Milestone_3_prep for M4_yujiaochen_brianho_jonjay

Jonathan Jay

4/19/2017

Looking ahead to Milestone 4, we've started preparing our dataset for CNN on movie poster files. EDA on the color of those posters is attached elsewhere in this submission. Here is how we've constructed our dataset for Milestone 4 (consistent with our plan from Milestone 2), with a few visualizations of the dataset.

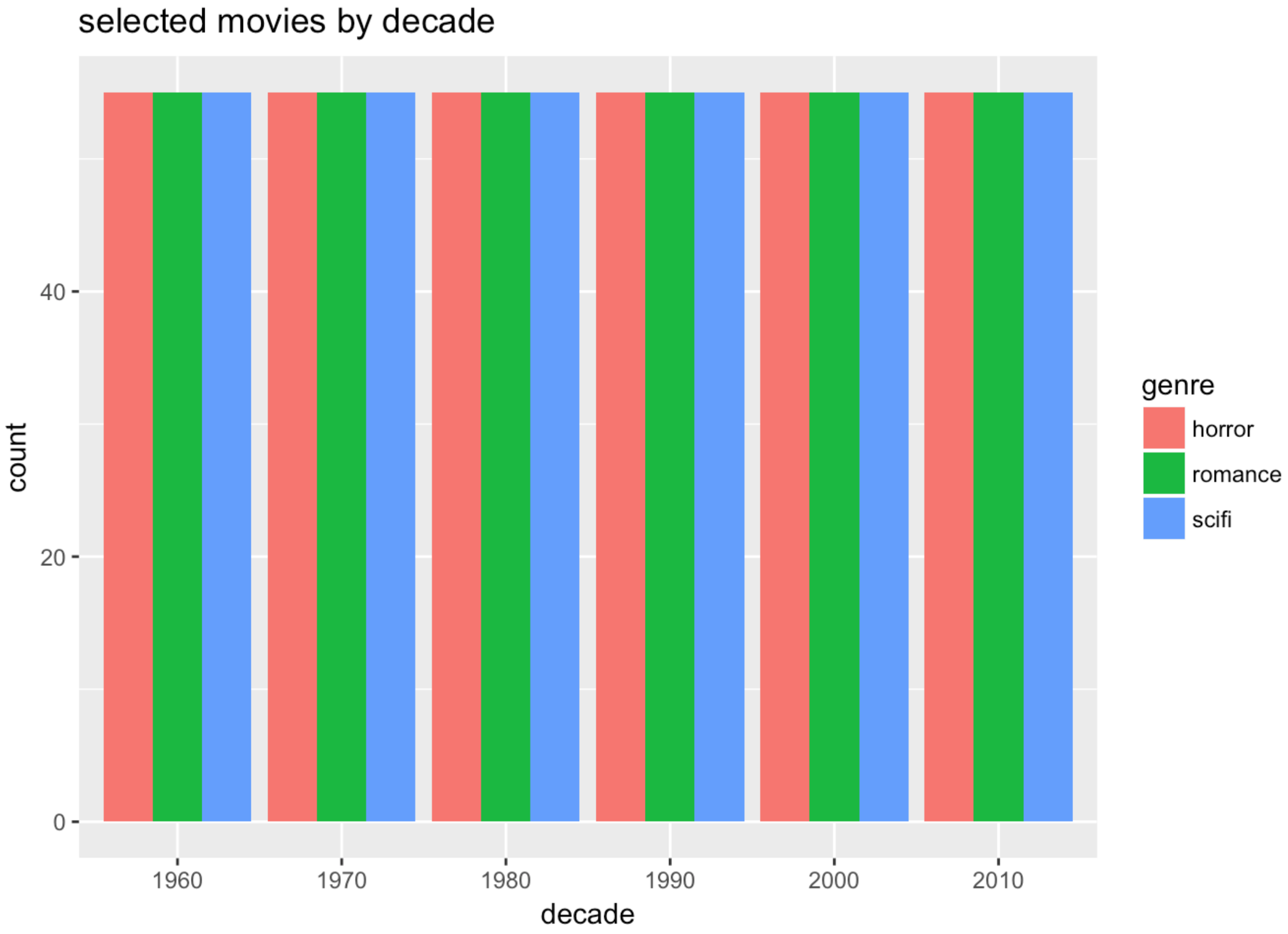
As discussed in Milestone 2, we aimed for balance across decades (to help control for time-related differences in movie poster design) and sought higher popularity, expecting that the posters for more-popular movies would be better representatives of their genre.

