SERVICIO NACIONAL DE APRENDIZAJE

LABORATORIO DE INTELIGENCIA ARTIFICIAL SENA - HUAWEI

Reporte de Analítica de datos e inteligencia artificial

Científicos e ingenieros de datos

John Jairo Castro Maldonado, M.Sc, Ph.D (C)

Jesús Arbey Benavides, M.Sc, Ph.D (C)

Fredy Alberto Jaimes Muñoz, M.Sc

CENTRO DE SERVICIOS Y GESTIÓN EMPRESARIAL- REGIONAL ANTIOQUIA

MEDELLÍN, ANTIOQUIA

2022

Tabla de contenido

1. Identificación del problema 3
1.1 Planteamiento del problema 3
1.2 Pregunta de investigación 3
2. Objetivo 3
1.1 Objetivo general 3
3. Descripción de los datasets 4
3.1 Resumen de los datasets y diseño de tablas 4
3.2 Arquitectura de la solución 7
4. Análisis exploratorio de los datos 7
5. Modelamiento 23
5.1 Resultados de los modelos 23
6. Desarrollo de la aplicación 37
7. Referencias Bibliográficas 39

1. Identificación del problema

1.1 Planteamiento del problema

BETTER WITH DATA es una empresa especializada en marketing digital, que brinda servicios corporativos que van desde la asesoría hasta la producción de material publicitario y contenido creativo. Su modelo de negocio tiene como objetivo conectar a sus clientes con sus audiencias convirtiendo los datos en emociones y las emociones en resultados. Para ello, BETTER WITH DATA ofrece formas de diversificar el negocio de sus clientes modelos a través de estrategias de marketing inteligente. La compañía también cuenta con una unidad de negocios dedicada a atender las necesidades de los *influencers*, apelando a estrategias como *storytelling* y narrativas de la vida real para potenciar el vínculo de los seguidores con las marcas de sus clientes.

1.2 Pregunta de investigación

¿Cómo realizar una segmentación eficiente del mercado para los clientes de BETTER WITH DATA? Esto incluye identificar a los clientes potenciales de las empresas y comprender cómo hacer que sus comunidades digitales crezcan más rápido.

Teniendo en cuenta como referencia la anterior pregunta, para la investigación se platea los siguientes objetivos:

2. Objetivo

1.1 Objetivo general

Teniendo en cuenta las necesidades comerciales que están relacionadas con la comprensión de la audiencia, los nuevos territorios y el crecimiento de la comunidad, es importante determinar las mejores variables en los conjuntos de datos que se pueden modelar para responder a las expectativas de la agencia, soportando con los datos no estructurados de las redes sociales como Facebook, Instagram, Twitter y YouTube.

3. Descripción de los datasets

Esta sección describe todos los conjuntos de datos utilizados para el desarrollo de la solución propuesta. Utilizamos dos clientes de BETTER WITH DATA y 4 redes sociales para desplegar todo el contenido y EDA para la aplicación.

3.1 Resumen de los datasets y diseño de tablas

La siguiente tabla resume las principales características de los conjuntos de datos que tendremos en cuenta a lo largo del desarrollo de nuestro proyecto. En este caso es importante señalar la base de datos adicional de Youtube para Alberto Linero y solo 3 redes sociales para Juanita Kremer.



Alberto Linero

Alberto Linero: Es periodista, escritor, docente y conferencista internacional. Con sus mensajes a través de los distintos medios de comunicación y las redes sociales, logra impactar diariamente a millones de personas. Además de todo el trabajo que realiza presencialmente en las distintas ciudades con sus conferencias nacionales e internacionales.



