TASK 2

For the second part of the project, we imported the dataset from the previous task. With the anime dataset, we performed 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 explored the dataset through multiple graphs and analyzed the data by observing the correlation between variables. Hence, we created a scatter plot for which the correlation between the two chosen variables on rank (y) and score (x) is negative and equals to -0,9761, which reflects a high correlation.

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 to identify the average number of episodes grouped by type. As expected, the TV series have the higher number of episodes.

Afterwards, we made two different models using numerical variables from the dataset in order to predict the number of likes. Therefore, we decided to make a scatterplot that shows how the variable “likes” is correlated with “popularity” and “scored\_by”. From the first graph, we discovered that the most popular anime are also the least liked ones. On the other side, the second scatterplot is more linear.

Before we created the models, we plotted the distribution of the target variable “likes”. Moreover, 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”, that are not considered numerical variables. Then, we decided a number of predictors and a target variable (“likes”) to create the regression model and normalized the distribution of the target variable through the standard scaler function.

After, we performed a train-test split where 25% is for the test group and 75% for the train group. We created three different models that range from one more interpretable and less precise to another that is more precise but less interpretable. The first multivariate linear regression had a R squared equal to 0.7360. The second regression model includes a Random Forest (RF) 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 (Xgboost), 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, as reflected by the table below, corresponding to the second linear regression with cross validation.

|  |  |
| --- | --- |
| Type of regression model | R squared |
| Multivariate linear regression | 0.7360141405427469 |
| RF | 0.8154562187061573 |
| RF with cross validation | 0.8328898966127319 |
| Xgboost | 0.8119881157925052 |
| Xgboost with cross validation | 0.8102276182949104 |