# **Group Members**

Campeny, Eloi

Chriki, Fatima Zohra

Dai, Zhongkai

González, Victor

Moure, Ximena

Xu, Ange

# **Data collection**

The dataset can be retrieved from this url: <https://www.kaggle.com/datasets/gauthamp10/google-playstore-apps>

It’s about a .csv file that contains 2.312.944 observations of 24 variables. As the dataset is quite large, we decided to reduce it to 20.000 observations.

The final dataset, as well as the script used to obtain it, can be found at:

<https://drive.google.com/drive/folders/1B-6fwHjjXAG-JSsAIGmQOihYE80M2_3m?usp=sharing>

# **Context**

Our data contains details about 20 thousand Android applications from the Google Playstore. Some important variables are the name of the app, the category it belongs to, the number of installs, whether it’s free or not, the size in Megabytes, etc.

Our goal is to find out which aspects can influence the downloads and rating of an app (our target variables in this dataset) and thus to help developers to better understand the current mobile applications market demand.

# **Basic structure of our data**

Our data has 20.000 records and 24 variables: 5 of them are numerical, 4 of them are binary, 15 of them are qualitative. Some of the qualitative variables are going to be transformed into numerical (*Installs, Size, Minimum.Android*), in order to accomplish the requirement of the minimum of 7 numerical variables.

Numerical variables

* Rating 344 missing 1.72%
* Rating.Count 344 missing 1.72%
* Minimum.Installs 1 missing 0.005%
* Maximum.Installs 0 missing 0%
* Price 0 missing 0%

Binary variables

* Free 0 missing 0%
* Ad.Supported 0 missing 0%
* In.App.Purchases 0 missing 0%
* Editors.Choice 0 missing 0%

Qualitative variables

* App.Name 0 missing 0%
* App.Id 0 missing 0%
* Category 0 missing 0% → Factor
* Installs 1 missing 0.005%
* Currency 1 missing 0.005% → Factor
* Size 0 missing 0% -> Numerical
* Minimum.Android 49 missing 0.245% → Factor
* Developer.Id 0 missing 0%
* Developer.Website 6097 missing 30.485%
* Developer.Email 0 missing 0%
* Released 750 missing 3.75% → Date
* Last.Updated 0 missing 0% → Date
* Content.Rating 0 missing 0% → Factor
* Privacy.Policy 3324 missing 16.62%
* Scraped.Time 0 missing 0%

There are 10911 missing values in this dataset, for a total of 480000 values. So, 2.27% of the values are missing.