INTRODUCTION TO MACHINE LEARNING (NPFL054)
A template for Homework #1

Name: Zouhar Vilém

School year: 2nd

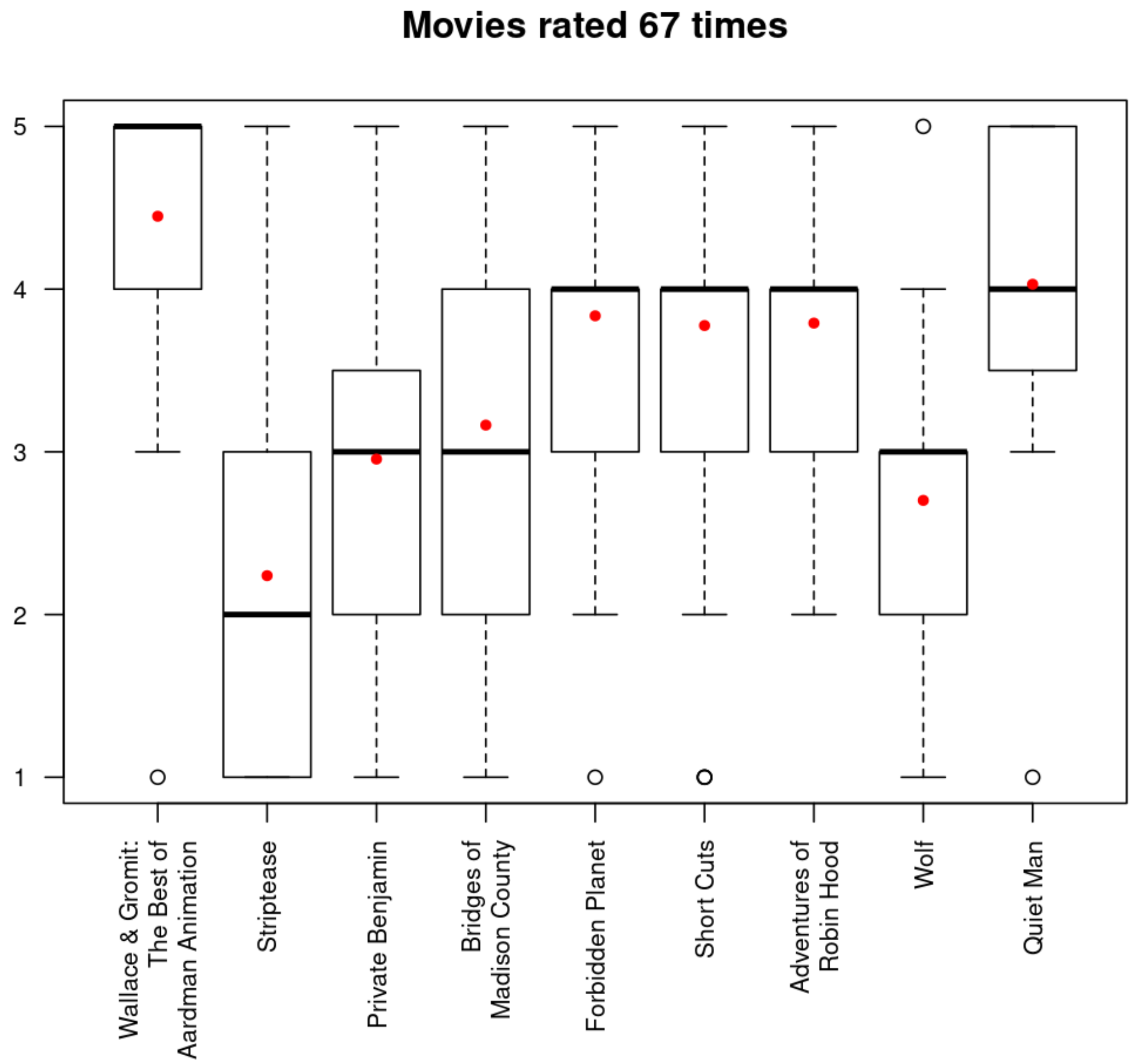
- **Provide answers for the exercises (1) - (3).**
- **For each exercise, your answer cannot exceed one sheet of paper.**

1. Conditional entropy [1pt]

H(OCCUPATION|RATING) = H(OCCUPATION, RATING) – H(RATING) =

= 5.877486 – 2.117503 = 3.759984

2. Boxplots of ratings of the movies rated 67 times [2pt]



The outliers (more than 1.5 IQR away from the median) are rather irrelevant in this case, because we have only 5 possible values and large data sample for each movie, so it is very probable, that the most bottom and top outliers will be 1 and 5 (regardless the movie quality, there will be some who will hate/love it). The smaller the IQR, the more concentrated the points are around the median. The mean is tended from the median in the opposite direction of most values. From this we can conclude, that:

Wallace & Gromit is very popular (small IQR and bottom whisker, high median)
Quiet Man is also very popular (similar reasons)
Short Cuts, Forbidden Planet and Adventures of Robin Hood are all similarly favored, although the last one hasn't received any 1 star rating (possibly because it received generally less ratings than the others)
Striptease is the least favored, because of the relatively small median, although there were some who liked it.

3. Clustering the users [7pt]

Cluster information:

```
   no_users   age
1       105 25.01
2        33 54.33
3        65 22.43
4        96 32.36
5        82 43.15
6        15 57.60
7        48 35.44
8       142 28.53
9        79 38.57
10       65 47.71
11       82 20.05
12        1  7.00
13       37 17.35
14       46 50.70
15       12 60.25
16       14 14.07
17       11 63.82
18        2 10.50
19        7 69.14
20        1 73.00
```

No duplicities found by comparing age, ratings and clusters. However, when utilizing solely the rating distribution, there were 3 pairs:

```
user age occupation    ONE  TWO  THREE FOUR FIVE cluss
139  20 student        0.00 0.04  0.21 0.46 0.29    11
147  40 librarian      0.00 0.05  0.10 0.45 0.40     9
558  56 writer         0.00 0.05  0.10 0.45 0.40     2
588  18 student        0.06 0.10  0.22 0.28 0.34    13
605  33 engineer       0.06 0.10  0.22 0.28 0.34     4
799  49 administrator  0.00 0.04  0.21 0.46 0.29    10
```