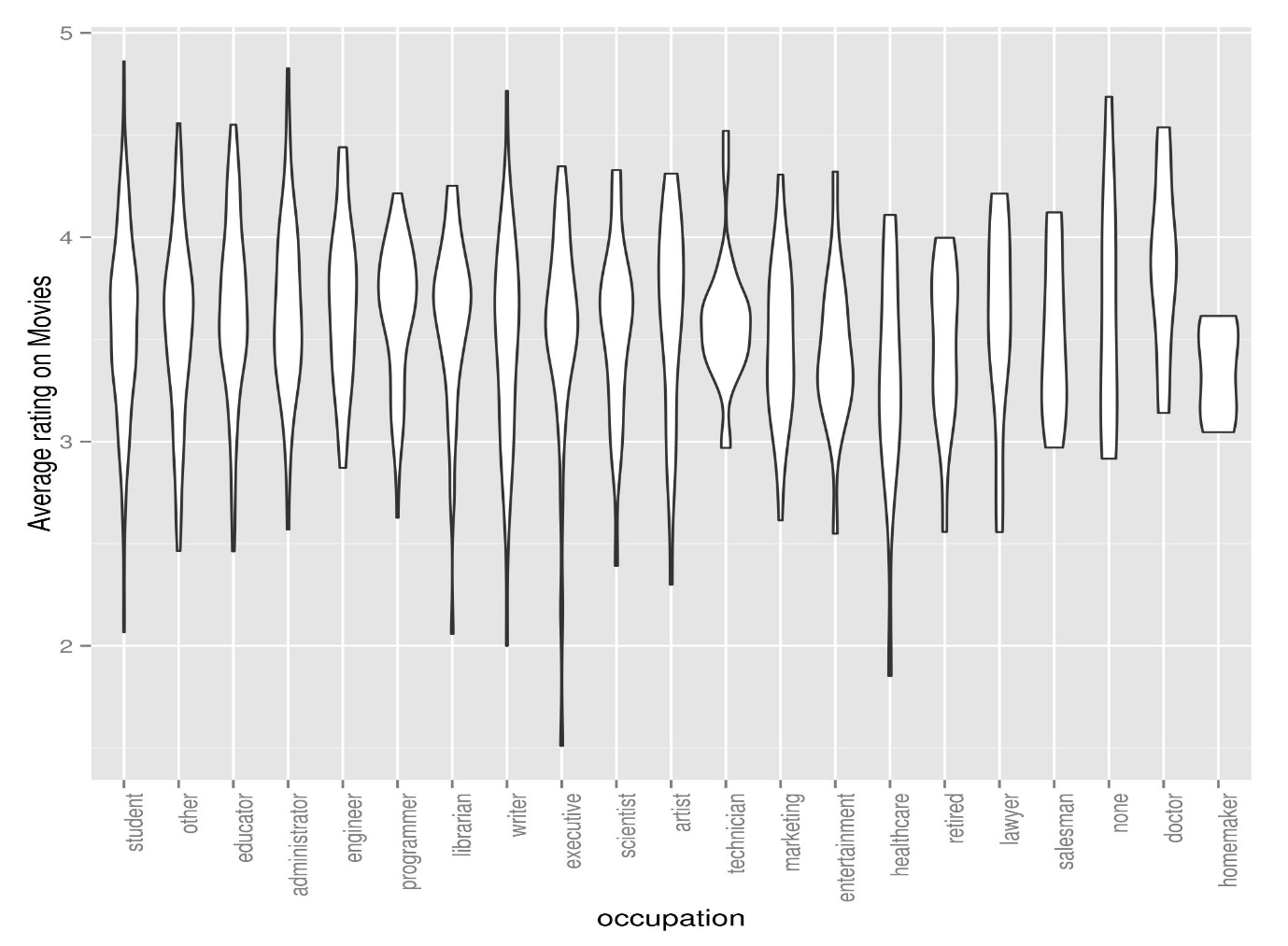
**Investigate the users with respect to profession that contributed to the dataset**

* Total numbers of each profession that contributed to the dataset
* Gender bias in each profession
* How the professions tend to rank movies



