Medición del éxito de un videojuego en Google Play

Danny Sebastián Díaz Padilla
Facultad de ingeniería en sistemas
Escuela Politécnica Nacional
Andalucía y av. Ladrón de Guevara, 170525, Quito, Ecuador
danny.diaz@epn.edu.ec

***Resumen*- Este documento presenta la definición formal del proyecto junto a las estrategias de minería de datos que se utilizarán que permitan identificar patrones en bases de datos gigantes para la posterior generación de conocimiento y sabiduría con el fin de ayudar al resto de personas que deseen iniciar un negocio por medio de un videojuego publicado en la plataforma Google Play. Se analiza también los indicadores claves de trabajos anteriores y relacionados con esta temática.**

***Palabras Clave*:** Videojuego, Minería de datos, Google Play, Web Scrapping.

I. INTRODUCCIÓN

A. Antecedentes

Debido a la adquisición de teléfonos inteligentes, desde el 2008 el mercado de aplicaciones en tiendas como App Store y Google Play ha incrementado [1]. Al momento de diseñar una aplicación existe una incertidumbre de si va a ser exitosa o no.

Tal y como lo expresa Nayebi et al. [2], la mitad de los desarrolladores tienen una clara estrategia de lanzamiento para tener una buena retroalimentación de los usuarios, sin embargo, esta estrategia se basa en sus recursos cognitivos, una lista del top 10 de aplicaciones más descargadas y su instinto. Así que este proyecto es un complemento a la estrategia de lanzamiento que tenga un desarrollador o un emprendedor, independientemente de la madurez que tenga la estrategia.

B. Descripción de la problemática

La creación de un videojuego para móviles supone un gran costo sobre todo cuando al ser publicado no genera los beneficios esperados [3].

Los emprendedores ante grandes dudas sobre la idea que será más rentable invertir, necesitan de una referencia para optimizar su toma de decisiones.

Además, el uso de las aplicaciones en general, se ve reducido a medida que transcurre el tiempo, incluso se estima que el 65% de los usuarios dejan de utilizar una app 3 meses después de instalarla [1]. De forma que es importante analizar también cuáles son los videojuegos más "adictivos" (de gran enganche) de forma que el emprendimiento sea sostenible a largo plazo.

C. Propósito del proyecto

El proyecto busca crear un modelo para predecir el éxito de un videojuego en base a las siguientes características: calificación, vistas, categoría y relación número de instalaciones-revisiones de los usuarios.

Tiene un propósito netamente comercial dirigido hacia emprendedores y programadores independientes que comprenden muy poco o nada acerca del mercado de videojuegos en teléfonos inteligentes.

Para este trabajo es importante definir el término éxito como un número de instalaciones elevado. Adicionalmente se consideran dos coeficientes: IRC el cuál es resultado de la división de reviews entre instalaciones totales y eso multiplicado por la calificación. RI es un valor resultante de dividir las reviews para el número de instalaciones.

D. Importancia del proyecto

El proyecto ayuda a la toma de decisiones. Con el fin de elegir la mejor alternativa para desarrollar una aplicación dependiendo del estado del mercado en un tiempo predeterminado o en tiempo real.

La importancia radica en la evasión de proyectos sin un futuro claro y por ende una pérdida de dinero/tiempo considerable para el inversor.

Aumentará las probabilidades de que el producto a crear sea exitoso, mas no dará una recomendación de un producto infalible, porque el diseño y la creatividad de cada persona en el desarrollo de los mismos difiere ya sea en los que producirán el producto o el nicho de mercado que vaya a consumirlo.

II. TRABAJOS PREVIOS Y RELACIONADOS

Los siguientes trabajos se enmarcan en el análisis de aplicaciones en las distintas tiendas. No existe un trabajo serio y fiable acerca del nicho específico de los videojuegos, sin embargo, estos trabajos aportan conclusiones muy interesantes y determinan en general indicadores claves que pueden reutilizarse en este trabajo.

A. **Bachelor of Science In computer Science and Engineering Exploratory Data Analysis and Success Prediction of Google Play Store Apps, Abdul Mueez et al. [4]**

En este trabajo se analiza la inmensa competencia dentro del mercado de las aplicaciones de Google Play para informar a un desarrollador para que sepa si está avanzando en la dirección correcta.
Se destaca sus conclusiones: los indicadores principales son la cantidad de instalaciones y las calificaciones de los usuarios que ha recibido a lo largo de su vida útil en lugar de los ingresos que generó. Descubrió características específicas como, por ejemplo, el número de palabras en el nombre de una aplicación afecta las instalaciones, y usarlos para averiguar qué aplicaciones tienen más probabilidades de tener éxito.
Además del análisis del número de aplicaciones que se han hecho muchas veces acorde a una categoría como lo muestra en la Fig. 1.

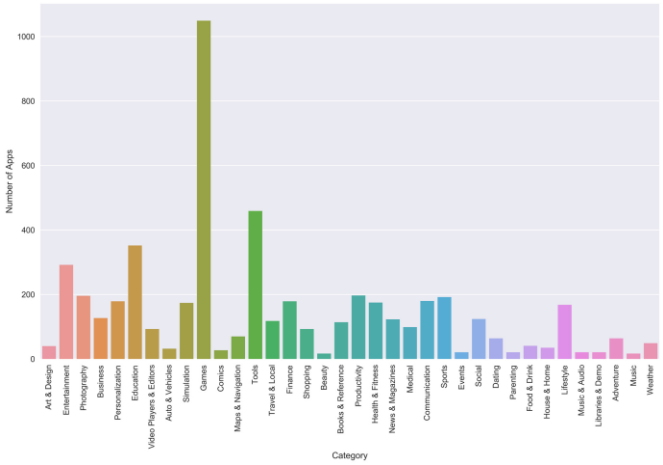


Fig. 1 Gráfica de barras de categorías contra número de aplicaciones

Y su análisis de correlaciones entre instalaciones y características de la aplicación como se muestre en un extracto en la Tabla 1.

Tabla 1

Tabla de Correlación entre número de instalaciones y otras características

Correlation of installs versus	Correlation value
Installs	1.000000
No_of_Ratings	0.141699
Size	0.060767
Rating	0.058889
Subjectivity_Mean	0.032624
Sentiment_Mean	-0.042139
Recent_Rating_Mean	-0.045481

B. **Causal Impact Analysis for App Releases in Google Play William Martin et al [3]**

En este trabajo se analiza el impacto de los lanzamientos de software propios y de competidores. Utilizando 38,858 aplicaciones populares de Google Play, durante un período de 12 meses.

El trabajo encontró que el 33% de estos lanzamientos causaron un cambio estadísticamente significativo en las calificaciones de los usuarios.
Evaluaron a 56 empresas y el 78% estuvo de acuerdo con la evaluación causal, de las cuales el 33% afirmó que su compañía consideraría cambiar su estrategia de lanzamiento de la aplicación como resultado. de estos hallazgos.

En sus conclusiones se tiene que la frecuencia general de lanzamiento no está correlacionada con el éxito posterior de la aplicación, pero que hay evidencia de que el precio y el tamaño del texto de lanzamiento y el contenido juegan un papel importante. Es más probable que los lanzamientos con precios más altos sean significativos y, quizás sorprendentemente, que tengan una calificación de efecto positivo. Todo esto resumido en la Fig. 2.

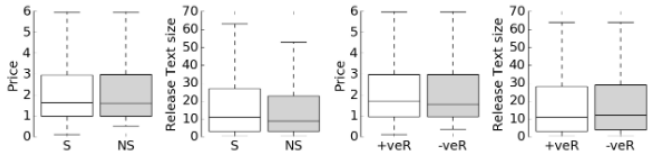


Fig. 2, Diagrama de cajas de precio y tamaño del texto de lanzamiento, comparando (S) emisiones significativas y (NS) no significativas, y emisiones que aumentaron la calificación (+ veR) y disminuyeron la calificación (-veR).