```
library(lubridate)
library(ggplot2)

movies <- read.csv("~/Documents/DrPH_open/Data Science 2/Final project/Milestone2/movies
_from_1930.csv")

#add year and decade fields
movies$date <- ymd(movies$release_date)
movies$year <- year(movies$date)
movies <- movies[movies$year >= 1960,] # most pre-1960 were romance, might have confused
 network
movies$decade <- (movies$year %/% 10) * 10


#filter by genre (only one of ours): romance = 10749, horror = 27, scifi = 878
# note: there weren't enough movies from only our genres, so I just dropped ones that ov
erlapped genres
movies$genres <- gsub("\\[|\\]", "", movies$genre_ids)

romance <- movies[grepl("10749", movies$genres), ]
romance <- romance[!grepl("27", romance$genres), ]
romance <- romance[!grepl("878", romance$genres), ]
romance$genre <- "romance"

horror <- movies[grepl("27", movies$genres), ]
horror <- horror[!grepl("10749", horror$genres), ]
horror <- horror[!grepl("878", horror$genres), ]
horror$genre <- "horror"

scifi <- movies[grepl("878", movies$genres), ]
scifi <- scifi[!grepl("27", scifi$genres), ]
scifi <- scifi[!grepl("10749", scifi$genres), ]
scifi$genre <- "scifi"

# filter to 330 of each genre, with 55 from each decade
horror <- horror[order(horror$decade, -horror$popularity),] #sort by decade, popularity
horror.ft <- data.frame()
#pull top 55 per decade