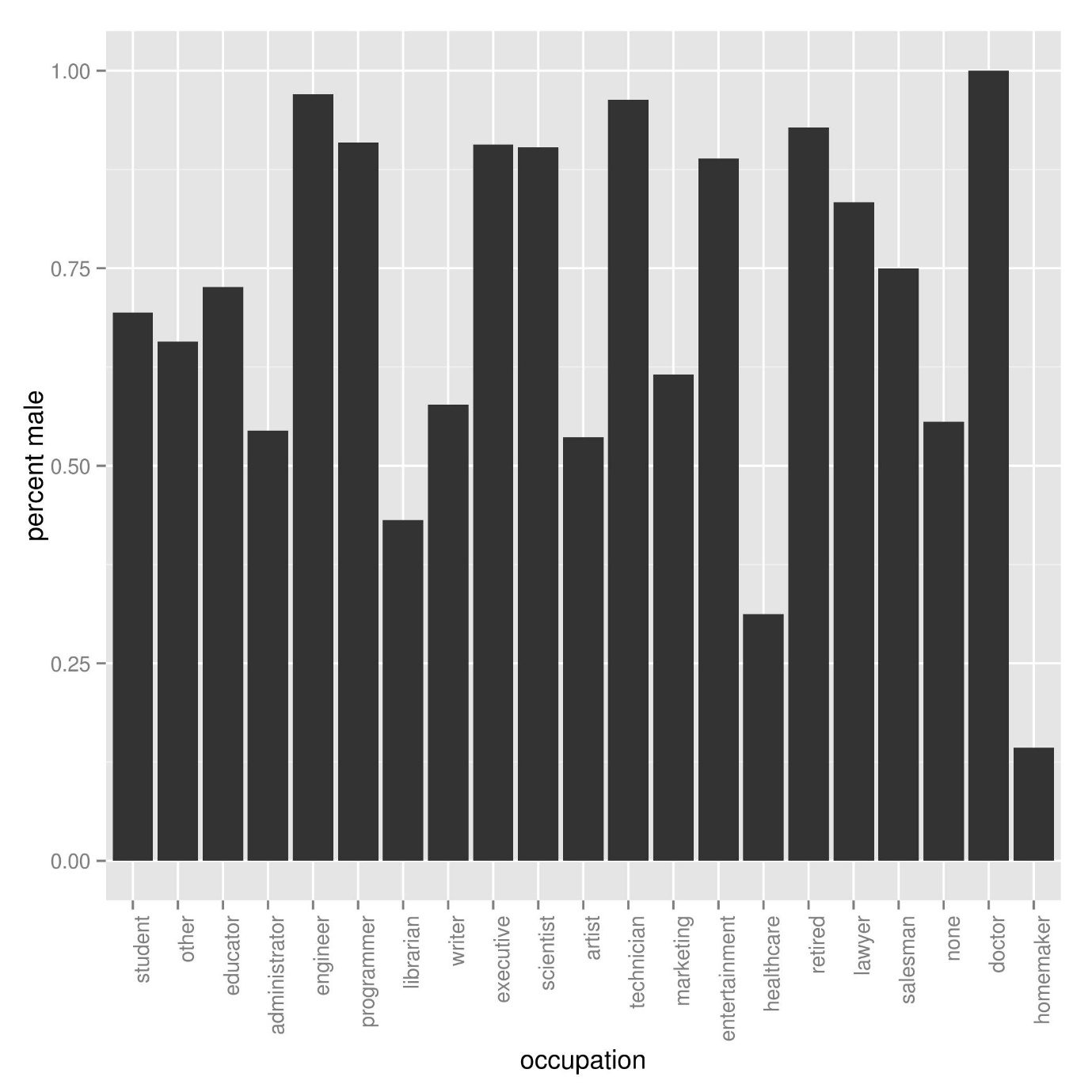
Code Snippet:

|  |
| --- |
| #Profession based analysis block block graph  #sorts by number of users  userPlot <- ggplot(mlDat\_user, aes(x=reorder(occupation,occupation,  function(x)-length(x)), fill = gender)) + geom\_bar()  #fix axis  userPlot <- userPlot + theme(axis.text.x = element\_text(angle = 90, hjust = 1))  userPlot <- userPlot + ylab("number of users") + xlab("occupation")  #flip axis to make professions easier to read  #userPlot <- userPlot + coord\_flip()  ggsave(filename = "userPlot.jpg") |