Juanita Kremer

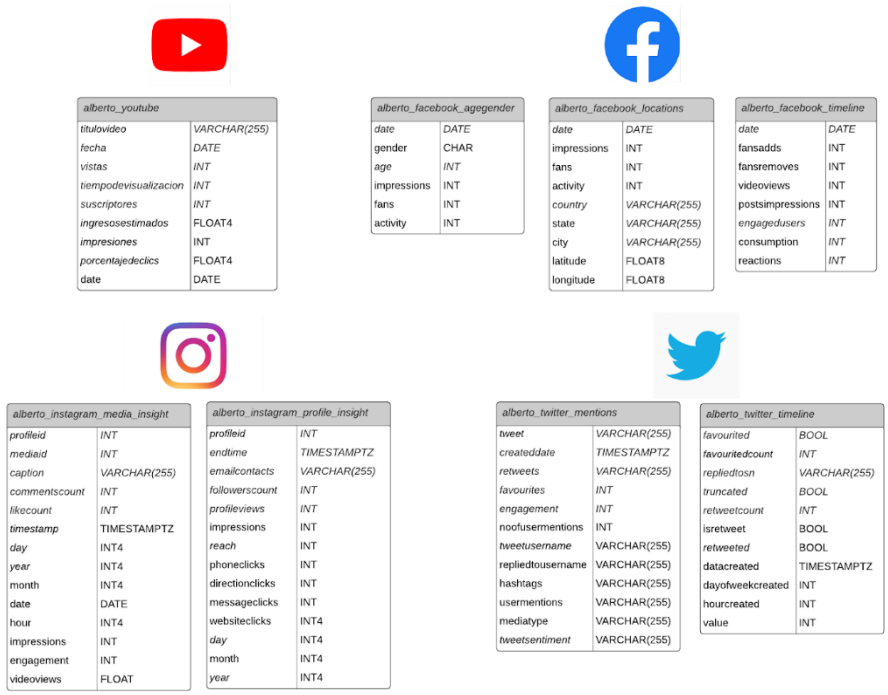
Juanita Kremer: Es periodista y locutora, con experiencia en el mundo tecnológico, en el que se ha convertido en referente. Actualmente es una de las líderes de opinión en tecnología del país, participando en un importante programa de una de las emisoras con más audiencias a nivel nacional. También comparte sobre este tema en sus redes sociales.

Name	From	To	Daily	Table name
Alberto Linero Youtube	2011	2021	X	alberto_youtube
Alberto Linero Facebook Age Gender	2019	2021	X	alberto_facebook_agegender
Alberto Linero Facebook Locations	2019	2021	X	alberto_facebook_locations
Alberto Linero Facebook Timeline	2019	2021	X	alberto_facebook_timeline
Alberto Linero Instagram Media Insight	2019	2021	X	alberto_instagram_media_insight
Alberto Linero Instagram Profile Insight	2019	2021	X	alberto_instagram_profile_insight
Alberto Linero Twitter Metions	2021	2021	X	alberto_twitter_mentions
Alberto Linero Twitter Timeline	2021	2021	X	alberto_twitter_timeline
Juanita Kremer Facebook	2018	2021	X	juanita_facebook
Juanita Kremer Instagram Media Insight	2017	2021	X	juanita_instagram_media_insight
Juanita Kremer Instagram Profile Insight	2017	2021	X	juanita_instagram_profile_insight

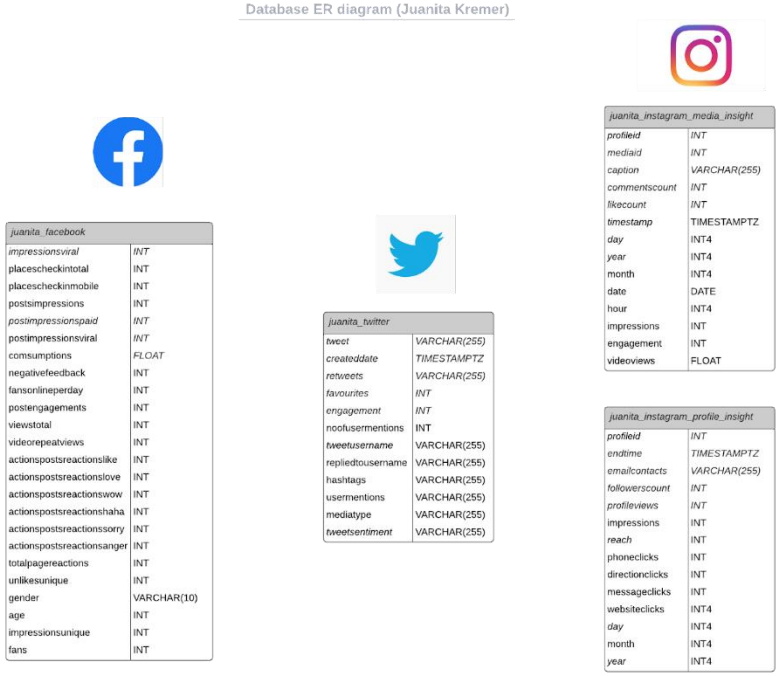
Tabla 1. Representa principales características de los conjuntos de datos

