Un premier projet de bout en bout

Après avoir lu le chapitre 2 du livre *Hands-On Machine Learning with Scikit-Learn & TensorFlow*, nous vous demandons de mettre en pratique la matière de ce chapitre avec un nouveau jeu de données contenant des informations sur des kits Légo.

Voici quelques informations sur ce jeu de données.

# Context

Have you ever wondered what is the most expensive lego set in the world? Or how many pieces does the Taj Mahal set contain? What about, how these features might interact with each other? I wanted to answer some of these questions and more, so I scraped lego set and price information from Lego's website.

# Content

This dataset contains lego sets scraped from lego.com. Each observation is a different Légo set and there are features like how many pieces are in the set, how much the set sells for, etc. This dataset contains lego sets from all the different countries they sell online to (except South Korea).

# Train.csv

Vous trouverez sur mooVin un fichier « train.csv », ce fichier contient des informations à propos de 9195 kits Légo. Chaque kit est décrit à l’aide des informations suivantes[[1]](#footnote-1) :

1. ages
2. num\_reviews
3. piece\_count
4. play\_star\_rating
5. prod\_desc
6. prod\_id
7. prod\_long\_desc
8. review\_difficulty
9. set\_name
10. star\_rating
11. theme\_name
12. val\_star\_rating
13. country
14. list\_price

# But de l’exercice

Le but de cet exercice est d’entrainer, à l’aide des informations contenues dans le fichier « train.csv », un premier « machine learning model » afin de prédire le prix de nouveaux kits Légo.

Le fichier « test.csv » contient les mêmes informations, à propos de 3066 kits Légo, que le fichier « train.csv ». Excepté la colonne « list\_price » qui n’est pas présente! Une fois votre modèle entraîné, c’est à vous de deviner le prix de ces 3066 kits.

# Travail à remettre

Pour le mardi 01/10/2019 à 23h59, ceux qui le souhaitent peuvent soumettre sur mooVin un fichier csv contenant pour chaque kit le prix prédit. Pour chaque kit, votre fichier csv contiendra une ligne reprenant uniquement le prix prédit. Les prix doivent évidemment apparaître dans le même ordre que les ligne du fichier « test.csv ». Nous évaluerons vos prédictions à l’aide du « Root Mean Square Error ».

1. Attention ce fichier n’est peut-être pas très « propre », pensez à le nettoyer [↑](#footnote-ref-1)