Se demuestra finalmente que el análisis causal puede ser una herramienta útil para los desarrolladores.

C. **Trabajo de fin de grado de C. E. Carazo de la Universidad de la Rioja: Citylok, una app de eventos global análisis de éxito. [5]**

En este trabajo se identifica claves para crear una app teniendo en cuenta la oportunidad en el sector de las apps y las dificultades de rentabilidad.

Se analiza solamente un caso a profundidad, Citylok que es una app que reúne la oferta cultural, gastronómica y de ocio y la presenta de una forma geolocalizada, calendarizada y categorizada en función del tipo de evento. El caso se analiza desde el punto de vista de la oferta y la demanda.

Se concluye que el punto fuerte de Citylok es su idea de negocio que se valida como relevante tanto desde el punto de vista de sus creadores como de sus usuarios ya que las 161 personas encuestadas en ese trabajo así lo han declarado. Por medio de encuestas determinan también si la idea de negocio de la aplicación es rentable y útil como se muestra en la Fig. 3.

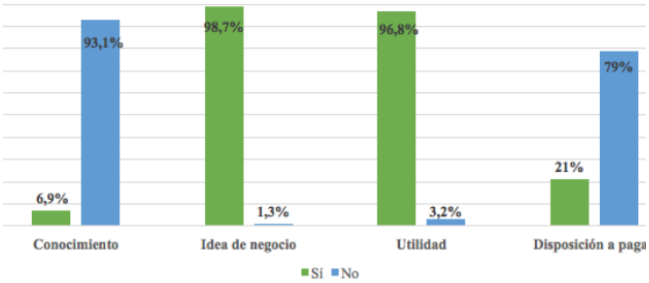


Fig. 3. Variables conocimiento, idea de negocio, utilidad y disposición a pagar

Y recomiendan para todos los emprendedores en nuevas tecnologías que utilicen las herramientas que les brinda el marketing móvil y de contenidos para crecer, consolidarse consiguiendo una comunidad del tamaño suficiente que les permita ser rentables.

D. **A Survey of App Store Analysis for Software Engineering William Martin et al [6]**

App Store Analysis estudia información sobre aplicaciones extraídas de tiendas de aplicaciones. Incorpora esta información no técnica con información técnica para conocer tendencias y comportamientos dentro de estas formas de repositorios de software. Esta encuesta describe y compara las áreas de investigación que se han explorado hasta ahora, extrayendo nuevas direcciones que la investigación futura debe tomar para abordar problemas y desafíos abiertos.

Muestran también las distribuciones de campos analizados en todos los estudios recopilados de las tiendas de aplicaciones

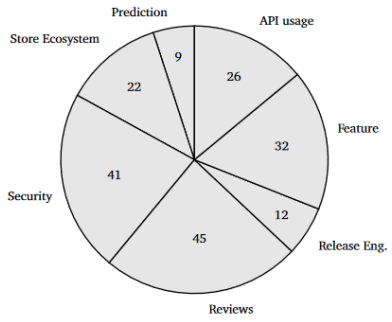


Fig. 4 Gráfico circular que muestra las distribuciones generales de los subcampos que muestran el período comprendido entre 2010 y el 27 de noviembre de 2015.

Como conclusiones rescatan que para la ingeniería de software se han identificado los subcampos clave de App Store Analysis hasta la fecha: análisis de API, análisis de características, ingeniería de versiones, análisis de revisión, análisis de seguridad, comparación de ecosistemas de tiendas y análisis de predicción de tamaño y esfuerzo.

Han observado la aparición de nuevas áreas de Análisis de App Store, y la progresión de ideas conceptuales a estudios empíricos prácticos que aplican y los refinan.

III. DESCRIPCIÓN DE LA PROPUESTA DEL PROYECTO

A. ***Objetivo general y específicos***

Objetivo general

Crear una guía para determinar las variables de éxito de un videojuego por medio de los procesos de Rapid Miner para un desarrollador de aplicaciones o emprendedor de forma que pueda determinar la tasa de éxito de su producto.

Objetivos específicos

- Crear un “dashboard” con los indicadores clave más importante del éxito de un videojuego en Google Play.
- Entrenar varios modelos matemáticos para la predicción de una variable de éxito la cual es dependiente de otras variables: puntuación de usuarios, descargas respecto revisiones de usuarios, revisiones de usuarios y la categoría.
- Seleccionar el mejor modelo matemático e incluir 50 ejemplos de prueba en el simulador.

B. ***Detalle***

Para el proyecto se utilizará una cantidad de 267mil aplicaciones, de las cuales se filtrará únicamente los que tengan categorías relacionados a los videojuegos.

Para la selección de esa cantidad de aplicaciones se ha indagado en más de 15 bases de datos y se han filtrado aquellas bases de datos que serán útiles para el análisis.

Además, como un extra se pretende recolectar con Web Scraping información de los videojuegos más exitosos dado que no existe una base de datos confiable que reúna características de este entorno en específico

Posteriormente se realizará una limpieza de datos utilizando una herramienta open source llamada Open Refine y usando un lenguaje de programación (Python) se obtendrá un conjunto de datos de calidad con desviación estándar aceptable e información confiable.

Con la base de datos limpia se la incluirá en la herramienta Rapid Miner y se procederá a realizar filtros especiales y gráficas para encontrar patrones.

Se utilizará la herramienta de auto modelamiento para seleccionar el mejor algoritmo de predicción en base a las características que se determinen más importantes del análisis anterior. Y utilizando el simulador se incluirá 50 ejemplos de prueba para ver qué factores tienen más “peso” al momento de crear un videojuego.

Este mismo proceso se realizará para la recolección de los datos utilizando Web Scraping. Se determinará por dos vías

diferentes los mejores factores, los mejores algoritmos y finalmente se recomendará un claro conjunto de campos para que el desarrollador y el emprendedor realicen una elección correcta sobre su próxima idea de negocios.

IV. HERRAMIENTAS QUE SE PRETENDE USAR

A. Open Refine (De Google)

Es una herramienta originalmente creada por Google para el manejo de bases de datos. Permite limpiar bases de datos, exportarlas en diferentes formatos, y arreglar y manejar las bases para un mejor uso. Actualmente el proyecto ya no es financiado por Google y se encuentra como proyecto abierto.[7]

B. Rapid Miner

RapidMiner es una herramienta perfecta para crear modelos y a posterior la realización de análisis predictivos de grandes volúmenes de datos.

Es una solución que facilita el autoservicio de análisis predictivo permitiendo una avanzada analítica empleando solamente drag and drop y opcionalmente la generación de código.

Se utiliza para realizar análisis de minería de datos (Data Mining) en aplicaciones empresariales, gobierno y academias. [8]

C. Python

Aunque RapidMiner ofrece una forma de limpiar los datos, no es tan completa como la de Open Refine (excepto en la formación de terceras columnas con nueva data de forma sencilla).

Python es el lenguaje estándar usado en Open Refine para aplicar funciones en toda una columna de datos de forma que se pueda formatear las columnas y asignarles un tipo de valor en el proceso de limpieza.

Su segundo uso será para realizar web scrapping de forma masiva, esto es la recolección de datos del HTML público mostrado por cada página web de interés dentro de las categorías de juegos en la tienda de aplicaciones Google Play.

La dependencia necesaria para realizar esto es Beautiful Soup, misma que sirve para extraer información de ficheros HTML y XML [9]

D. Tableau public

Es un software gratuito de visualización de datos que a partir de simples archivos de Excel u otros orígenes de datos, permite generar visualizaciones de alto impacto gráfico e interactivo.[10]

V. EXPERIMENTACIÓN

Para la experimentación se utiliza el siguiente proceso de Bi Data: Adquisición/Transporte de datos, almacenamiento, limpieza de datos, indexación, analítica de datos (incluyendo un proceso para la limpieza de datos), visualización de los datos. Con el fin de producir una toma de decisión por parte del lector de esos datos.

