

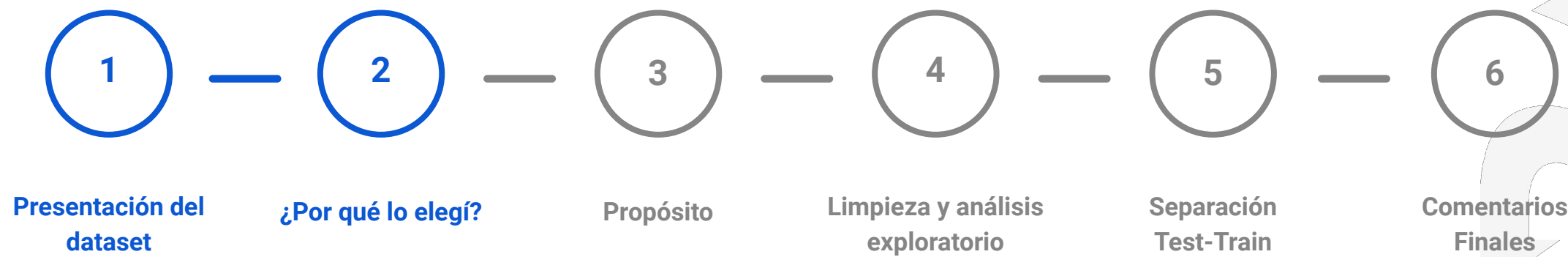


Instituto Tecnológico
de Buenos Aires

APPS EN GOOGLE PLAY STORE

Análisis Predictivo - 1Q 2022 - ITBA

Camila Collado



DATASET



Este dataset proviene de **KAGGLE** (también está en otros sitios como [OpenML](#) y [Github](#))

Contiene datos sobre **2.3M** de aplicaciones en Google Play Store

Fecha **JUNIO 2021**

PROPÓSITO

En Google Play Store hay una gran cantidad de apps que cumplen todo tipo de funciones.

A la hora de decidir que aplicación descargar la **calificación** es un dato importante para los usuarios.

Las calificaciones son brindadas por los usuarios pero podemos intentar anticiparnos a su pensamiento.



LIMPIEZA DEL DATASET

24 variables, 2.312.944 filas

VARIABLES QUE NO APORTAN INFORMACIÓN

Developer.Website

Developer.Email

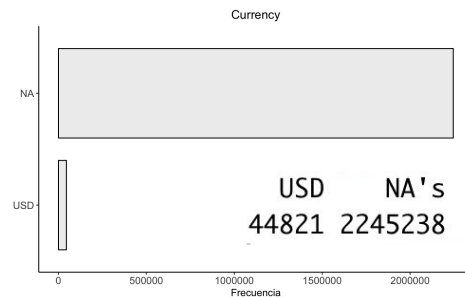
Privacy.Policy

Scraped.Time

App.Name ————— 199385 repetidos

Installs —————

Currency —————



VARIABLES MODIFICADAS

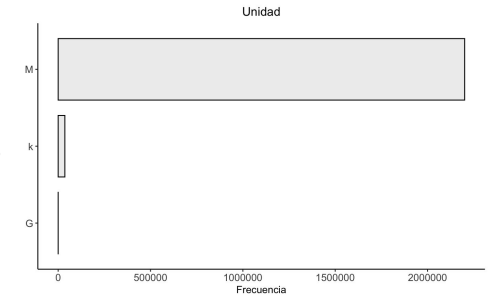
Minimum.Installs	Installs
<dbl>	<chr>
10	10+
5000	5,000+
50	50+
10	10+
100	100+
50	50+

Size

Released

Last.Updated

Unificar unidad de medida: MB
Cambiar “Varies with device”
por la mediana



Quitar vacíos, cambiar a
formato fecha y filtrar los
registros con
Last.Updated > Released.
Más adelante se cambió
por una resta de días.

