1. Make some initial transformations of the set:

1.1. Remove from variable"world\_sales\_gross"commas and sign "$"

and convert it to a numeric one.

1.2. Add a variable named "Sci\_Fi" which is TRUE when "genres:" contains "Sci-Fi" and FALSE otherwise.

1.3. Add a categorical variable that will have 4 values:

- the "1975-1984" value for films that were made in the years 1975-1984,

- the "1985-1994" value for films that were made in the years 1985-1994,

- similarly for the periods 1995-2004 and 2005-2014

Name me "year\_cat"; make sure that this variable is of type "factor".

1.4. Add a numeric variable named "world\_sales\_mln", which should be equal to world\_sales\_gross divided by million; values ​​in the new variable should be rounded to one decimal place

1.5. How many missing data are there for the variable "studio"?

2. Perform the following analyzes:

2.1. Enter the mean, median, standard deviation, and range for "rt\_score".

2.2. How many categories are there in the "studio" variable?

2.3. Which movie (with which title) made the most money in the four time frames specified by the new "year\_cat" variable?

2.4. Create a correlation matrix for the variables: "rt\_audience\_score", "rt\_score", "imdb\_rating" and "length". Which variables are most strongly correlated?

2.5. Films from which label (the "studio" variable) are rated the best on average by imdb, and which label are the worst rated from?

Zad 3. Visualizations.

3.1. Create a bar chart with each bar representing one category of the variable "year\_cat" and the height of the bars representing the number of sci-fi movies in the dataset produced during that period (use the new variable "Sci\_fi"). Add the appropriate axis labels and "color" the bars in different colors.

3.2. Create a histogram for the variable "world\_sales\_mln". The bars should be blue with a black border. Give the chart a title "Blockbuster world sales" and add the appropriate axis labels.

3.3. Create a scatterplot for the variables "rt\_score" and "imdb\_rating". Color the points in the graph green.

Zad.4. Perform a regression analysis in which the explained variable will be "world\_sales\_gross" and the variables

explaining "rt\_audience\_score", "rt\_score" and "imdb\_rating" (in other words, we are looking for a regression function,

which will model the relationship between different movie ratings and the returns from that movie). Which explanatory variables are

related to the dependent variable, i.e. for which explanatory variables is the regression coefficient statistically significant?

How big is the standard deviation of the residuals and the coefficient R2? What do these numbers show about the quality of our model?