A. **Adquisición de datos**

Para la adquisición de datos se utiliza una técnica para la recolección de datos a partir del HTML presentado en las direcciones URL relacionadas a las subcategorías de videojuegos.

El primer paso consiste en trazar un mapa que debe seguir el algoritmo hasta llegar a los detalles de interés de cada aplicación en particular.

Las URLs iniciales son provistas manualmente y los siguientes enlaces internos se obtienen de forma automática realizando una búsqueda dentro de las etiquetas HTML del tipo ‘<a href=”” />’ que establece un enlace con la siguiente página.

URLs iniciales:

- "https://play.google.com/store/apps/category/GAME?hl=en"
- "https://play.google.com/store/apps/top/category/GAME?hl=en"
- "https://play.google.com/store/apps/new/category/GAME?hl=en"
- "https://play.google.com/store/apps/category/GAME_ACTION"
- "https://play.google.com/store/apps/category/GAME_ADVENTURE"
- "https://play.google.com/store/apps/category/GAME_ARCADE"
- "https://play.google.com/store/apps/category/GAME_BOARD"
- "https://play.google.com/store/apps/category/GAME_CARD"
- "https://play.google.com/store/apps/category/GAME_CASINO"
- "https://play.google.com/store/apps/category/GAME_CASUAL"
- "https://play.google.com/store/apps/category/GAME_EDUCATIONAL"
- "https://play.google.com/store/apps/category/GAME_MUSIC"
- "https://play.google.com/store/apps/category/GAME_PUZZLE"
- "https://play.google.com/store/apps/category/GAME_RACING"

- "https://play.google.com/store/apps/category/GAME_ROLE_PLAYING"
- "https://play.google.com/store/apps/category/GAME_SIMULATION"
- "https://play.google.com/store/apps/category/GAME_SPORTS"
- "https://play.google.com/store/apps/category/GAME_STRATEGY"
- "https://play.google.com/store/apps/category/GAME_TRIVIA"
- "https://play.google.com/store/apps/category/GAME_WORD"..

Un ejemplo de recolección de enlaces se muestra en la Fig. 5, y aprovecha la reutilización de código acostumbrada en la programación WEB. En este ejemplo se busca una clase con nombre "xwY9Zc", que a su vez es reutilizada por otros contenedores para proveer una misma función: mostrar una lista horizontal de aplicaciones.

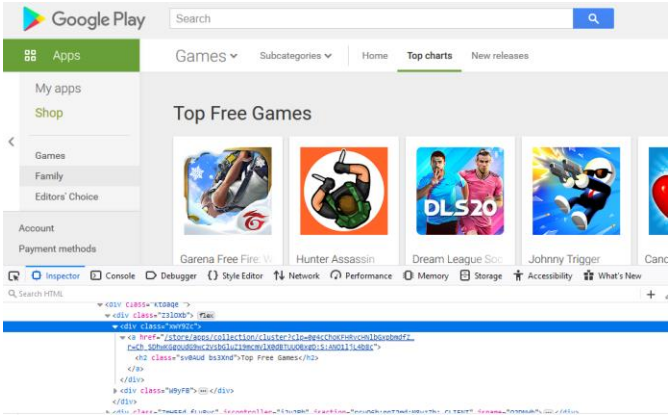


Fig. 5 Inicio del mapeo para web scrapping

El objetivo de la primera recolección de enlaces es desplegar la lista de aplicaciones de forma que se maximice la cantidad de objetos a analizar.

En la siguiente pantalla se busca una clase encargada de presentar los contenedores pequeños horizontales para cada aplicación (Fig. 6).

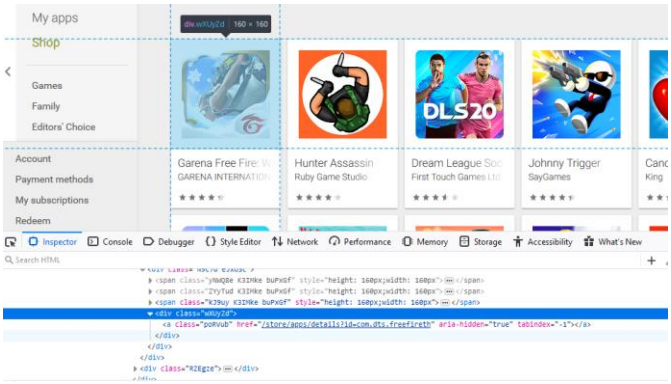


Fig. 6 Ingreso al catálogo usando la referencia interna

Es importante recalcar que el nombre de las clases usadas en las etiquetas por parte de Google cambia en el tiempo, dependiendo de su política de decoración y frameworks que establecen nombres de clases de forma automática.

En este experimento, además, se comprobó que 11 aplicaciones de 3500 no contienen el mismo nombre en las clases contenedoras como se muestra en la Fig. 7. Esto se sabe ya que, si al buscar la clase no existe, el programa ejecuta un error controlado y se descarta la aplicación.

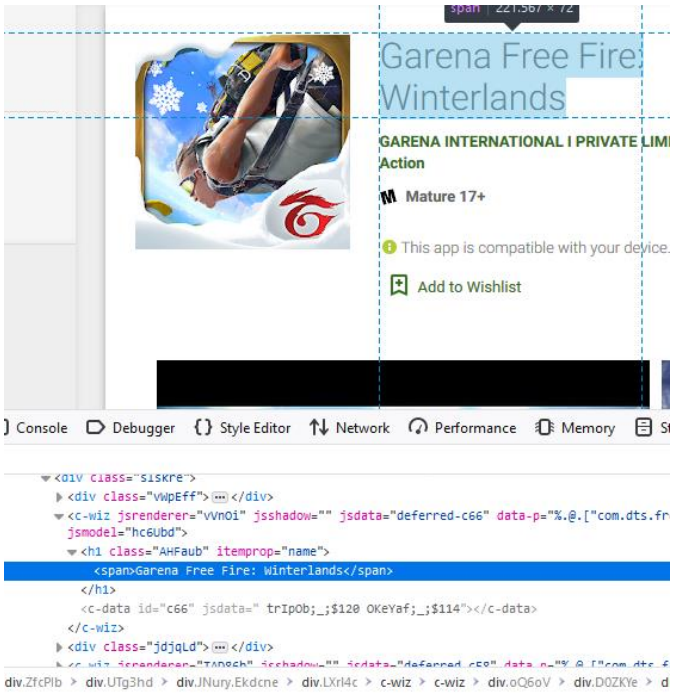


Fig. 7 Pantalla final con el detalle de la aplicación

Los enlaces de las siguientes tarjetas llevan directamente a los detalles de la aplicación de forma que el diámetro del mapa es de 3, entiéndase diámetro como longitud mínima de pasos que debe realizar un nodo A para llegar a un nodo B a través de un grafo de elementos.

El mapa de los enlaces entre cada página web funciona como un grafo en este experimento.

Cada tarea de recolección de enlaces para cada elemento, es repetitiva debido a la naturaleza de la programación explicada al inicio de este punto (A).

Se puede crear un mapa desde cualquier punto del dominio web a analizar, sin embargo, debido a que existían pocas categorías de videojuegos se optó por seleccionar manualmente las URLs de comienzo.

Un problema muy importante que ocurre al realizar la navegación entre enlaces es que el catalogo de aplicaciones se despliega de forma limitada, en espera de que el usuario realice un "scroll down" hasta el final de la página para mostrarle más aplicaciones.

Esto se debe a que es mucho más óptimo para la base de datos interna de Google mostrar poco a poco las aplicaciones (de forma incremental) que mostrar el 100% de aplicaciones desde la primera navegación.

La solución de este problema es: simular el “scroll down” interactuando directamente con el HTML (específicamente el botón o atributo de la página encargado de esta carga incremental).