Los conjuntos de datos se cargaron en tablas de Postgres en AWS RDS. Las tablas relacionales equivalentes están mapeadas en el resumen que se incluyó anteriormente y se muestran a continuación:

Database ER diagram (Alberto Linero)

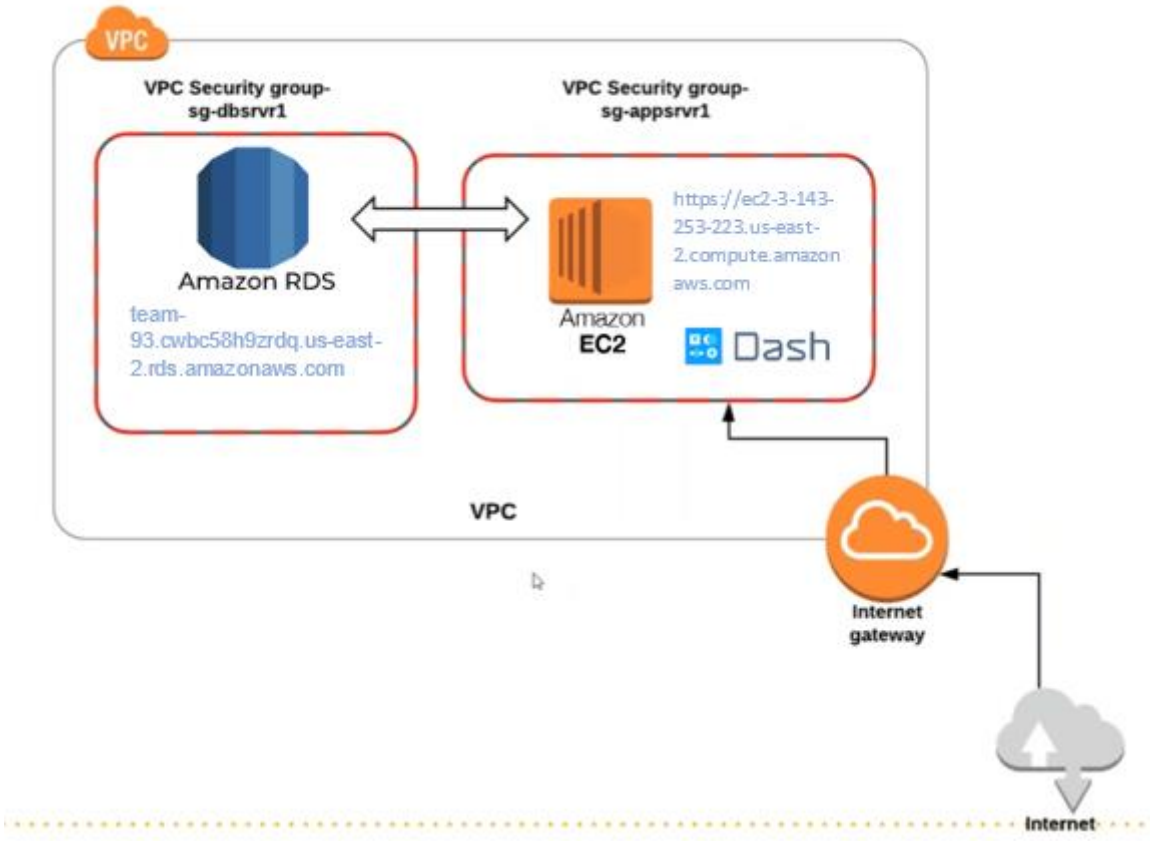


Database ER diagram (Juanita Kremer)



3.2 Arquitectura de la solución

La arquitectura fue un RDS de AWS para almacenar todo el conjunto de datos de Juanita Kremer y Alberto Linero, y la aplicación se desarrolló en AWS EC2 sobre Dash de la siguiente manera:



