**DANA Project Proposal**

**Team Members:**

* Aarushi Dhyani
* Andres Camilo Viloria Garcia
* Diana Ortiz
* Sameeksha Kiran Mendon

**DATA QUALITY**

1. Remove duplicates
2. Deal with the null values
3. Remove null values for all variables except rating
4. Remove foreign signs like $ from price and + from installs
5. Analyze the rating variable (because we have more than 1400+ null values)

- Divide the ratings for new and old apps

1. **Questions of interest**

1. How is the distribution of Rating? Can it be considered Normal distribution?

2. How is the distribution of prices? Can it be considered Normal distribution?

3. Which are the 5 most expensive apps in the play store dataset?

4. What is the correlation between Installs and Reviews?

5. Can we consider Type and Content Rating are strongly associated?

1. **Objectives:**

* **Population of interest:** All apps available on Google Play Store.
* **Sampling frame: All** apps are available to download on Google Play Store in the zone where the data was collected (unknown)
* **Sampling design:** Not random sampling
* **Sample size:** 10841 records.

1. **Variables:**

* **Categorical:**
* **Genres:** An app can belong to multiple genres (apart from its main category). For e.g., a musical family game will belong to.
* **Category:** Category the app belongs to
* **Type:** Paid or Free
* **Content Rating:** Age group the app is targeted at - Children / Mature 21+ / Adult.

* **Numerical:** 
  + **Price:** Price of the app (as when scraped)
  + **Size:** Size of the app (as when scraped)
  + **Rating:** Overall user rating of the app (as when scraped)
  + **Installs**: Number of user downloads/installs for the app (as when scraped)

Source of the Data Set:

<https://www.kaggle.com/datasets/lava18/google-play-store-apps>