Code Snippet:

|  |
| --- |
| rankPlot <- ggplot(mlDat\_user, aes(x=reorder(occupation,occupation,  function(x)-length(x)), mean\_rating)) + geom\_violin()  #fix axis  rankPlot <- rankPlot + theme(axis.text.x = element\_text(angle = 90, hjust = 1))  rankPlot <- rankPlot + ylab("Average rating on Movies") + xlab("occupation")  #flip axis to make professions easier to read  #rankPlot <- rankPlot + coord\_flip()  ggsave(filename = "rankPlot.jpg") |

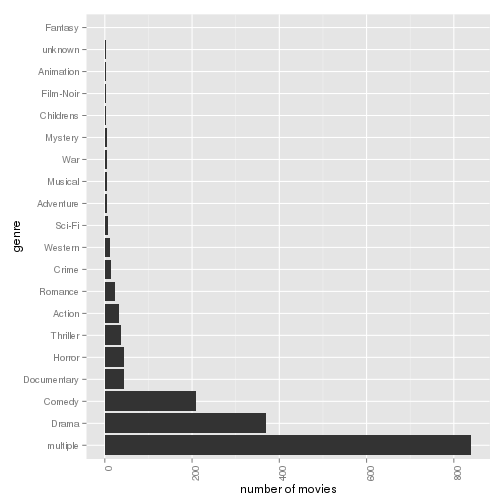


**Code Snippet:**

|  |
| --- |
| gender\_dat <- ddply(mlDat\_user, ~occupation, summarize, perc\_male = (length(gender[gender == "M"])/length(gender)), counts = -length(user\_id))  #sorts by number of users  genderPlot <- ggplot(gender\_dat, aes(x=reorder(occupation, counts), perc\_male)) + geom\_bar(stat="identity")  #fix axis  genderPlot <- genderPlot + theme(axis.text.x = element\_text(angle = 90, hjust = 1))  genderPlot <- genderPlot + ylab("percent male") + xlab("occupation")  #flip axis to make professions easier to read  #genderPlot <- genderPlot + coord\_flip()  ggsave(filename = "genderPlot.jpg") |

**Observations:**

* There are very few doctors and homemakers, we probably can't say anything about these groups with very much confidence
* Males make up of more of our sample. Some professions like engineering (rather unsurprisingly) are completely male dominated.
* The professions do not rank things evenly. Some appear more picky; for example executives seem to sometime rank movies very low and healthcare workers seem to have a very low average rating.

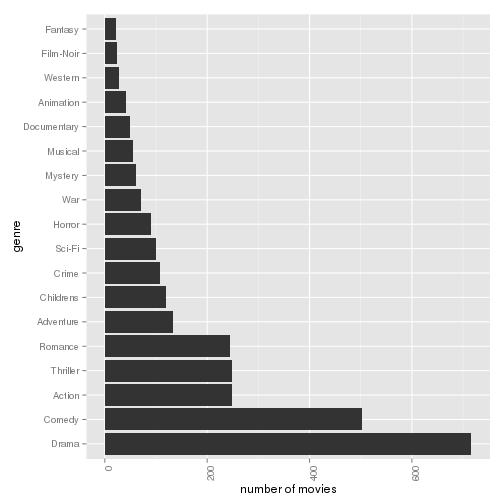
Investigate the movies with respect to users in the dataset

**Observation:**

A majority of the titles are multi genre. There really does not seem to be an even spread of titles, for example there are almost no pure fantasy titles.

**Code Snippet:**

|  |
| --- |
| genreCountPlot <- ggplot(mlDat\_movie, aes(x=reorder(genre,genre,  function(x)-length(x)))) + geom\_bar()  #fix axis  genreCountPlot <- genreCountPlot + theme(axis.text.x = element\_text(angle = 90, hjust = 1))  genreCountPlot <- genreCountPlot + ylab("number of movies") + xlab("genre")  genreCountPlot <- genreCountPlot + coord\_flip()  print(genreCountPlot)  ggsave(filename = "genreCountPlot.jpg") |

