- What types of data cleaning or transformation did you perform?

For numerical features:

- Delete the rows where “year” have values such as “N.V.” or “nan”; and replace “year” by constructing the new feature named “duration”, which indicates how long it has been since the wine harvested (duration=2022-year)
- Standardize the numerical features named “num_reviews”, “price”, and “duration” using log()

For categorical features:

- Change the data type of “body”, “acidity”, “rating” from float to string, as they are discrete rather than continuous values; therefore, this problem is regarded as an unbalanced multi-classification problem instead of a regression problem
- Impute the missing values of “body” and “acidity” using the most frequent values
- Drop the “country” feature since it only owns unique value and is not informative
- Bin the features such as “winery” and “region” based on total numbers of ratings they have
- Address the above categories using get_dummies methods

- What process did you take to build your predictive model?
- Import data and conduct exploratory data analysis
- Clean and transform the raw data
- Split the data into training set and test set
- Train the classifier using the training set and evaluate its generalization ability (calculate weighted AUC) using test set
- Get the ranked feature importance table

- What factors / features were predictive of wine quality?

- Number of reviews
- Price
- Duration

- What models and techniques did you test?

- Random Forest & Weighted AUC

- What were your results? Would you feel comfortable productioning this model, why or why not?

- We can make prediction for ratings of wines based on their historical data of number of reviews, prices, and duration
- No. These data only comes from the Spanish, it may not suit well for other countries.

- If you had more time, what other type of data would you want or techniques would you explore to improve your model fit?

- I would like to conduct dimension reduction, remaining most (80%) percent of information, but reducing dimension of the features.
- I may also split the training set into the training set and the validation set, and tune the hyper parameters of the random forests, such as the max depth of the random forest.
- Other features such as “text of reviews“, “detailed descriptions of wines” can also be informative, as we can use NLP techniques such as tf-idf to decide whether “some words” may be representative regarding the quality of the wine.

● Finally, what recommendation would you make to the CEO on the types of wines to target to maximize quality?

- I would suggest CEO to focus more on features such as “number of reviews”, “price”, “duration” to predict the quality of wine.
- From the feature importance table, we may also find the Rioja Red > Ribera Del Duro Red > Red > Toro Red > Priorat Red > Grenache > Verdejo > …