LIMPIEZA DEL DATASET

16 variables, 1.196.055 filas

VARIABLES CATEGÓRICAS

Quitar vacíos y NA

Category

Minimum.Android

Developer.Id

Content.Rating

• Cambiar los registros de "Varies with device" por la moda

• Filtrar la categoría "Unrated"

VARIABLES LÓGICAS

Quitar NA y cambiar a tipo de dato lógico

Free

Ad.Supported

In.App.Purchases

Editors.Choice

VARIABLES NUMÉRICAS

Quitar NA

Rating

Rating.Count

Minimum.Installs

Maximum.Installs

Price

ReleasedDA

Last.Updated.DA

Size.mb

• > 0

Rating= 0 (46,25%)

INSIGHTS

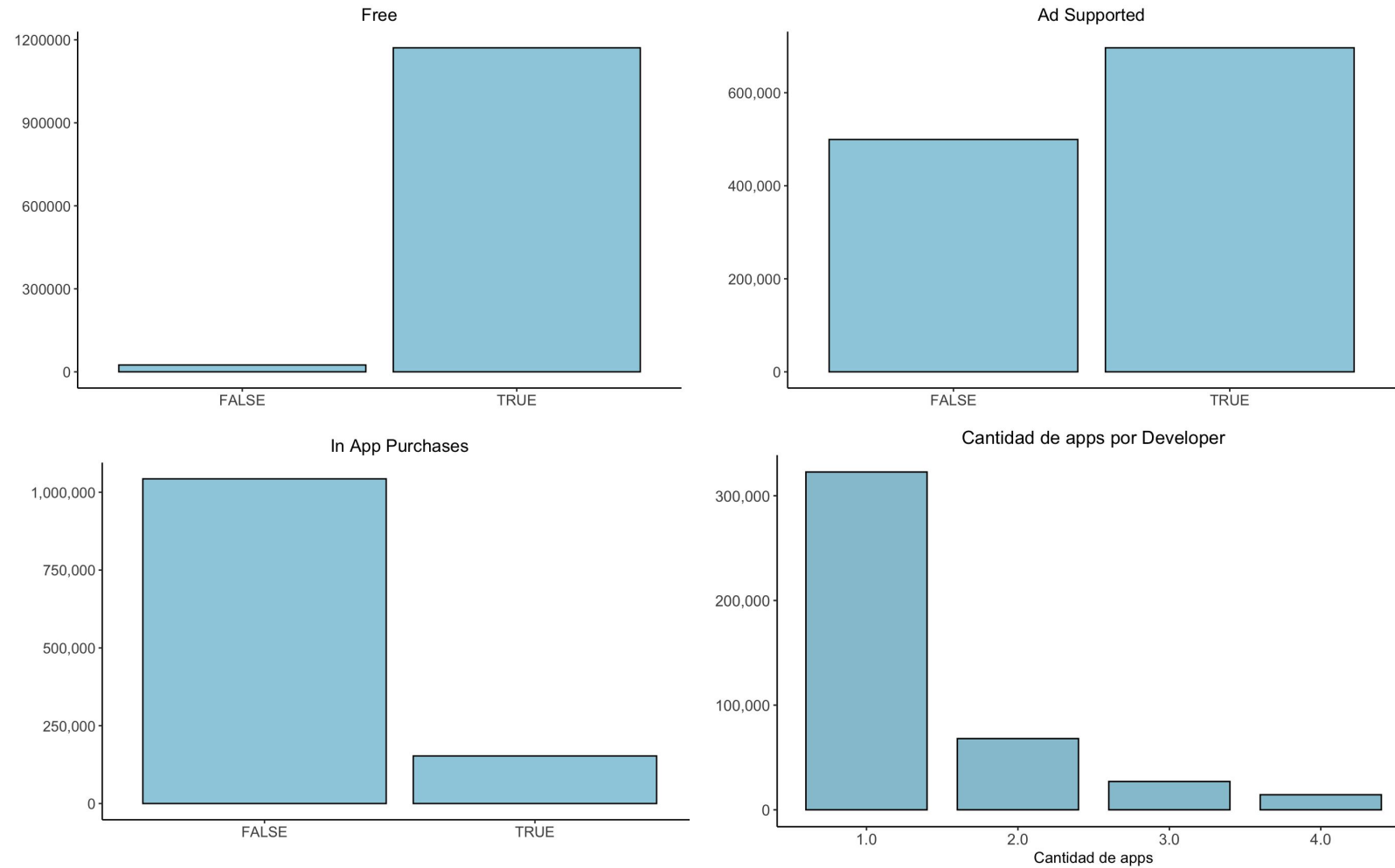
LA MAYORÍA DE LAS APPS ESTÁN
DISPONIBLE PARA **ANDROID 4.1**

SOLO EL **13,7%** DE LAS APPS
TIENE **RESTRICCIÓN DE EDAD**

TOP 5 CATEGORY

1. EDUCATION
2. TOOLS
3. ENTERTAINMENT
4. MUSIC & AUDIO
5. BOOKS & REFERENCE

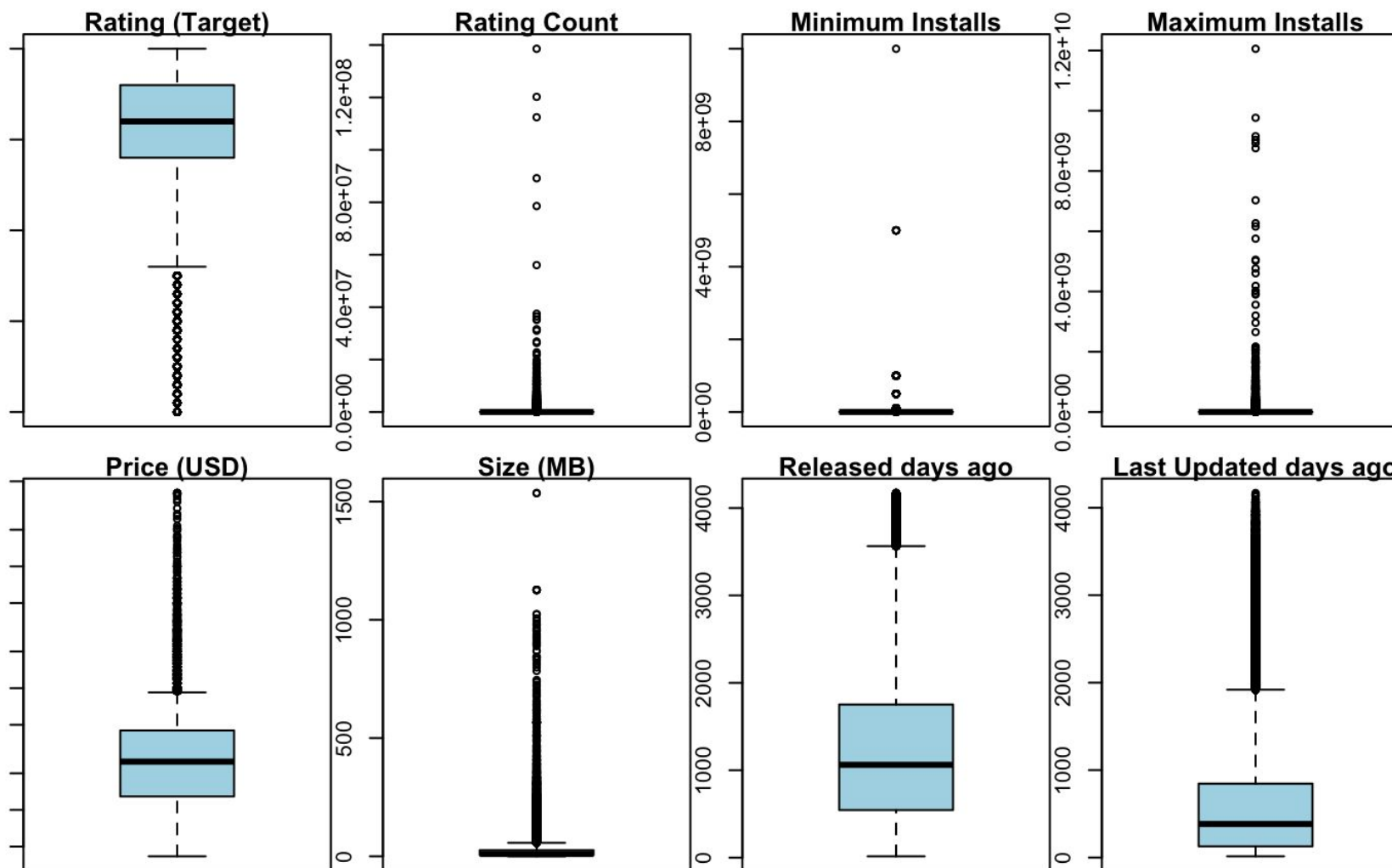
INSIGHTS



*sin outliers

OUTLIERS