**Total number of movies with a specific genre counted multiple times for multi genre movies: **

**Code Snippet:**

|  |
| --- |
| mlDat\_movie\_multi <- ddply(mlDat\_multi, ~movie\_title + release\_date + genre, summarize, mean\_rating = mean(rating))  genreCountPlot\_multi <- ggplot(mlDat\_movie\_multi, aes(x=reorder(genre,genre,  function(x)-length(x)))) + geom\_bar()  #fix axis  genreCountPlot\_multi <- genreCountPlot\_multi + theme(axis.text.x = element\_text(angle = 90, hjust = 1))  genreCountPlot\_multi <- genreCountPlot\_multi + ylab("number of movies") + xlab("genre")  genreCountPlot\_multi <- genreCountPlot\_multi + coord\_flip()  print(genreCountPlot\_multi)  ggsave(filename = "genreCountPlot\_multi.jpg") |

Investigate the trends in movies with respect to rating and other factors within the dataset

Does genre affect the rating of a movie? Does genre matter to the average male or female?

Pure Genre:



**Observation:**

Noir and Animation seem to be the highest rated surprisingly. Low sample sizes for some of these is a problem; the only reason fantasy and war seem like they make a difference in terms of gender is that the sample size for both is quite small. The only gender differnce we notice is maybe the fact that women seem to like musicals more than men.

