Assignment 02

Task 01:

In the lecture an example case of analysing the Movielense data was presented focusing on identifying the differences in highly rated movies based on females' and males' ratings. In this exercise you need to perform similar analysis but focusing on the variable age. You need to perform the following steps:

- Merge the three datasets (users, ratings, movies) the same way as it was done in the lecture example.
- Filter the resulting dataframe to keep ratings only from users who have at least 100 ratings available
- Filter the resulting dataframe to include information only on movies that have at least 200 ratings and that have genre 'Drama' or 'Comedy.
- Discretize the age column into four groups: (a) less than 22 years old, (b) between 22 and 33 years old, (c) between 33 and 42 years old, (d) older than 42 years.
- Determine and print the top 5 highest rated movies for the four groups separately.
- Count the number of movies that appear in the top 100 highest rated movies for all the age groups.

Task 02:

In this exercise you have to work with the data found in the file 'Credit.csv'. In the dataset you can find the following variables:

(a) Income
(b) Limit (credit limit of the person)
(c) Rating (credit rating)
(d) Cards (number of credit cards used by the person)
(e) Age
(f) Gender
(g) Student (whether the person is a student or not)
(h) Married (whether the person is married or not)
(i) Ethnicity
(j) Balance (positive values mean that the person has outstanding credit balance)

You have to perform some descriptive analysis tasks on this dataset:

- Visualization: create 4 plots of your choice; they can be histograms, boxplots etc. At least two of the plots should involve data from two separate columns (for example a boxplot of a column grouped by Student status or ethnicity, or a heatmap depicting the relationship of values in two columns).

- Data understanding: you need to focus on the column Balance, and try to understand what other variables have an impact on/are related to it. You are free to explore the data with any of the tools we used in the course, but include (i) some plots on the relationship between Balance and some other variables, (ii) calculate the correlation between Balance and other variables (when applicable). The output of this analysis should be a list of four variables that you identify to be related to Balance, with the choice of the variables justified using the results of your analysis.

Task 03:

In this exercise, you will have to continue working the the Credit dataset, and prepare it for further modeling through the following steps.

- Remove outliers: for each numeric column, remove the top 5% of values (0.95 quantile)

- Create a categorical version of Cards column with two categories:

  (i) users with more than 3, and (ii) users with at most 3 cards.

- Create a categorical version of Age with 4 categories. You may freely choose the intervals, but name the categories Very young, Young, Middle-aged and Elderly

- Scale the remaining numeric columns using the min-max transformation.

- Create dummy variables (one-hot encoding) for all the categorical columns.