ProcessingData.R

- Add 2 means: zipcode\_mean\_ratings\_train & occupation\_mean\_ratings\_train

- Remove rows with variable "Unknown" == TRUE (Lots of NA in both Validation and Train)

- Remove the variable "Unknown"

- Remove the variable "age" (Already have age\_band)

- Remove the variable "video\_release\_date" (All NA)

- Remove the variable "imdb\_url" (Can not do analysis with URL)

- Reorder the columns in the following order:

* First 3: user\_id, item\_id, rating
* Ratings by users (1 to 5)
* Ratings by IMDB (1 to 10)
* Vote numbers
* Genres
* User information: gender, occupation, zip\_code, age\_band
* Movie information: timestamp, movie\_title, release\_date, item\_imdb\_mature\_rating, item\_imdb\_length

- Create new variables: realease\_date\_year, release\_date\_month, timestamp\_month (No year for timestamp because it has only 2 years with balance counts)

- Remove variables: release\_date, timestamp (Extracted the year and month out)

- Remove variable: movie\_title (Do not know how and what to do with texts)

**Need to do list:**

- Find a way on how to deal with missing values (imputation)