TASK 2

In the second task, we perform different methods of data visualization and data manipulation on which the data analysis based. First of all, we imported the necessary libraries in order to perform the data analysis. We created a scatter plot for which the correlation between the two variables on rank (y) and score (x) is negative and equals to -0,9761. Then, we decided to group the anime according to their “type” in a pie chart, where the largest percentage of values belong to the TV category. Accordingly, we classified on another plot the average number of episodes grouped by type which again shows that the majority is TV. Afterwards, another scatterplot based on popularity and likes showed that the most popular anime are also the least liked ones.

Moreover, the regression model was made by plotting distribution of target variable “likes”. Firstly, we checked whether there was a presence for null values. We looked for instances with null values in the rank column and deleted the instances with null values, respectively “rank”, “score” and “scored\_by”. We decided a number of predictors and a target variable (“likes”) to create the regression model and normalized the distribution of the target variable likes. After, we performed a train-test split where 25% is for the test group and 75% for the train group. The first multivariate linear regression had a R squared equal to 0.7360. The second regression model includes a Random Forest regressor. Since we only identified a few thousands of observations, we used cross validation that increased to R squared from 0.8155 to 0.8329. The last regression model was developed with the Extreme Gradient boosting, that had R squared equal to 0.8120. This also required cross validation, which gave a R squared of 0.8102. The model that reflects the highest R squared is therefore corresponding to the second linear regression.