El transporte de los datos se realizó por medio de peticiones HTTP. Y la velocidad del mismo rondaba los 100 milisegundos. Mientras que el procesamiento de los datos en promedio rondaba los 5 segundos por cada aplicación con las características mostradas en la Fig. 8. (Considerar 6GB de RAM asignadas al proceso y un 100% de los recursos del procesador solo fueron usados alrededor de 42%).

Mientras más RAM se asigne mucho más rápido terminará la recolección de datos.

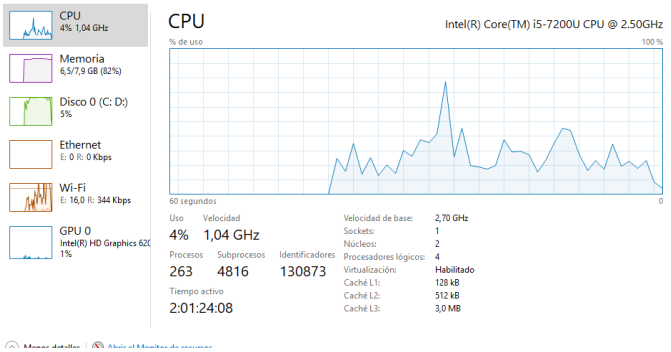


Fig. 8 Uso del programa

B. **Almacenamiento de datos**

No se utilizó un motor de almacenamiento como Neo4J, Hadoop, MySQL o asociados.

Debido a la cantidad de datos y la versatilidad de los documentos .xlsx dentro de las herramientas Rapid Miner y Tableau. Se optó por almacenar la información en un archivo Excel (Fig. 9).

tamanio_	LongitudD	PalabrasP	nombre	Coeficien	numero_r	instalacio	Coeficien	calificacio	precio	calificacio	categoria	purchase_	purchase_
49,2	2416,0	16,0	Candy Cru	,1	########	########	,3	4,6	,0	Everyone	Casual	1,0	150,0
49,2	3879,0	14,0	Bubble Sh	,0	728124,0	########	,0	4,2	,0	Everyone	Casual	1,0	100,0
92,0	1550,0	6,0	Roblox	,1	########	########	,5	4,5	,0	Everyone	Adventure	,5	200,0
99,0	3377,0	25,0	Mobile Le	,2	########	########	,7	4,4	,0	Teen	Action	1,0	100,0
73,0	2176,0	17,0	Hill Climb	,0	########	########	,1	4,4	,0	Everyone	Racing	2,5	123,0
48,0	114,0	21,0	Bubble Sh	,0	135737,0	########	,1	4,6	,0	Everyone	Casual	1,0	100,0
40,0	1456,0	29,0	Garena Fr	,4	########	########	1,7	4,3	,0	Mature 17	Action	1,0	110,0
95,0	434,0	14,0	Subway Su	,0	########	########	,1	4,5	,0	Everyone	Arcade	1,0	105,0
49,2	1339,0	9,0	Minecraft	,3	########	########	1,3	4,4	7,0	Everyone	Arcade	1,0	50,0
82,0	2522,0	17,0	Stick War:	,0	########	########	,1	4,4	,0	Teen	Strategy	2,5	140,0
24,0	588,0	3,0	Pou	,0	########	########	,1	4,4	,0	Everyone	Casual	1,0	25,0
27,0	1114,0	22,0	Carrom Po	,0	########	########	,1	4,4	,0	Everyone	Sports	1,0	100,0
99,0	2184,0	10,0	Gacha Life	,2	########	########	,6	4,1	,0	Everyone	Casual	2,0	20,0
45,0	580,0	19,0	World Soc	,0	########	########	,1	4,2	,0	Everyone	Sports	2,0	6,0
32,0	826,0	25,0	Anger of s	,0	911168,0	########	,0	4,3	,0	Teen	Action	1,0	105,0
49,2	3477,0	16,0	Bubble Sh	,0	80401,0	########	,0	4,2	,0	Everyone	Arcade	1,0	105,0
38,0	829,0	32,0	Worms Zo	,0	260485,0	########	,1	4,4	,0	Everyone	Action	2,0	100,0
49,2	1896,0	14,0	Lep's Wor	,0	860321,0	########	,0	4,3	,0	Teen	Arcade	1,0	105,0
31,0	439,0	24,0	Supreme I	,0	215852,0	########	,1	4,3	,0	Teen	Action	,0	,0
50,0	965,0	10,0	Standoff 2	,2	########	########	,7	4,3	,0	Mature 17	Action	1,0	70,0
49,2	440,0	10,0	slither.io	,1	########	########	,2	4,2	,0	Everyone	Action	4,0	4,0
92,0	2228,0	21,0	Mini Worl	,1	########	########	,5	4,3	,0	Everyone	Adventure	1,0	100,0
42,0	922,0	11,0	Robbery B	,0	931652,0	########	,1	4,3	,0	Everyone	Action	1,0	8,5
77,0	3315,0	18,0	Zombie Ca	,0	########	########	,1	4,4	,0	Everyone	Action	3,0	50,0
31,0	1657,0	23,0	2 3 4 Playe	,0	299227,0	########	,0	4,3	,0	Everyone	Simulatio	,0	,0
130,0	2910,0	11,0	Brawl Star	,1	########	########	,4	4,3	,0	Everyone	Action	1,0	100,0

Fig. 9 Excel generado por el programa

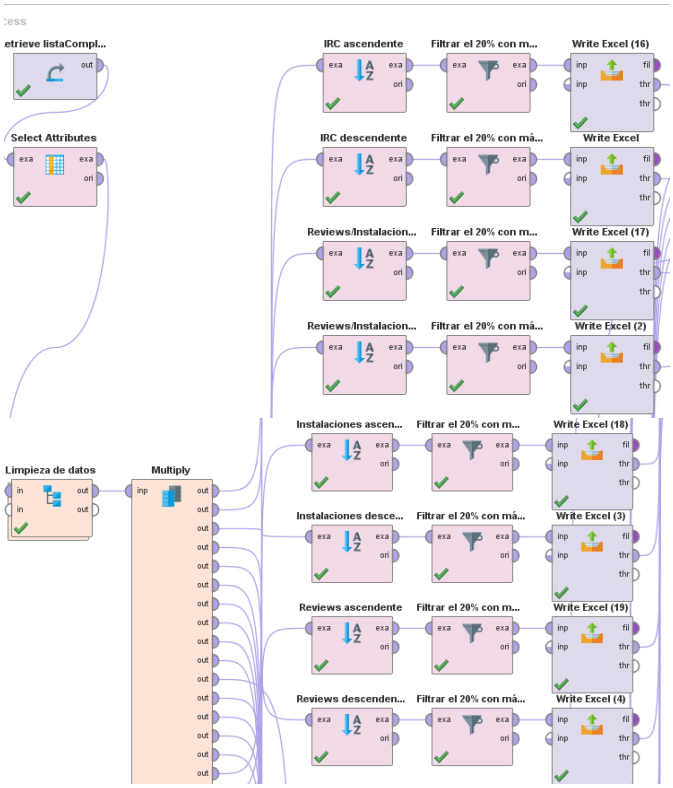
C. **Indexación de datos**

La indexación de datos es recomendable realizarla en grandes volúmenes de datos para facilitar la consulta posterior de elementos. Las aplicaciones de videojuegos recolectadas rondan las 3500; este **volumen** no se consideró lo suficientemente grande para aplicar una estrategia de indexación en un almacén de datos distinto a un archivo Excel.

D. **Analítica de datos**

Al realizar Web Scraping se realizó una limpieza al mismo tiempo que recibía los datos de la aplicación pero como extra se usó OpenRefine para corroborar posibles errores que pudieron haber pasado por alto.

En Rapid Miner se prepara primero los datos eliminando duplicados y seleccionando atributos específicos. Además, una extracción del 20% mayor y menor respecto a los temas: Coeficiente IRC, coeficiente RI, instalaciones, reviews, calificaciones, longitud del título y longitud de la descripción. Como agregaciones solo se realizó promedio por categoría para: coeficiente IRC, coeficiente RI, instalaciones, reviews y calificaciones. Como se ve en la Fig. 10 particionada.