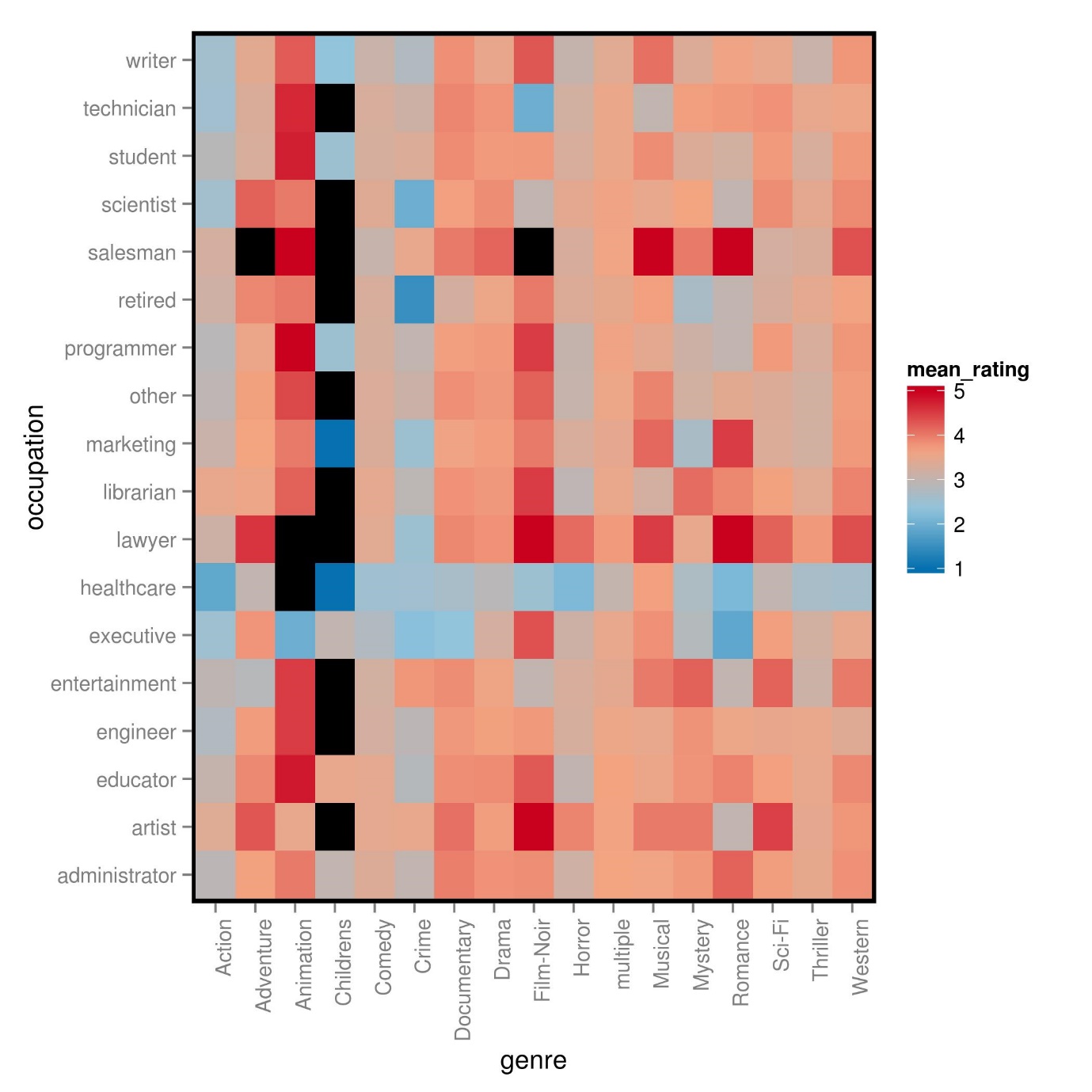
**Multiple Genres:**



The low sample sizes seem to be fixed. Things noticed:

* Horror movies do not get good ratings. It fits my expectations since horror movies tend to typically focused on cheap thrills but are not typically impactful.
* There are very few fantasy films and they tend to rank low. It could be that no one wants to take risks on making these types of movies because of this.

Investigate trends in profession with respect to movie genre preference



Code Snippet:

|  |
| --- |
| mlDat\_genre\_occup <- ddply(mlDat, ~genre + occupation, summarize, mean\_rating = mean(rating))  mlDat\_genre\_occup <- droplevels(subset(mlDat\_genre\_occup, occupation != "homemaker"))  mlDat\_genre\_occup <- droplevels(subset(mlDat\_genre\_occup, genre != "unknown"))  mlDat\_genre\_occup <- droplevels(subset(mlDat\_genre\_occup, genre != "Fantasy"))  mlDat\_genre\_occup <- droplevels(subset(mlDat\_genre\_occup, occupation != "none"))  mlDat\_genre\_occup <- droplevels(subset(mlDat\_genre\_occup, genre != "War"))  mlDat\_genre\_occup <- droplevels(subset(mlDat\_genre\_occup, occupation != "doctor"))  heatMapPalette <- colorRampPalette(rev(brewer.pal(4, "RdBu")))  #get data ready for heatmap  goHeat <- ggplot(mlDat\_genre\_occup, aes(x = genre, y = occupation, fill = mean\_rating))  #rotate labels  goHeat <- goHeat + geom\_tile() + theme(axis.text.x = element\_text(angle = 90, hjust = 1))  #add colours  goHeat <- goHeat + scale\_fill\_gradientn(colours = heatMapPalette(100))  #change background  goHeat <- goHeat + theme(panel.background = element\_rect(fill='black'),  panel.grid.major = element\_blank(),  panel.grid.minor = element\_blank())  print(goHeat)  ggsave(filename = "genreOccupHeatMap.jpg") |

Note: Black tiles are missing data.

So now we are noticing some other interesting trends:

* The retired (and likely elderly) do not like pure crime movies. Maybe they prefer not to think about what crimes could happen.
* People who work in healthcare (not doctors) have extremely high standards.
* Executives tend to dislike many movie genres but still like noir films.
* Lawyers like noir and romance films.

Note: Some professions are underrepresented (for instance there are only 7 doctors in this dataset who also happen to be male). Also, Pooling together multiple genre movies greatly reduced possible the sample sizes for genres.

**Actual Code with Comments**

**----**

library(ggplot2)

library(plyr)

library(RColorBrewer)

library(grid)

**#load single genre file**

mlDat <- read.csv("Results/unifiedMLData.csv")

**#fix dates field**

mlDat$release\_date <- as.Date(mlDat$release\_date, "%Y-%m-%d")

**#sanity checking**

str(mlDat)

summary(mlDat)

head(mlDat)

tail(mlDat)

#load multi genre file

mlDat\_multi <- read.csv("Results/unifiedMLDataMulti.csv")

**#fix dates field**

mlDat\_multi$release\_date <- as.Date(mlDat\_multi$release\_date, "%Y-%m-%d")

str(mlDat\_multi)

summary(mlDat\_multi)

head(mlDat\_multi)

tail(mlDat\_multi)

**#prepare table for analysis of users**

mlDat\_user <- ddply(mlDat, ~user\_id + age + gender + occupation, summarize, mean\_rating = mean(rating))

agePlot <- ggplot(mlDat\_user, aes(age)) + geom\_histogram(aes(y=..density..), binwidth=1, colour="black", fill="white")

agePlot <- agePlot + geom\_density(alpha=.2, fill="#666666")

print(agePlot)

ggsave(filename = "agePlot.jpg")

