Part 1. Homework 1: Linear Models

Write a report that contains the results of the computations that you are asked to carry out below, as well as the explanation of what you are doing. The main text should include pieces of source code and graphical and numerical output. Upload your answers in a .pdf document (use LaTeX or R Markdown, for instance), as well as the source code (*.R or *.Rmd, for instance). Your work must be reproducible.

1. Linear Model for IMDB data

The IMDB dataset contains information of 940 films between 2000 and 2016. The data has been obtained form the `http:\\www.imdb.com` webpage. This is the list of the available variables:

```
1. movietitle
2. gross: in dollars
3. budget: in dollars
4. duration: in minutes
5. titleyear: 2000-2016
6. directorfl: Director Facebook likes
7. actor1fl: Actor 1 Facebook likes
8. actor2fl: Actor 2 Facebook likes
9. actor3fl: Actor 3 Facebook likes
10. castfl: Cast Facebook likes
11. facenumberinposter: Number of faces that appears in the poster
12. genre: Action/Comedy/Drama/Terror
```

The `titleyear` variable must be substituted by a a categorial variable (`yearcat`) with 3 levels: 2000-2005, 2006-2010 and 2011-2016. The idea is to predict how much the film will **gross** by knowing its characteristics.

1. Do the Exploratory Data Analysis for this dataset. Explain the most interesting conclusions.
2. With the original variables, fit the complete model including as predictors all the numerical variables, the two categorical variables and the interaction between numerical-categorical and categorical-categorical.
3. Use the stepwise procedure, by using the BIC criterion, to select the significant variables
4. Check the presence of multicollinearity. If there is some non-interaction multicollinearity in the model, make the corresponding corrections.
5. Validate the model by checking the assumptions
6. Interpret the final model