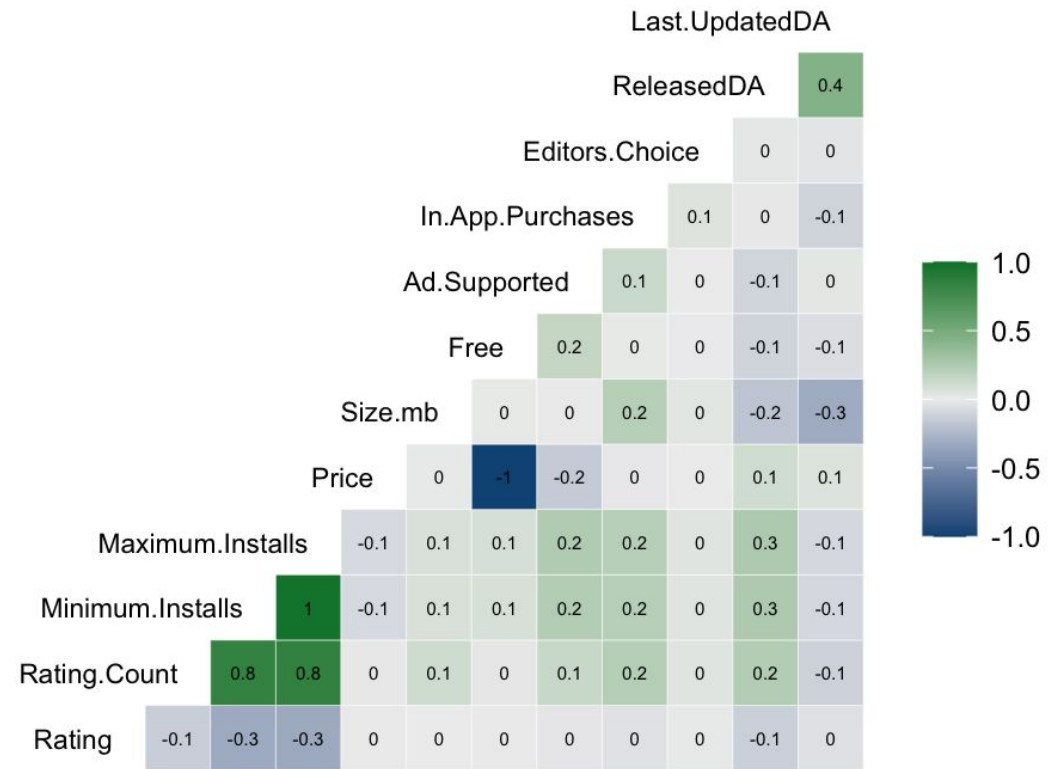
VARIABLES NUMÉRICAS



*log=Price

CORRELACIONES

PARA VARIABLES NUMÉRICAS Y
CATEGÓRICAS

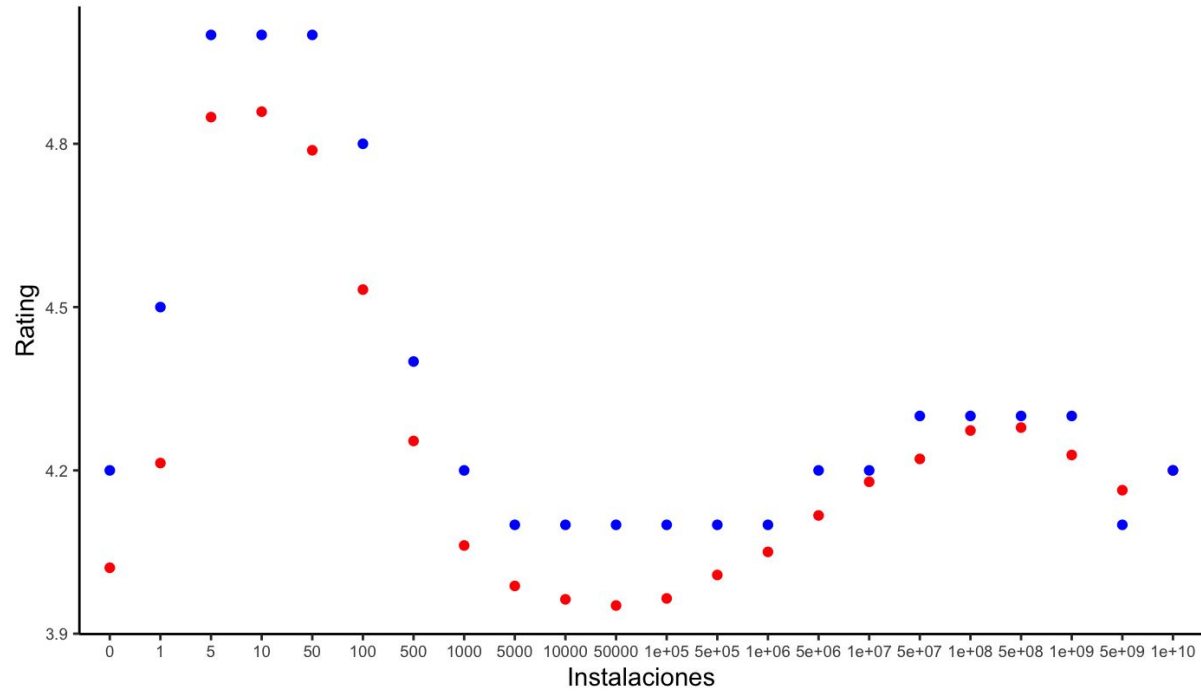


	Category	Minimum.Android	Developer.Id	Content.Rating
Category	1.0000000	0.0724406	0.8277207	0.2766208
Minimum.Android	0.0724406	1.0000000	0.2766208	0.0381142
Developer.Id	0.8277207	0.8191162	1.0000000	0.8004715
Content.Rating	0.2766208	0.0381142	0.8004715	1.0000000