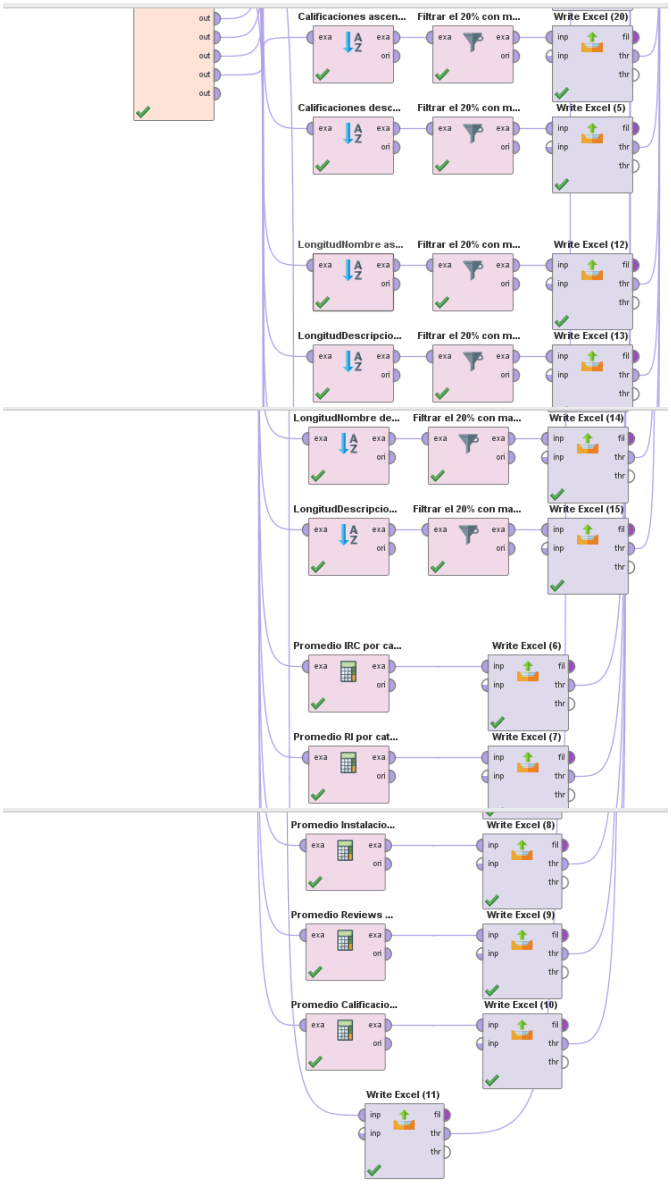


Fig. 10 Proceso creado para el procesamiento de datos

La salida de este proceso es un conjunto de 20 archivos con los datos procesados y filtrados por propósito previamente mencionado como se ve en la Fig. 11:

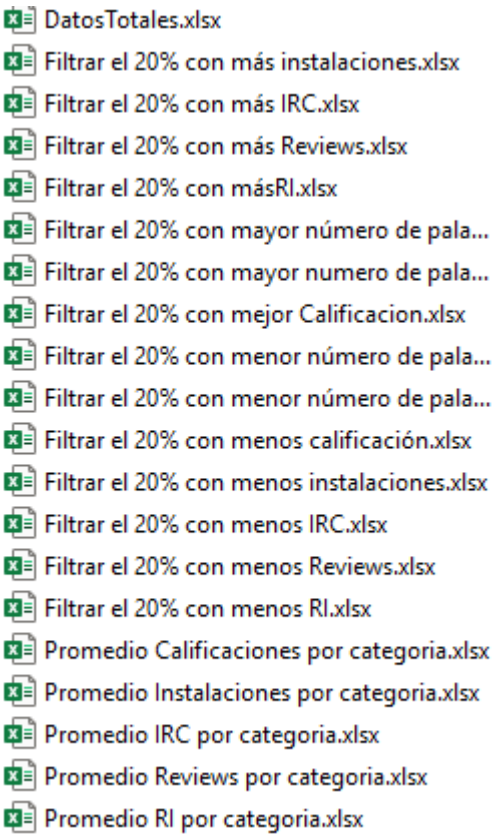


Fig. 11 Salida del proceso de RapidMiner

E. **Visualizaciones**

Para la visualización se consideró el conjunto de datos totales, de forma que se creó los siguientes tres dashboard:

Un análisis de la influencia de las palabras para el número de instalaciones como se ve en la Fig. 12.

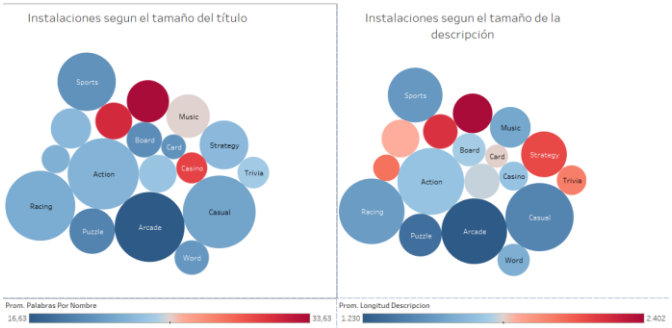


Fig. 12 Tamaño de las palabras y descripciones por categoría

Un análisis separado para la calificación del contenido respecto al número de instalaciones (Fig. 13)

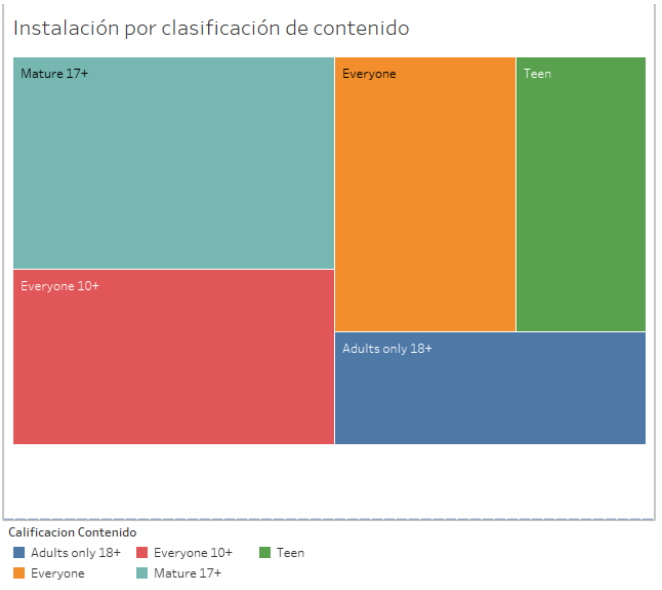


Fig. 13 Instalaciones por cada clasificación de contenidos

Y finalmente un dashboard completo con indicadores que pueden ser importantes para el lector.