for(i in unique(movies$decade)){
  horror.ft <- rbind(horror.ft, horror[horror$decade==i, ][1:55, ])
}

romance <- romance[order(romance$decade, -romance$popularity),]
romance.ft <- data.frame()
for(i in unique(movies$decade)){
  romance.ft <- rbind(romance.ft, romance[romance$decade==i, ][1:55, ])
}

scifi <- scifi[order(scifi$decade, -scifi$popularity),]
scifi.ft <- data.frame()
for(i in unique(movies$decade)){
  scifi.ft <- rbind(scifi.ft, scifi[scifi$decade==i, ][1:55, ])
}
```

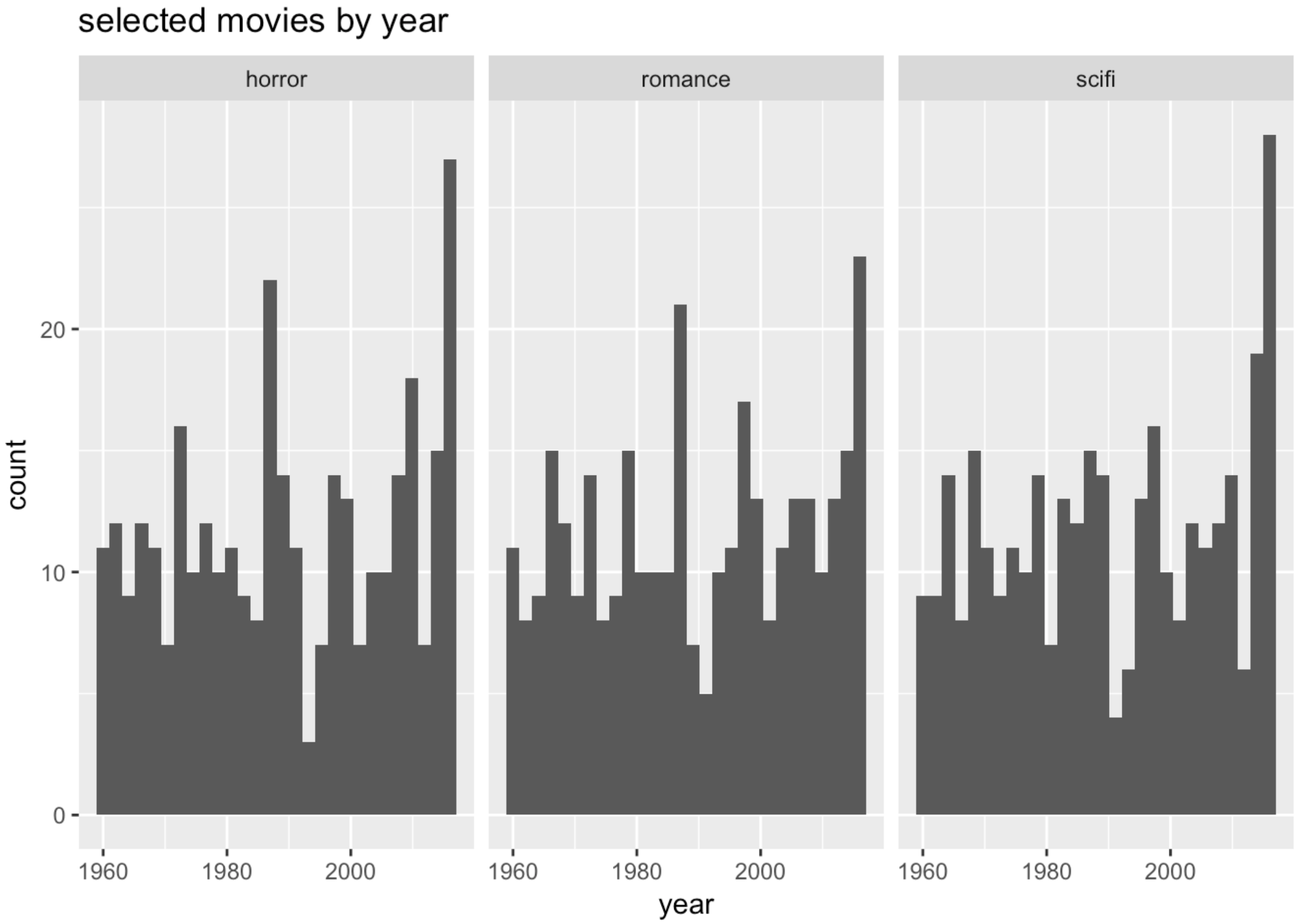
```
#combine filtered datasets and save
newmovies <- rbind(romance.ft, horror.ft, scifi.ft)

write.csv(newmovies, file = "Movie subset for poster analysis_990 movies.csv")