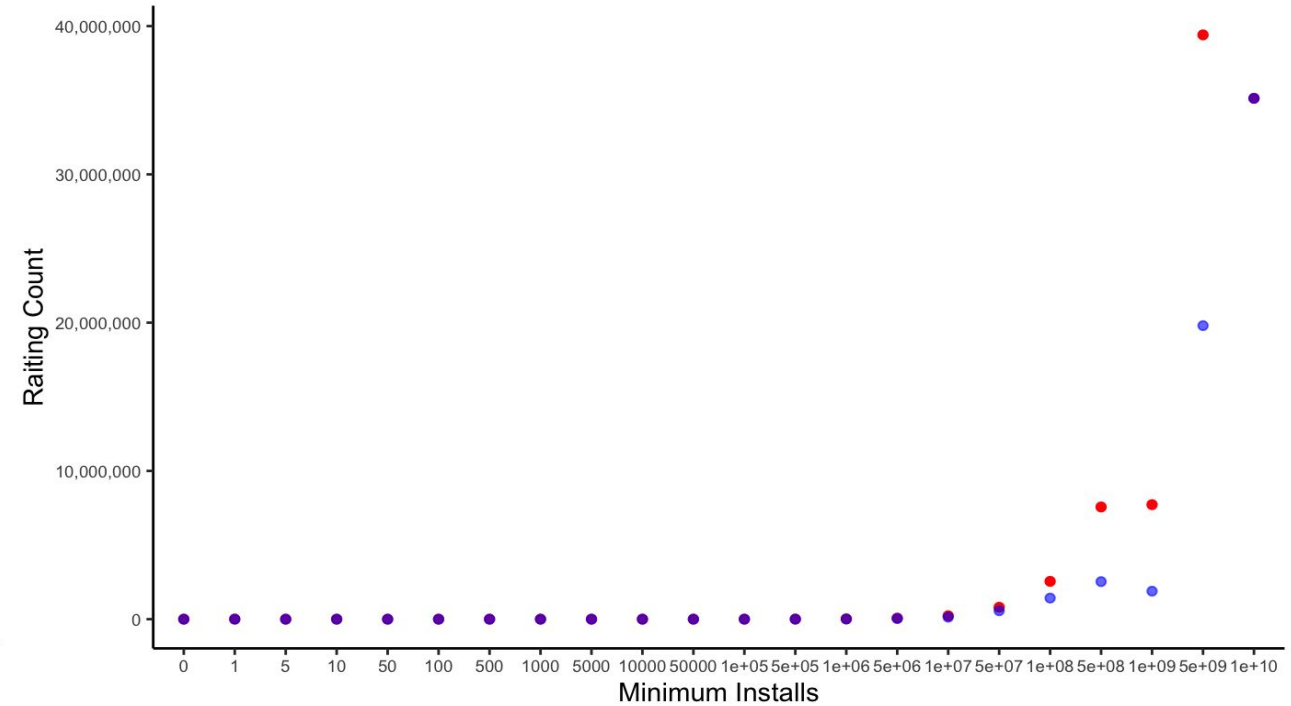
VARIABLE RATING

Correlación Rating y Minimum.Installs
Spearman: -0.3
MIC = 0.16

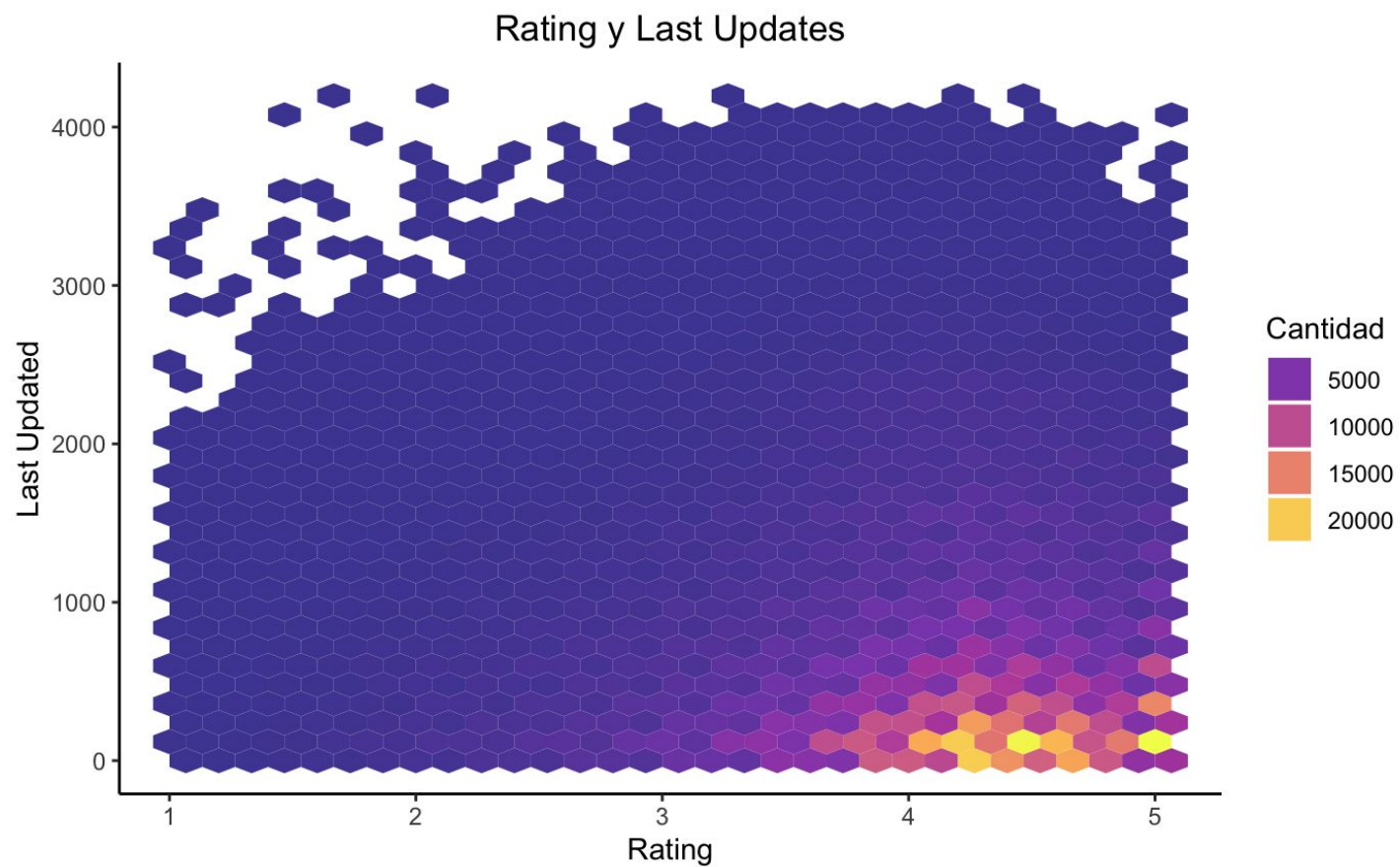
Media y Mediana de Rating por instalaciones



Media y Mediana de cantidad de rating por instalaciones



VARIABLE RATING



SEPARACIÓN TRAIN Y TEST

Tipo de predicción
Regresión

Variable target
Rating



Train 80% **Test 20%**

n = 965844

n = 239211

COMENTARIOS FINALES

Extras:

- Analizar más a fondo los nombre de las apps, explorando duplicados y uso de emojis.
- Reducir la cantidad de variables, filtrando algunas como free o minimum.installs.
- Probar clusters.
- Conseguir nuevos datos haciendo un scrapping del Google Play Store hoy.