### About the Datasets

These datasets come from [Kaggle](https://www.kaggle.com/datasets) and have been upvoted quite consistently by the community. The [Apps Dataset](https://www.kaggle.com/lava18/google-play-store-apps) comes from web scraping the Google Play Store on 10,000 apps and 64,000 reviews – though it seems to have some repeated lines on it, which effectively makes it a 8,900 apps dataset. And the [Wines Dataset](https://www.kaggle.com/zynicide/wine-reviews/home) comes from web scraping the [WineEnthusiast](https://www.winemag.com/?s=&drink_type=wine) website – and it also suffers from repeated lines: it goes from 150,000 reviews to 90,000 ones. All this data offers valuable insights about the apps and wines markets, some of which we will discuss here.

In order to study these datasets, I've taken the time of refreshing my Statistics background with an incredible – he manages to make a very simple practical and easy to learn approach to the subject – Udemy course by Jose Portilla: [Probability and Statistics for Business and Data Science](https://www.udemy.com/probability-and-statistics-for-business-and-data-science/). The statistical tests discussed in the second half of the course will be useful to attest our findings. I've also made available my course notes and a mind map, along with my Python code, in this Github Repo.

### Paid Apps vs Free Apps

First, let's take a look at the general characteristics of the data. From the graphs below, histograms that show the percentage of apps for each range of rating, we can clearly see that the majority of the apps stand within the range of 4.0 to 4.5 ratings (a little bit more than 50%), and the apps that manage to get really good scores represent at best 7% of the market.

Now, when it comes to prices, this dataset is, as is reality, heavily skewed towards free apps, since they represent most of the Google Play Store. Effectively,