```

Some visualizations of the dataset makeup:

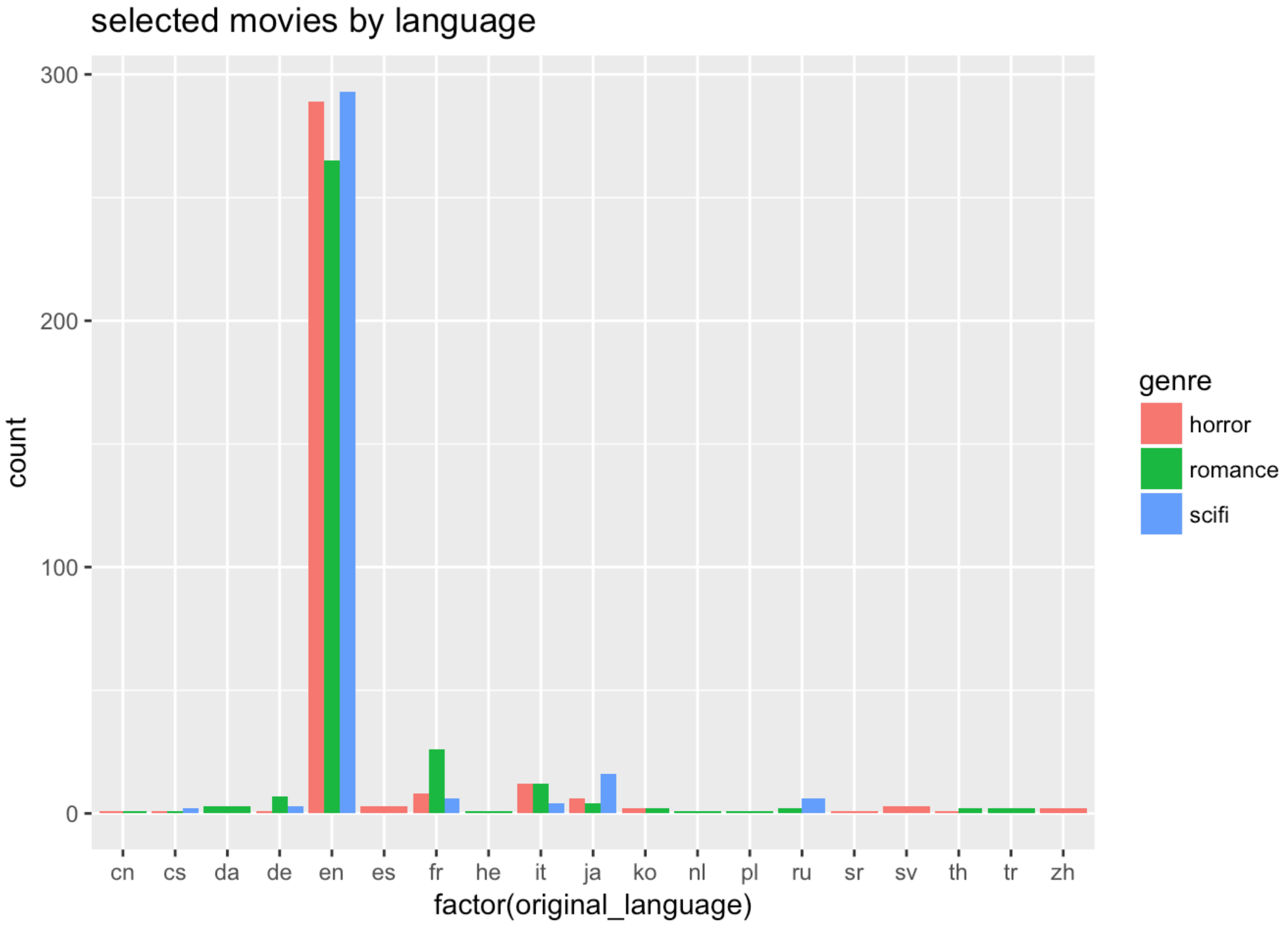
```
ggplot(newmovies, aes(x=factor(decade), fill=genre)) + geom_bar(stat="count",
position="dodge") +
  labs(x="decade", title="selected movies by decade")
```



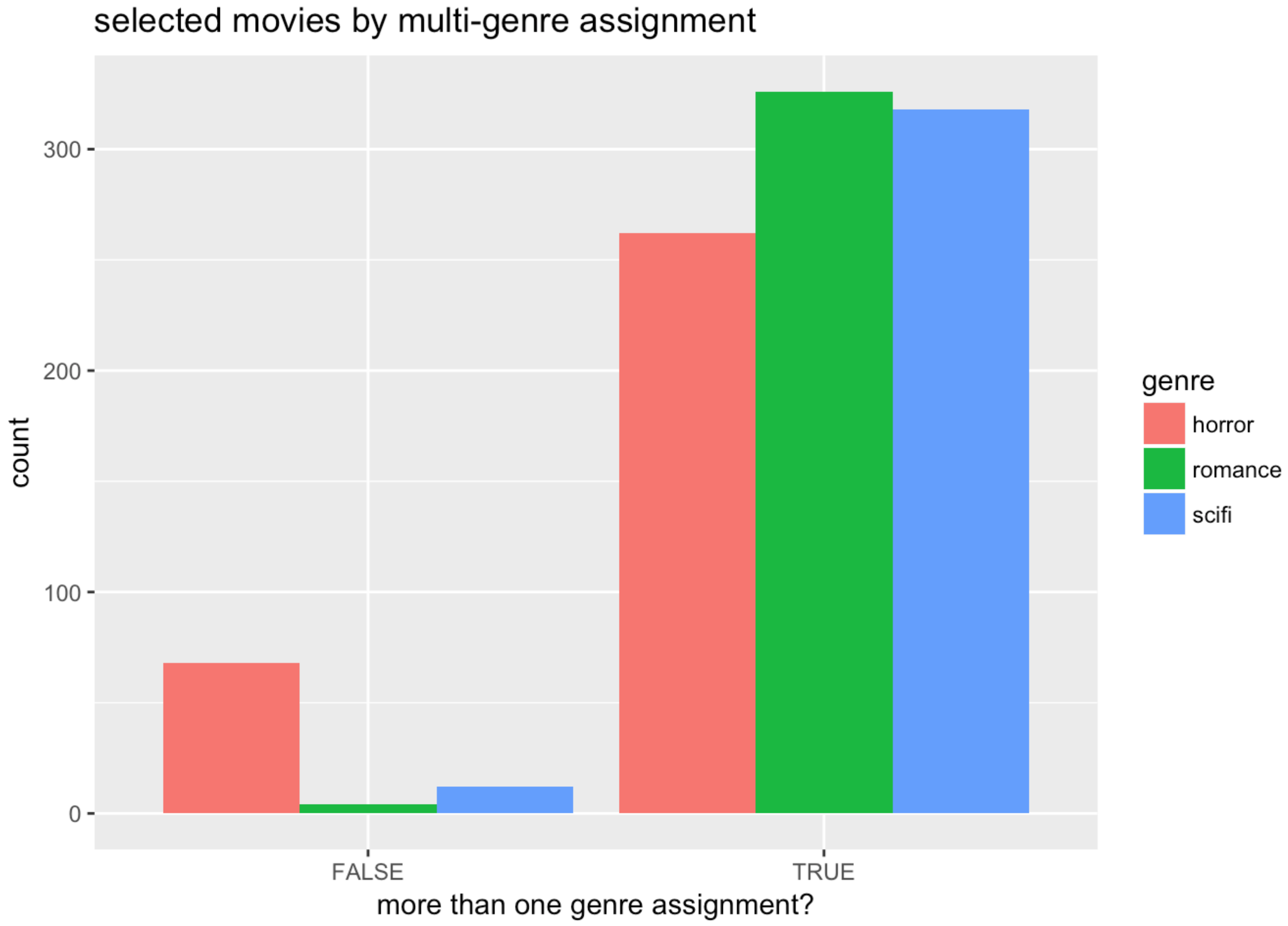
```
ggplot(newmovies, aes(x=year)) + geom_histogram(bins=28) + facet_wrap(~genre) + labs(tit
le="selected movies by year")
```



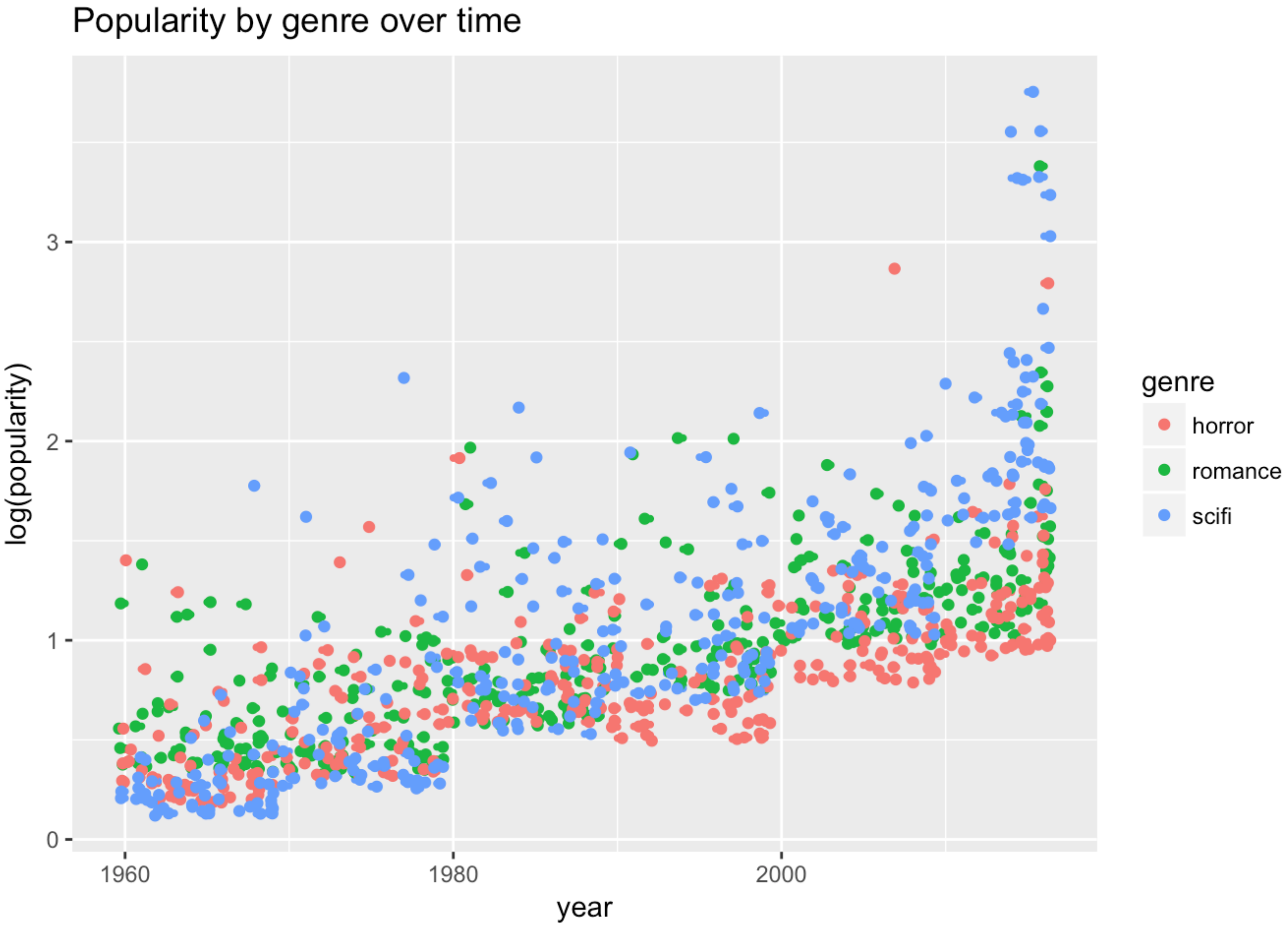
```
ggplot(newmovies, aes(x=factor(original_language), fill=genre)) + geom_bar(stat="count",
 position="dodge") +
  labs(title="selected movies by language")
```



```
ggplot(newmovies, aes(x=grepl(",", newmovies$genres), fill=genre)) + geom_bar(stat="coun
t", position="dodge") + labs(x="more than one genre assignment?", title="selected movies
 by multi-genre assignment")
```



```
ggplot(newmovies, aes(x=year, y=log(popularity), color=genre)) + geom_point(size=0.7) +
geom_jitter() +
  labs(title="Popularity by genre over time")
```



We can see the dataset is balanced by genre and decade. A large majority of movies have more than one genre assignment (though, by stipulation, none has more than one of our genres of interest). The general trend in popularity is rising over time, although it appears that scifi went from trailing horror to surpassing it around 1990.