4. Análisis exploratorio de los datos

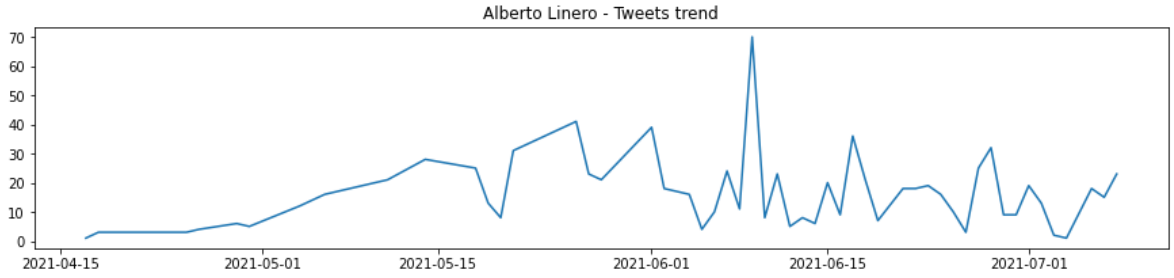
El principal objetivo de este análisis exploratorio de datos (AED) de nuestro proyecto es analizar y comprender a partir de la generación de diferentes gráficos estadísticos y cálculo de determinadas métricas, patrones de comportamiento y características conjuntas de determinados 'talentos' como Alberto Linero y Juanita Kremer, en las diferentes redes sociales que gestionan (Twitter, Instagram, Facebook y YouTube).

Estos son algunos resultados de nuestro interés por red social:

Exploración de bases de datos de Twitter:

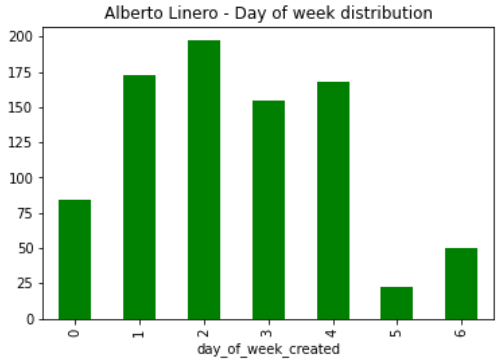
Alberto Linero – Cronología

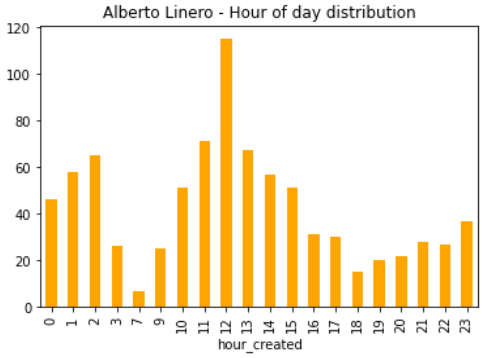
Tendencia de tweets por fecha



El conjunto de datos de la línea de tiempo corresponde a un período de cuatro meses desde mediados de abril hasta la primera semana de mayo, donde podemos ver una tendencia de aumento de abril a junio y hay un pico significativo de tweets (alrededor de 70) en la segunda semana de junio. Después de ese período, el comportamiento de los tweets tiene una tendencia a la baja y podemos ver otros picos en días particulares.