**#prepare table for analysis of movies**

mlDat\_movie <- ddply(mlDat, ~movie\_title + release\_date + genre, summarize, mean\_rating = mean(rating))

datesPlot <- ggplot(mlDat\_movie, aes(release\_date)) + geom\_histogram(aes(y=..density..), binwidth=500, colour="black", fill="white")

**#alter axis**

datesPlot <- datesPlot + geom\_density(alpha=.2, fill="#FF6633")

print(datesPlot)

ggsave(filename = "datesPlot.jpg")

# enna pannalam idhuku approm !

**#Profession based analysis block block graph**

**#sorts by number of users**

userPlot <- ggplot(mlDat\_user, aes(x=reorder(occupation,occupation,

function(x)-length(x)), fill = gender)) + geom\_bar()

**#fix axis**

userPlot <- userPlot + theme(axis.text.x = element\_text(angle = 90, hjust = 1))

userPlot <- userPlot + ylab("number of users") + xlab("occupation")

**#flip axis to make professions easier to read**

#userPlot <- userPlot + coord\_flip()

ggsave(filename = "userPlot.jpg")

gender\_dat <- ddply(mlDat\_user, ~occupation, summarize, perc\_male = (length(gender[gender == "M"])/length(gender)), counts = -length(user\_id))

**#sorts by number of users**

genderPlot <- ggplot(gender\_dat, aes(x=reorder(occupation, counts), perc\_male)) + geom\_bar(stat="identity")

#fix axis

genderPlot <- genderPlot + theme(axis.text.x = element\_text(angle = 90, hjust = 1))

genderPlot <- genderPlot + ylab("percent male") + xlab("occupation")

**#flip axis to make professions easier to read**

#genderPlot <- genderPlot + coord\_flip()

ggsave(filename = "genderPlot.jpg")

################

rankPlot <- ggplot(mlDat\_user, aes(x=reorder(occupation,occupation,

function(x)-length(x)), mean\_rating)) + geom\_violin()

#fix axis

rankPlot <- rankPlot + theme(axis.text.x = element\_text(angle = 90, hjust = 1))

rankPlot <- rankPlot + ylab("Average rating on Movies") + xlab("occupation")

#flip axis to make professions easier to read

#rankPlot <- rankPlot + coord\_flip()

ggsave(filename = "rankPlot.jpg")

**######################### genre analysis**

genreCountPlot <- ggplot(mlDat\_movie, aes(x=reorder(genre,genre,

function(x)-length(x)))) + geom\_bar()

#fix axis

genreCountPlot <- genreCountPlot + theme(axis.text.x = element\_text(angle = 90, hjust = 1))

genreCountPlot <- genreCountPlot + ylab("number of movies") + xlab("genre")

genreCountPlot <- genreCountPlot + coord\_flip()

print(genreCountPlot)

ggsave(filename = "genreCountPlot.jpg")

**######################## genre analysis - multiple**

mlDat\_movie\_multi <- ddply(mlDat\_multi, ~movie\_title + release\_date + genre, summarize, mean\_rating = mean(rating))

genreCountPlot\_multi <- ggplot(mlDat\_movie\_multi, aes(x=reorder(genre,genre,

function(x)-length(x)))) + geom\_bar()

#fix axis

genreCountPlot\_multi <- genreCountPlot\_multi + theme(axis.text.x = element\_text(angle = 90, hjust = 1))

genreCountPlot\_multi <- genreCountPlot\_multi + ylab("number of movies") + xlab("genre")

genreCountPlot\_multi <- genreCountPlot\_multi + coord\_flip()

print(genreCountPlot\_multi)

ggsave(filename = "genreCountPlot\_multi.jpg")

**##################### gender based analysis whether this affects or not anything!**

#mlDat\_avgRating <- ddply(mlDat, ~genre, summarize, gender = "Both", rating = mean(rating))

mlDat\_gender <- ddply(mlDat, ~genre + gender, summarize, rating = mean(rating))

