

# CORE: Exploring and understanding the Data

no more than 2.5 pages!!!!!!!

•(20 marks) Highlight the findings of your dataset exploration. You should identify four important patterns (e.g. large correlation between variables), and discuss the potential consequence this may have on your results. To achieve a high mark, you should consider more complicated patterns, such as feature interactions. Use your judgement and justify what is an important pattern.

Due to there are lots of datasets in there, so the first thing that I do is to merge them into a whole dataset called merged\_dataset except for the testing-instances dataset. Then, I use pandas-profile to generate the report based on this whole merged\_dataset. By observing the report and googling the definition for each field( [1] and [2] ), for feature columns, I find that although there are no missing values by using python, but there are some potential missing values for several features. The initial findings for each feature are described at below:

- 'instance\_id' is unique, which means the percentage of distinct is 100%
  - For avoiding the over-fit issue during training, I decide to delete this column at start, since it is absolutely unnecessary feature.
- For features that has potential missing value: (on the right screenshot)
  - 'artist\_name' is the category variable which has empty\_field as the individual category to capture the missing values. ~~so leave it alone missing value~~
    - Through the observation, it is uniform distributed across all datasets(~ 5% missing for each), which cause the total missing percentage is less than 5% as well. Therefore, I think it is MCAR, which means rows can be deleted where artist\_name is missing, which is the operation that I do on this feature.
  - Undoubtedly, 'duration\_ms' is a numerical feature that values should always be positive. However, there are about 10% rows that has -1 as its value, which indicates they are missing.
    - The distribution of missing also is the same as previous, which means it is MCAR as well. However, from slide14, I obtain that Deletion approach should only be applied when it is MCAR and missing percentage is less than 5%.
    - So, I choose to use the Imputation approach, which try to use the global mean value to fill the NA.
  - From [1] and [2], I know that 'tempo' should be the numerical feature, but due to it has '?' represents the missing value, so python recognize it as the categorical feature. It is also the MCAR which is the same reason as previous. There are also about 10% missing.
    - The same as previous 'duration\_ms' one, I choose to use global mean to fill the NA
- For the rest of features, there are no potential missing values.
- 'track\_hash', 'track\_name', 'popularity', 'acousticness', 'danceability', 'energy', 'instrumentalness', 'key', 'liveness', 'loudness', 'mode', 'speechiness', 'obtained\_date', 'valence', 'music\_genre'
- 

```
missing feature and values:
artist_name    2489
duration_ms    4939
tempo          4980
dtype: int64
```

•(20 marks) Visualisation is an important aspect of this task. Please illustrate at least one important finding of your work using visualisation. For full marks, you should be expected to use more than a simple scatter plot.

## Completion: Developing and testing your machine learning system

You should refine your machine learning system a number of times (at least 3, including the initial system) based on the performance you achieve on the public leaderboard. Your submitted report should contain up to 4 pages regarding the Completion component

(15 marks) Discuss the initial design of your system, i.e. before you have submitted any predictions to the Kaggle competition. Justify each decision you made in its design, e.g. reference insight you gained in the Core part.

My initial design

(25 marks) Discuss the design of one or more of your intermediary systems. Justify the changes you made to the previous design based on its performance on the leaderboard, and from any other additional investigation you performed.

(10 marks). Use your judgement to choose the best system you have developed — this may not necessarily be the most accurate system on the leaderboard. Make sure you select this submission as your final one on the competition page before the deadline. Explain why you chose this system, and note any particularly novel/interesting parts of it. You should submit screen captures and/or the source and executable code required to run your chosen submission so that the tutors can verify its authenticity.

## Challenge: Reflecting on your findings

Until now, we have been focusing on achieving the best performance possible — but there should be some other aspects that ML tool users should consider, e.g. the interpretability of the model.

You should consider the interpretability of your final chosen model in this part. Your report ( **1 page** on the Challenge component) should address the following questions:

- (10 marks) How easy is it to interpret your chosen machine learning model, i.e. how easy to comprehend why certain predictions have been made?
- If your model is difficult to interpret, do you see any problems with this? (e.g. whether users will trust your model? whether it is difficult for the deployment of your solution or to use the model? and so forth.) How would it compare to a simpler model, e.g. a simple K-Nearest neighbour?
- If your model is easy to interpret, what are its limitations? (e.g. whether it can catch the underlying relationship in the data? whether it can provide accurate predictions?) How would it compare to a more complex model, such as a ensemble method (e.g. random forest)

## Reference:

1. <https://towardsdatascience.com/is-my-spotify-music-boring-an-analysis-involving-music-data-and-machine-learning-47550ae931de>
2. <https://rpubs.com/PeterDola/SpotifyTracks>
3. <https://stackoverflow.com/questions/13413590/how-to-drop-rows-of-pandas-dataframe-whose-value-in-a-certain-column-is-nan>
  - o It contains everything about different ways of delete NA
- 4.