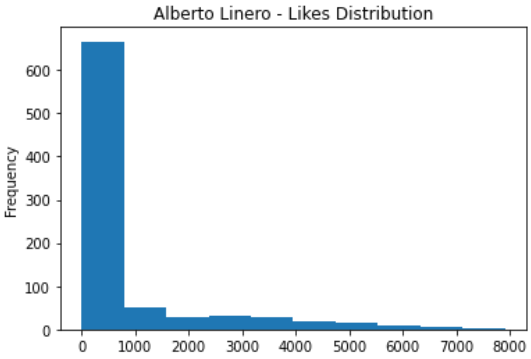
Tweets trend by day of week and hour

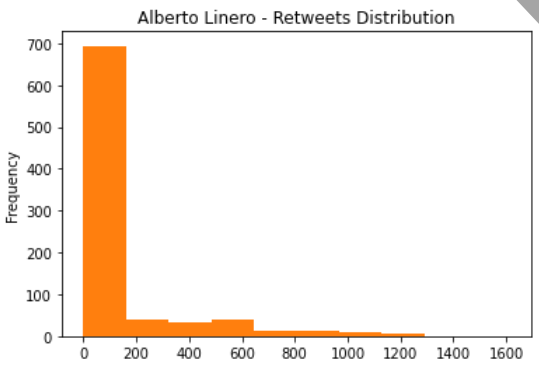




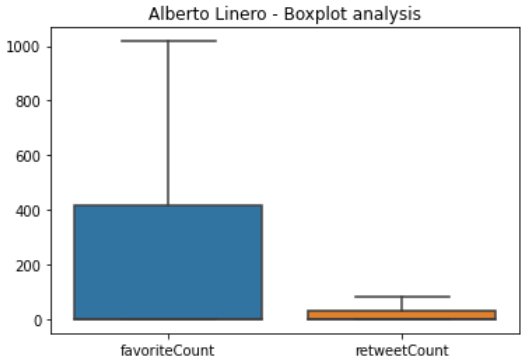
Cuando analizamos el comportamiento de los tweets por día de la semana y hora, hay algunos hallazgos interesantes. En cuanto al día de la semana, hay un mayor nivel de tweets en los días laborables en comparación con el fin de semana, donde el nivel de tweets cae significativamente. El día con más publicaciones es el miércoles, por otro lado, cuando analizamos el comportamiento horario nos dimos cuenta que Alberto Linero solía twittear en horario laboral, de 10:00h a 15:00h con un pico importante al mediodía y volvía a postear en horas muy tempranas como como 01:00h.

Distribución de Me gusta y Retweets de Tweets



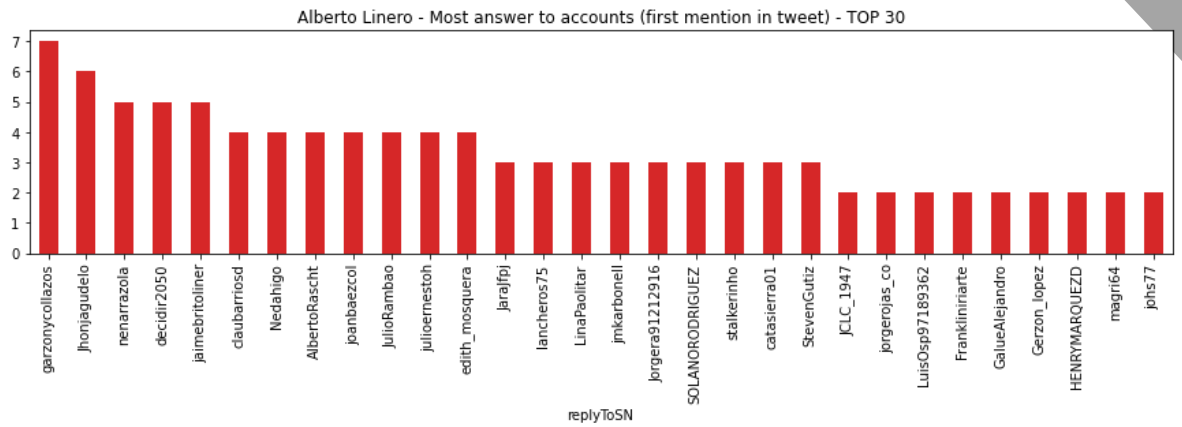


	favoriteCount	retweetCount
count	849.000000	849.000000
mean	674.305065	106.987044
std	1420.956142	242.029637
min	0.000000	0.000000
25%	0.000000	0.000000
50%	2.000000	0.000000
75%	414.000000	33.000000
max	7912.000000	1613.000000



Las variables de me gusta y retweets tienen una distribución muy sesgada hacia la derecha, como podemos ver en los gráficos. Los me gusta se concentran por debajo de 1.000 reacciones y los retweets se concentran por debajo de 200. Cuando analizamos las descripciones estadísticas, confirmamos que las distribuciones de las variables debido a la media son superiores a la mediana lo que explica el comportamiento sesgado a la derecha encontrado. Como podemos ver en los valores máximos y el diagrama de caja, hay valores atípicos en ambas variables, lo que significa que hay tweets con una gran participación debido a la gran cantidad de me gusta y retweets.