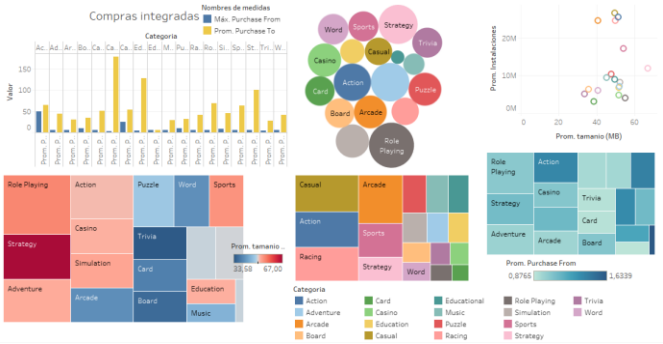


Fig. 14 Dashboard de variables de interés al generar la aplicación

En la Fig. 14 se observa 6 gráficas, se las explicará de izquierda a derecha, fila a fila:

- El primer gráfico marca cuales son los juegos con mayores ganancias por las compras integradas siendo el casino la categoría que más caro pone a los objetos dentro de la aplicación.
- El segundo gráfico indica que tan grande es el coeficiente IRC (en promedio) por cada categoría.
- El tercer gráfico analiza si el tamaño promedio de la aplicación es un factor importante, se observa que la mayoría de categorías están en un rango cercano en cuando a MB.
- El cuarto gráfico es un mapa de calor que marca a lo más azul como las aplicaciones más livianas en términos de MB y el tamaño implica el número de instalaciones.
- El quinto gráfico indica el número de instalaciones por cada categoría.
- Por último, se marca el número de instalaciones con el color marcando el valor promedio

VI. **RESULTADOS**

Los que más palabras tienen en el título:

numero_reviews	Real	0	Min 70	Max 11889010	Average 371845.593
instalaciones	Real	0	Min 5000	Max 100000000	Average 9151200
CoeficienteCalidadIRC	Real	0	Min 0.003	Max 2.298	Average 0.222
calificacion_aplicacion	Real	0	Min 1.700	Max 4.900	Average 4.327

Los que menos palabras tienen en el título:

numero_reviews	Real	0	Min 10	Max 24554300	Average 690462.368
instalaciones	Real	0	Min 5000	Max 500000000	Average 16779175
CoeficienteCalidadIRC	Real	0	Min 0.003	Max 1.800	Average 0.271
calificacion_aplicacion	Real	0	Min 1.800	Max 4.900	Average 4.282

Los que tienen más palabras en la descripción:

CoeficienteRI	Real	0	Min 0.001	Max 0.505	Average 0.069
numero_reviews	Real	0	Min 167	Max 24554300	Average 616827.855
instalaciones	Real	0	Min 5000	Max 500000000	Average 12643150
CoeficienteCalidadIRC	Real	0	Min 0.004	Max 2.186	Average 0.301
calificacion_aplicacion	Real	0	Min 2.900	Max 4.900	Average 4.328
precio	Real	0	Min 0	Max 18.990	Average 0.726

Los que tienen menos palabras en la descripción:

CoeficienteRI	Real	0	Min 0.001	Max 0.262	Average 0.039
numero_reviews	Real	0	Min 10	Max 31883070	Average 492376.135
instalaciones	Real	0	Min 5000	Max 1000000000	Average 20058550
CoeficienteCalidadIRC	Real	0	Min 0.003	Max 1.205	Average 0.170
calificacion_aplicacion	Real	0	Min 1.800	Max 4.800	Average 4.241
precio	Real	0	Min 0	Max 13.990	Average 0.178

Los que tienen mayor cantidad de instalaciones:

tamanio_(MB)	Real	0	Min 2.200	Max 155	Average 58.624
LongitudDescripcion	Integer	0	Min 4	Max 4037	Average 1664.185
PalabrasPorNombre	Real	0	Min 2	Max 50	Average 21.310
numero_reviews	Real	0	Min 20000	Max 51308241	Average 2348244.685
instalaciones	Real	0	Min 10000000	Max 1000000000	Average 60760484.673
CoeficienteCalidadIRC	Real	0	Min 0.006	Max 1.734	Average 0.173
calificacion_aplicacion	Real	0	Min 3.500	Max 4.800	Average 4.324
precio	Real	0	Min 0	Max 6.990	Average 0.021
calificacion_contenido	Nominal	0	Least Adults only 18+ (0)	Most Everyone (233)	Values Everyone (233), Teen (103), ...[3 more]
categoria	Nominal	0	Least Travel & Local (0)	Most Action (88)	Values Action (88), Arcade (60), ...[23 more]
purchase_from	Real	0	Min 0	Max 24.990	Average 1.231
purchase_to	Real	0	Min 0	Max 399.990	Average 79.705

Los que tienen menor cantidad de instalaciones:

tamanio_(MB)	Real	0	Min 0.723	Max 214	Average 40.317
LongitudDescripcion	Integer	0	Min 94	Max 4054	Average 1650.680
PalabrasPorNombre	Real	0	Min 4	Max 50	Average 19.600
nombre	Nominal	0	Least Skyscan [...] ental (0)	Most 0h h1 (1)	Values 0h h1 (1), 0h n0 (1), ...[2332 more]
CoeficienteRI	Real	0	Min 0.001	Max 0.543	Average 0.112
numero_reviews	Real	0	Min 10	Max 54300	Average 6160.882
instalaciones	Real	0	Min 5000	Max 100000	Average 61675
CoeficienteCalidadIRC	Real	0	Min 0.003	Max 2.573	Average 0.481
calificacion_aplicacion	Real	0	Min 1.700	Max 4.900	Average 4.219
precio	Real	0	Min 0	Max 19.990	Average 2.309
calificacion_contenido	Nominal	0	Least Adults only 18+ (0)	Most Everyone (211)	Values Everyone (211), Teen (120), ...[3 more]
categoria	Nominal	0	Least Travel & Local (0)	Most Role Playing (64)	Values Role Playing (64), Puzzle (49), ...[23 more]
purchase_from	Real	0	Min 0	Max 49.990	Average 0.816
purchase_to	Real	0	Min 0	Max 349.990	Average 21.919

Las que tienen más RI:

tamanio_(MB)	Real	0	Min 2.600	Max 223	Average 49.611
LongitudDescripcion	Integer	0	Min 4	Max 6410	Average 1881.757
PalabrasPorNombre	Real	0	Min 4	Max 50	Average 20.707
nombre	Nominal	0	Least Skyscan [...] ental (0)	Most 10000000 (1)	Values 10000000 (1), 2048 Number puzzle game
CoeficienteRI	Real	0	Min 0.096	Max 0.705	Average 0.179
numero_reviews	Real	0	Min 502	Max 51308241	Average 966394.080
instalaciones	Real	0	Min 5000	Max 500000000	Average 6418675
CoeficienteCalidadIRC	Real	0	Min 0.307	Max 3.241	Average 0.786
calificacion_aplicacion	Real	0	Min 3	Max 4.900	Average 4.389
precio	Real	0	Min 0	Max 19.990	Average 1.895
calificacion_contenido	Nominal	0	Least Adults only 18+ (0)	Most Everyone (162)	Values Everyone (162), Teen (131), ...[3 more]
categoria	Nominal	0	Least Travel & Local (0)	Most Role Playing (83)	Values Role Playing (83), Adventure (42), ...[23 r
purchase_from	Real	0	Min 0	Max 4.990	Average 0.876
purchase_to	Real	0	Min 0	Max 399.990	Average 52.606

Los que tienen menos RI:

tamanio_(MB)	Real	0	Min 0.687	Max 161	Average 42.372
LongitudDescripcion	Integer	0	Min 2	Max 8378	Average 1488.995
PalabrasPorNombre	Real	0	Min 4	Max 50	Average 23.040
nombre	Nominal	0	Least Skyscan [...] ental (0)	Most 100 Doors 2015 Pro (1)	Values 100 Doors 2015 Pro (1), 100 Door [...] om
CoeficienteRI	Real	0	Min 0.001	Max 0.012	Average 0.007
numero_reviews	Real	0	Min 10	Max 1192356	Average 120037.673
instalaciones	Real	0	Min 5000	Max 100000000	Average 15450750
CoeficienteCalidadIRC	Real	0	Min 0.003	Max 0.060	Average 0.029
calificacion_aplicacion	Real	0	Min 1.700	Max 4.900	Average 4.110
precio	Real	0	Min 0	Max 0	Average 0
calificacion_contenido	Nominal	0	Least Adults only 18+ (0)	Most Everyone (309)	Values Everyone (309), Teen (55), ...[3 more]
categoria	Nominal	0	Least Travel & Local (0)	Most Puzzle (47)	Values Puzzle (47), Educational (45), ...[23 more]
purchase_from	Real	0	Min 0	Max 49.990	Average 1.343
purchase_to	Real	0	Min 0	Max 399.990	Average 25.288