#mlDat\_gender <- rbind(mlDat\_gender, mlDat\_avgRating)

genderRatingPlot <- ggplot(mlDat\_gender, aes(genre, rating)) + geom\_histogram(stat="identity")

genderRatingPlot <- genderRatingPlot + facet\_wrap(~ gender)

**#fix axis**

genderRatingPlot <- genderRatingPlot + theme(axis.text.x = element\_text(angle = 90, hjust = 1))

genderRatingPlot <- genderRatingPlot + coord\_flip()

print(genderRatingPlot)

ggsave(filename = "genderRatingPlot\_f\_m.jpg")

**#### both with avg rating point**

mlDat\_avgRating <- ddply(mlDat, ~genre, summarize, gender = "Both", rating = mean(rating))

mlDat\_gender <- ddply(mlDat, ~genre + gender, summarize, rating = mean(rating))

mlDat\_gender <- rbind(mlDat\_gender, mlDat\_avgRating)

genderRatingPlot <- ggplot(mlDat\_gender, aes(genre, rating)) + geom\_histogram(stat="identity")

genderRatingPlot <- genderRatingPlot + facet\_wrap(~ gender)

#fix axis

genderRatingPlot <- genderRatingPlot + theme(axis.text.x = element\_text(angle = 90, hjust = 1))

genderRatingPlot <- genderRatingPlot + coord\_flip()

print(genderRatingPlot)

ggsave(filename = "genderRatingPlot\_f\_m\_both.jpg")

**############### heatmap ?**

mlDat\_genre\_occup <- ddply(mlDat, ~genre + occupation, summarize, mean\_rating = mean(rating))

mlDat\_genre\_occup <- droplevels(subset(mlDat\_genre\_occup, occupation != "homemaker"))

mlDat\_genre\_occup <- droplevels(subset(mlDat\_genre\_occup, genre != "unknown"))

mlDat\_genre\_occup <- droplevels(subset(mlDat\_genre\_occup, genre != "Fantasy"))

mlDat\_genre\_occup <- droplevels(subset(mlDat\_genre\_occup, occupation != "none"))

mlDat\_genre\_occup <- droplevels(subset(mlDat\_genre\_occup, genre != "War"))

mlDat\_genre\_occup <- droplevels(subset(mlDat\_genre\_occup, occupation != "doctor"))

heatMapPalette <- colorRampPalette(rev(brewer.pal(4, "RdBu")))

#get data ready for heatmap

goHeat <- ggplot(mlDat\_genre\_occup, aes(x = genre, y = occupation, fill = mean\_rating))

#rotate labels

goHeat <- goHeat + geom\_tile() + theme(axis.text.x = element\_text(angle = 90, hjust = 1))

#add colours

goHeat <- goHeat + scale\_fill\_gradientn(colours = heatMapPalette(100))

**#change background**

goHeat <- goHeat + theme(panel.background = element\_rect(fill='black'),

panel.grid.major = element\_blank(),

panel.grid.minor = element\_blank())

print(goHeat)

ggsave(filename = "genreOccupHeatMap.jpg")

mlDat\_genre\_occup <- ddply(mlDat\_multi, ~genre + occupation, summarize, mean\_rating = mean(rating))

#choose divergent colour tones to try and make distintions between like and dislike

heatMapPalette <- colorRampPalette(rev(brewer.pal(4, "RdBu")))

**#get data ready for heatmap**

goHeat2 <- ggplot(mlDat\_genre\_occup, aes(x = genre, y = occupation, fill = mean\_rating))

#rotate labels

goHeat2 <- goHeat2 + geom\_tile() + theme(axis.text.x = element\_text(angle = 90, hjust = 1))

#add colours

goHeat2 <- goHeat2 + scale\_fill\_gradientn(limits = c(2.5,4.6), colours = heatMapPalette(100))

**#change background**

goHeat2 <- goHeat2 + theme(panel.background = element\_rect(fill='black'),

panel.grid.major = element\_blank(),

panel.grid.minor = element\_blank())

print(goHeat2)

ggsave(filename = "genreOccupHeatMap\_multi.jpg")

**References**

[**http://tutorials.iq.harvard.edu/R/Rintro/Rintro.html**](http://tutorials.iq.harvard.edu/R/Rintro/Rintro.html)

[**https://cran.r-project.org/web/packages/recommenderlab/index.html**](https://cran.r-project.org/web/packages/recommenderlab/index.html)