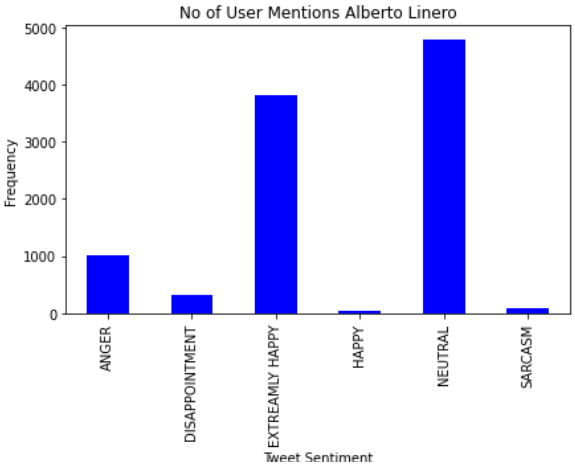
Cuentas más interactuadas

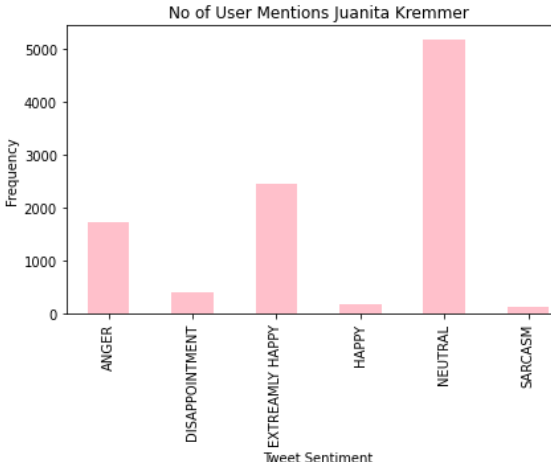


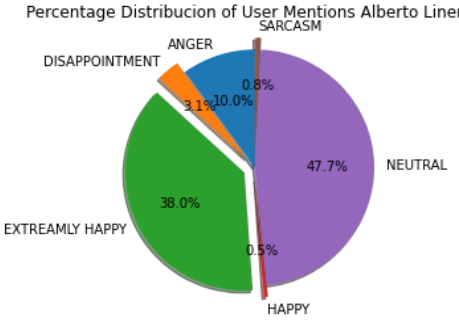
Alberto Linero mencionó primero en sus respuestas a la cuenta “@garzonycollazos” siete veces, seguido de “@Jhonjagudelo” quien fue mencionado seis veces. Como podemos ver no existe una cuenta que concentre de manera significativa el nivel de respuestas. Esto se distribuye a más de 400 usuarios a los que había respondido al menos una vez.

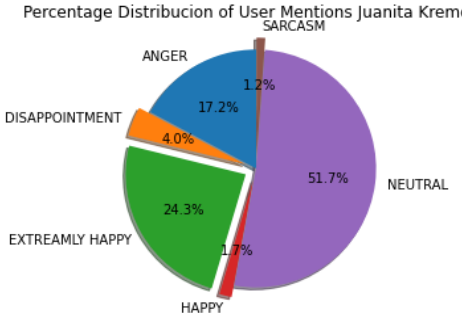
Alberto Linero/ Juanita Kremer – Menciones

