DAT 301 LAB 2

(due Tuesday, Oct 20)

0. Make sure you can create a pdf file using "Knit" button in R-Studio. You need to create a new Rmarkdown file (`File --> New File --> R Markdown...`). This file should then be knitted to create the corresponding pdf file. You need to submit **both** Rmd and pdf file. **Important!** You get 0 credit if your submitted Rmd file does not compile OR does not create the submitted pdf file. Also, you need to include all the code chunks that preduce answers to all the questions.

If you have any trouble, we can discuss it in class, or office hours. If you use any library (you should, at least `dplyr`), you need to load it in your file. Do NOT use `install.packages()` command!

In this lab you will explore data about movies collected from http://imdb.com/

These data are stored in the data set `movies` in the package `ggplot2movies`. So, install this package first by typing `install.packages('ggplot2movies')` in the console. Do **not** include this command in your Rmd file. Then, load the data (these should be the first lines of code in your submitted Rmd file)

```
library(ggplot2movies)
data(movies)
```

To see the documentation for this data set, either go to https://cran.r-project.org/web/packages/ggplot2movies/ggplot2movies.pdf or in the RStudio console type

```
?movies
```

Note the meanings of the columns. A data frame with 58788 rows and 24 variables

- title. Title of the movie.
- year. Year of release.
- budget. Total budget (if known) in US dollars
- length. Length in minutes.
- rating. Average IMDB user rating.
- votes. Number of IMDB users who rated this movie.
- r1-10. Multiplying by ten gives percentile (to nearest 10%) of users who rated this movie a 1.
- mpaa. MPAA rating.
- action, animation, comedy, drama, documentary, romance, short. Binary variables representing if movie was classified as belonging to that genre.

Questions to Answer

1. What is the range of years of production of the movies of this data set (i.e. what is the year of production of the oldest movie and of the most recent movie in this data set)?
2. What proportion of movies have their budget included in this data base, and what proportion doesn't? What are top 5 most expensive movies in this data set?
3. What are top 5 longest movies?
4. Of all short movies, which one is the shortest (in minutes)? Which one is the longest? How long are the shortest and the longest short movies?
5. How many movies of each genre (action, animation, comedy, drama, documentary, romance, short) are there in this data base? (use a bar plot)

6. What is the average rating of all movies within each genre? (use a bar plot)

7. What is the average rating of all movies within each genre that were produced in the years 2000-2005?

8. For each of the first 6 genres (not including short movies) consider only movies from 1990 until the last year recorded and plot a function of the number of movies in this data base of corresponding genre produced by year, for years from 1990 until the last year recorded. For each of the 6 genres you should have one curve, and plot all the curves in the same figure. Naturally, use different colors, and appropriate legend.

9. Finally, formulate 3 questions of your choice related to this dataset and answer them. At least one of the answers should include some plot. Use any kind of plot system (base, plotly, or ggplot2). Impress me and the grader with your answers!