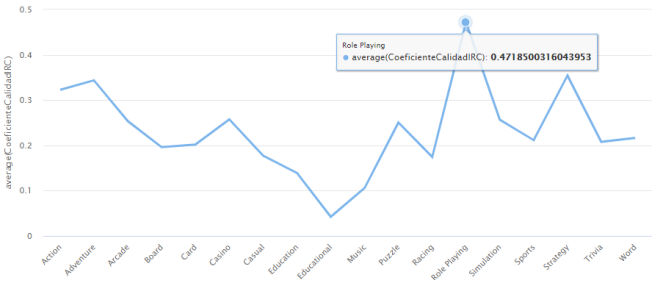
Los que tienen más IRC:

tamanio_(MB)	Real	0	Min 2.600	Max 223	Average 50.015
LongitudDescripcion	Integer	0	Min 4	Max 6410	Average 1864.720
PalabrasPorNombre	Real	0	Min 4	Max 50	Average 20.750
nombre	Nominal	0	Least Skyscan [...] ental (0)	Most 10000000 (1)	Values 10000000 (1), 2048 Number puzzle game
CoeficienteRI	Real	0	Min 0.087	Max 0.705	Average 0.178
numero_reviews	Real	0	Min 502	Max 51308241	Average 966760.073
instalaciones	Real	0	Min 5000	Max 500000000	Average 6427987.500
CoeficienteCalidadIRC	Real	0	Min 0.417	Max 3.241	Average 0.787
calificacion_aplicacion	Real	0	Min 3.200	Max 4.900	Average 4.411
precio	Real	0	Min 0	Max 19.990	Average 1.880
calificacion_contenido	Nominal	0	Least Adults only 18+ (0)	Most Everyone (166)	Values Everyone (166), Teen (129), ...[3 more]
categoria	Nominal	0	Least Travel & Local (0)	Most Role Playing (82)	Values Role Playing (82), Puzzle (42), ...[23 more
purchase_from	Real	0	Min 0	Max 4.990	Average 0.876
purchase_to	Real	0	Min 0	Max 399.990	Average 52.442

Los que tienen menos IRC:

tamanio_(MB)	Real	0	Min 0.687	Max 161	Average 42.782
LongitudDescripcion	Integer	0	Min 2	Max 8378	Average 1480.818
PalabrasPorNombre	Real	0	Min 4	Max 50	Average 23.030
nombre	Nominal	0	Least Skyscan [...] ental (0)	Most 100 Doors 2015 Pro (1)	Values 100 Doors 2015 Pro (1), 100 Door [...] om
CoeficienteRI	Real	0	Min 0.001	Max 0.016	Average 0.007
numero_reviews	Real	0	Min 10	Max 1170092	Average 113971.235
instalaciones	Real	0	Min 5000	Max 100000000	Average 14934125
CoeficienteCalidadIRC	Real	0	Min 0.003	Max 0.051	Average 0.029
calificacion_aplicacion	Real	0	Min 1.700	Max 4.800	Average 4.080
precio	Real	0	Min 0	Max 1.990	Average 0.005
calificacion_contenido	Nominal	0	Least Adults only 18+ (0)	Most Everyone (307)	Values Everyone (307), Teen (58), ...[3 more]
categoria	Nominal	0	Least Travel & Local (0)	Most Educational (45)	Values Educational (45), Puzzle (45), ...[23 more]
purchase_from	Real	0	Min 0	Max 49.990	Average 1.382
purchase_to	Real	0	Min 0	Max 399.990	Average 25.313

Promedio de IRC por categoría:



Los que más calificación tienen:

tamanio_(MB)	Real	0	Min 0.723	Max 147	Average 52.679
LongitudDescripcion	Integer	0	Min 4	Max 8378	Average 1767.910
PalabrasPorNombre	Real	0	Min 3	Max 50	Average 22.140
nombre	Nominal	0	Least Skyscan [...] ental (0)	Most 0h h1 (1)	Values 0h h1 (1), 0h n0 (1), ...[2332 more]
CoeficienteRI	Real	0	Min 0.003	Max 0.705	Average 0.094
numero_reviews	Real	0	Min 68	Max 51308241	Average 940049.655
instalaciones	Real	0	Min 5000	Max 1000000000	Average 16154287.500

CoeficienteCalidadIRC	Real	0	Min 0.013	Max 3.241	Average 0.438
calificacion_aplicacion	Real	0	Min 4.500	Max 4.900	Average 4.652
precio	Real	0	Min 0	Max 17.990	Average 0.740
calificacion_contenido	Nominal	0	Least Adults only 18+ (1)	Most Everyone (239)	Values Everyone (239), Teen (104), ...[3 more]
categoria	Nominal	0	Least Travel & Local (0)	Most Puzzle (65)	Values Puzzle (65), Word (37), ...[23 more]
purchase_from	Real	0	Min 0	Max 9.990	Average 1.027
purchase_to	Real	0	Min 0	Max 399.990	Average 64.020

Los que menos calificación tienen:

tamanio_(MB)	Real	0	Min 2	Max 278	Average 45.802
LongitudDescripcion	Integer	0	Min 4	Max 4057	Average 1513.065
PalabrasPorNombre	Real	0	Min 4	Max 50	Average 20.665
nombre	Nominal	0	Least Skyscan [...] ental (0)	Most 100 Crypts (1)	Values 100 Crypts (1), 100 Doors (1), ...[2332 mo
CoeficienteRI	Real	0	Min 0.001	Max 0.469	Average 0.040
numero_reviews	Real	0	Min 10	Max 6177428	Average 156457.270
instalaciones	Real	0	Min 5000	Max 100000000	Average 8047925
CoeficienteCalidadIRC	Real	0	Min 0.003	Max 1.735	Average 0.151
calificacion_aplicacion	Real	0	Min 1.700	Max 4.100	Average 3.810
precio	Real	0	Min 0	Max 14.990	Average 0.651
calificacion_contenido	Nominal	0	Least Adults only 18+ (0)	Most Everyone (263)	Values Everyone (263), Teen (93), ...[3 more]
categoria	Nominal	0	Least Travel & Local (0)	Most Simulation (50)	Values Simulation (50), Arcade (38), ...[23 more]
purchase_from	Real	0	Min 0	Max 9.990	Average 1.205
purchase_to	Real	0	Min 0	Max 399.990	Average 38.335

Los que tienen más reviews:

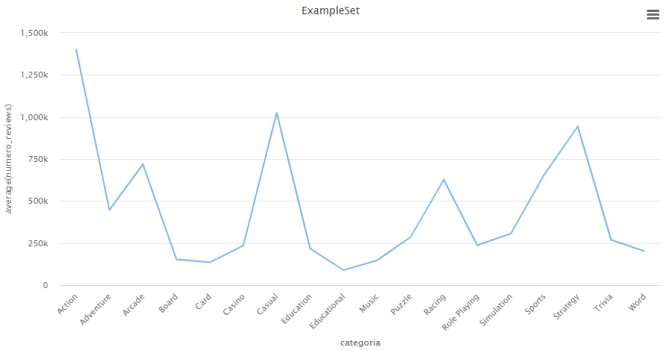
tamanio_(MB)	Real	0	Min 2.200	Max 155	Average 63.054
LongitudDescripcion	Integer	0	Min 4	Max 4037	Average 1818.320
PalabrasPorNombre	Real	0	Min 2	Max 50	Average 20.785
nombre	Nominal	0	Least Skyscan [...] ental (0)	Most 2048 Num [...] game (1)	Values 2048 Number puzzle game (1), 4 Pics 1
CoeficienteRI	Real	0	Min 0.007	Max 0.705	Average 0.069
numero_reviews	Real	0	Min 537987	Max 51308241	Average 2591573.295
instalaciones	Real	0	Min 1000000	Max 1000000000	Average 56025484.673
CoeficienteCalidadIRC	Real	0	Min 0.031	Max 3.241	Average 0.306
calificacion_aplicacion	Real	0	Min 3.600	Max 4.800	Average 4.386
precio	Real	0	Min 0	Max 6.990	Average 0.026
calificacion_contenido	Nominal	0	Least Adults only 18+ (1)	Most Everyone (211)	Values Everyone (211), Teen (109), ...[3 more]
categoria	Nominal	0	Least Travel & Local (0)	Most Action (69)	Values Action (69), Arcade (44), ...[23 more]
purchase_from	Real	0	Min 0	Max 4.990	Average 1.128
purchase_to	Real	0	Min 0	Max 399.990	Average 94.135

Los que tienen menos reviews:

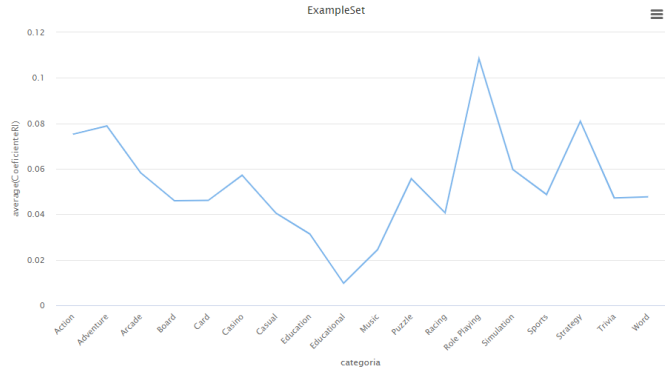
tamanio_(MB)	Real	0	Min 0.687	Max 128	Average 39.941
LongitudDescripcion	Integer	0	Min 47	Max 8378	Average 1644.767
PalabrasPorNombre	Real	0	Min 4	Max 50	Average 21.015
nombre	Nominal	0	Least Skyscan [...] ental (0)	Most 0h h1 (1)	Values 0h h1 (1), 0h n0 (1), ...[2332 more]
CoeficienteRI	Real	0	Min 0.001	Max 0.534	Average 0.069
numero_reviews	Real	0	Min 10	Max 7443	Average 2944.468
instalaciones	Real	0	Min 5000	Max 5000000	Average 222425

CoeficienteCalidadIRC	Real	0	Min 0.003	Max 2.298	Average 0.297
calificacion_aplicacion	Real	0	Min 1.700	Max 4.900	Average 4.183
precio	Real	0	Min 0	Max 18.990	Average 1.504
calificacion_contenido	Nominal	0	Least Adults only 18+ (0)	Most Everyone (258)	Values Everyone (258), Teen (93), ...[3 more]
categoria	Nominal	0	Least Travel & Local (0)	Most Puzzle (49)	Values Puzzle (49), Simulation (45), ...[23 more]
purchase_from	Real	0	Min 0	Max 49.990	Average 0.902
purchase_to	Real	0	Min 0	Max 399.990	Average 20.251

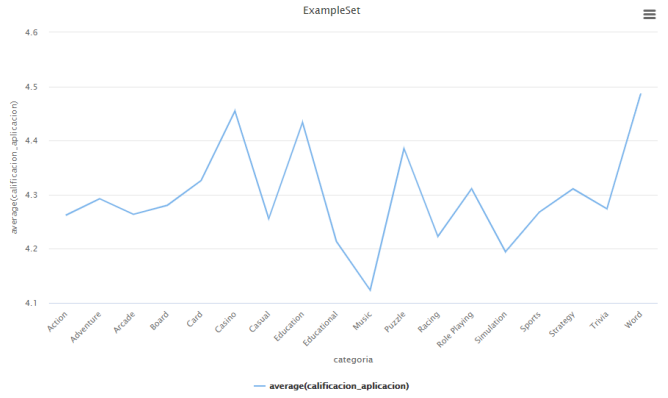
Promedio de Reviews por categoría:



Promedio de RI por categoría:



Promedio de calificaciones por categoría:



VII. CONCLUSIONES

Se concluye que este proyecto al abarcar un segmento específico de las aplicaciones otorgará un nuevo aporte o al menos una nueva perspectiva en el desarrollo y emprendimiento.

De trabajos anteriores y relacionados se rescata que los indicadores principales son: la cantidad de instalaciones, las calificaciones de los usuarios que ha recibido a lo largo de su vida útil, número de palabras en el nombre de la aplicación, el precio, el contenido (la idea de negocio y su utilidad).

Además, la herramienta RapidMiner ofrece un auto modelamiento de inteligencia artificial y limpieza de datos, pero con menos características que la limpieza de datos que ofrece open refine.

Toda página posee un archivo "robots.txt" o similares que indican un mapa completo del sitio en orden de aumentar su posicionamiento por parte de Search Engine Optimization, se puede usar esto también para realizar una recolección más completa de datos.

AGRADECIMIENTOS

Un agradecimiento a Gautham Prakash[1], el cual ha recolectado la principal base de datos a utilizarse en el proyecto usando Web Scrapping junto con Python.

REFERENCIAS

[1] I. Group, "Factores que marcan el éxito o el fracaso de una app", *Digitaltransformation.ituser.es*, 2016. [Online]. Available: https://digitaltransformation.ituser.es/noticias/2016/09/factores-que-marcan-el-exito-o-el-fracaso-de-una-app. [Accessed: 19- Nov- 2019].

[2] M. Nayebi, B. Adams, and G. Ruhe. Mobile appreleases – a survey research on developers and usersperception. InIEEE 23rd International Conference onSoftware Analysis, Evolution and Reengineering (SANER'16). IEEE, 2016.

[3] Martin, W., Sarro, F. and Harman, M. (n.d.). *Causal Impact Analysis for App Releases in Google Play*. [online] ucl. Available at: http://www0.cs.ucl.ac.uk/staff/f.sarro/resource/papers/Martin_FSE_Causal_PrePrint.pdf [Accessed 19 Nov. 2019].

[4] A. Mueez, K. Ahmed, T. Islam and W. Iqba, "Exploratory Data Analysis and SuccessPrediction of Google Play Store Apps", *Dspace.bracu.ac.bd*, 2018. [Online]. Available: http://dspace.bracu.ac.bd/xmlui/bitstream/handle/10361/11407/15101108%2C15101020%2C15101109%2C15141002_CSE.pdf?sequence=1&isAllowed=y. [Accessed: 19- Nov- 2019].

[5] C. Escudero Carazo, "Citylok, una app de eventos global", *Biblioteca.unirioja.es*, 2018. [Online]. Available: https://biblioteca.unirioja.es/tfe_e/TFE003059.pdf. [Accessed: 19- Nov- 2019].

[6] W. Martin, F. Sarro, Y. JiaYuanyuan Zhang and M. Harman, "A Survey of App Store Analysis for Software Engineering", *Cs.ucl.ac.uk*, 2016. [Online]. Available: http://www.cs.ucl.ac.uk/fileadmin/UCL-CS/research/Research_Notes/RN_16_02.pdf. [Accessed: 19- Nov- 2019].

[7] "Open Refine", *https://es.schoolofdata.org/2014/06/30/openrefine/*, 2014. [Online]. Available: https://es.schoolofdata.org/2014/06/30/openrefine/. [Accessed: 19- Nov- 2019].

[8] B. Fernández García, "Rapidminer: software data mining | Clarcat", *Clarcat*. [Online]. Available: https://www.clarcat.com/rapidminer/. [Accessed: 19- Nov- 2019].

[9] "Introducción a Beautiful Soup 4 con Python", *Clouding*, 2019. [Online]. Available: https://clouding.io/kb/introduccion-a-beautifulsoup. [Accessed: 19- Nov- 2019].

[10]"¿Qué es Tableau public?", tableauperu.com. [Online]. Available: https://tableauperu.com/que-es-tableau-public/ [Accessed: 27- Nov- 2019].

[1] Su perfil https://www.kaggle.com/gauthamp10 y la base de datos que recolectó https://www.kaggle.com/gauthamp10/google-playstore-apps