Número de menciones de usuarios por sentimiento de tweet

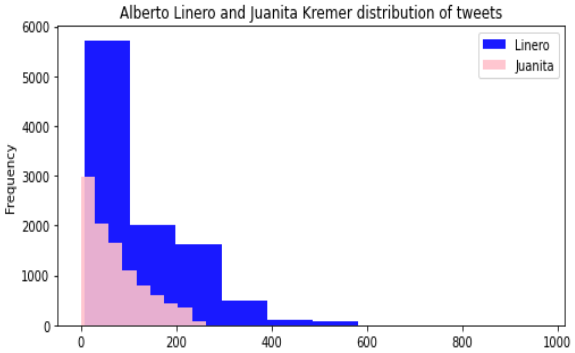






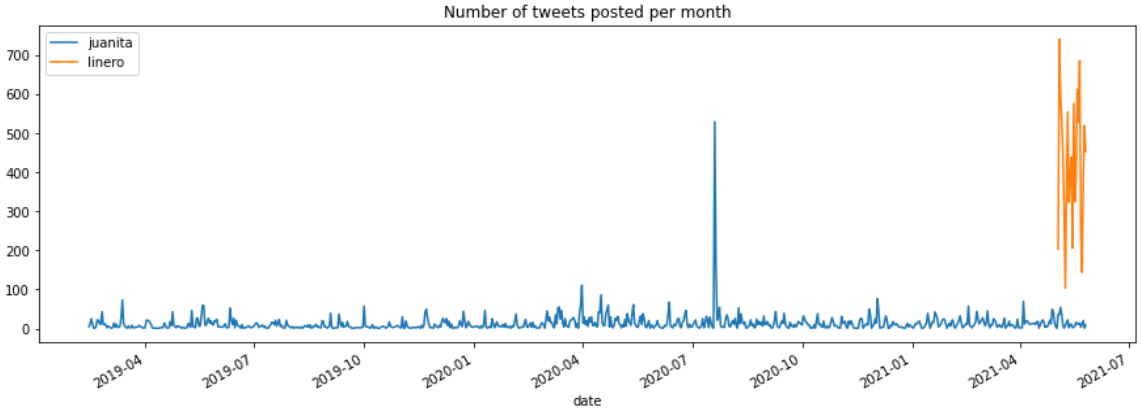


Haciendo una comparación entre el número de menciones de usuarios por tipo de sentimiento, se observa que en general tanto Alberto como Juanita son muy similares, sin embargo, hay una diferencia notable en cuanto a la categoría extremadamente feliz a favor de Alberto L. con una diferencia porcentual de aproximadamente 14%. Tanto para Alberto como para Juanita, aproximadamente la mitad de sus tuits se expresan de forma neutral. Finalmente, tus tweets que expresan algo negativo puntúan bajo.




```
Tweets_Linero
mean    118.623435
std     111.990553
Name: Tweet, dtype: float64
Tweets_Juanita
mean    73.584064
std     60.605663
Name: Tweet, dtype: float64
```

En el gráfico anterior se puede observar que la distribución para ambos talentos tiene un sesgo hacia la derecha bastante pronunciado, donde el promedio de tuits para Alberto es de 118, mientras que para Juanita es de aproximadamente 73 tuits.



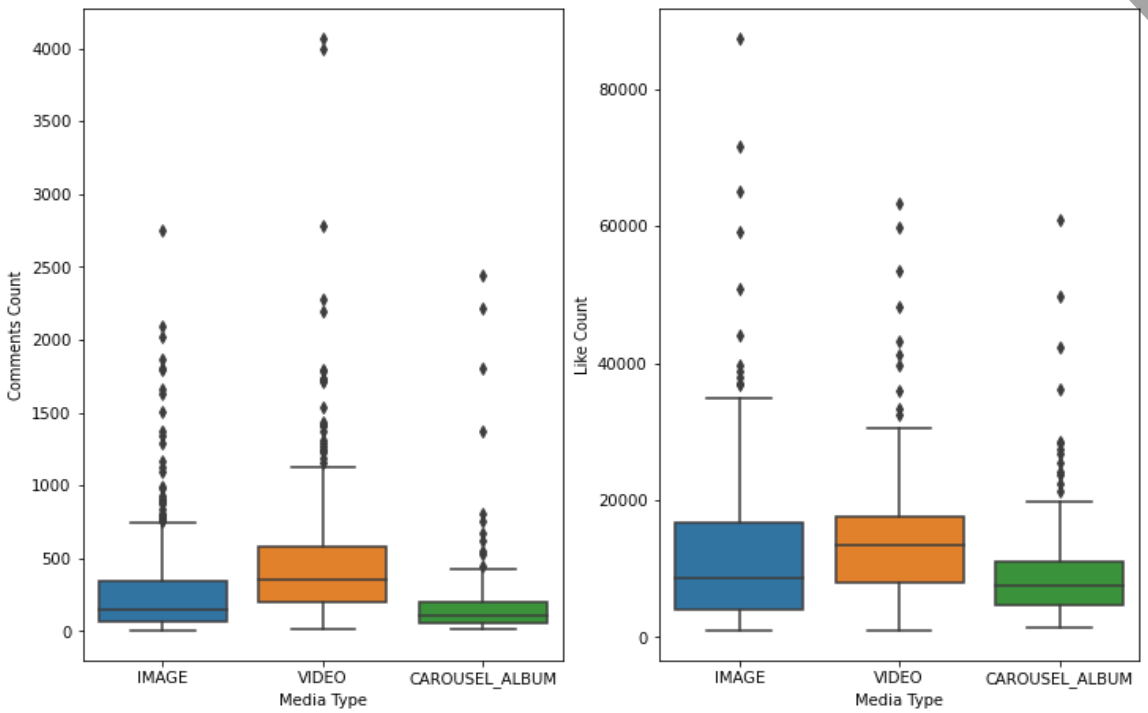
Se puede observar un perfil de actividad diferente para cada talento. Juanita ha mantenido una actividad constante por debajo de los 100 tuits mensuales durante todo el periodo estudiado, sin embargo, tuvo un ascenso bastante pronunciado en agosto de 2020 por alguna razón. Alberto, que empezó en el mundo del twitter dos años más tarde que Juanita, pero se sabe que publica con mucha más frecuencia.

1. Instagram databases exploration:

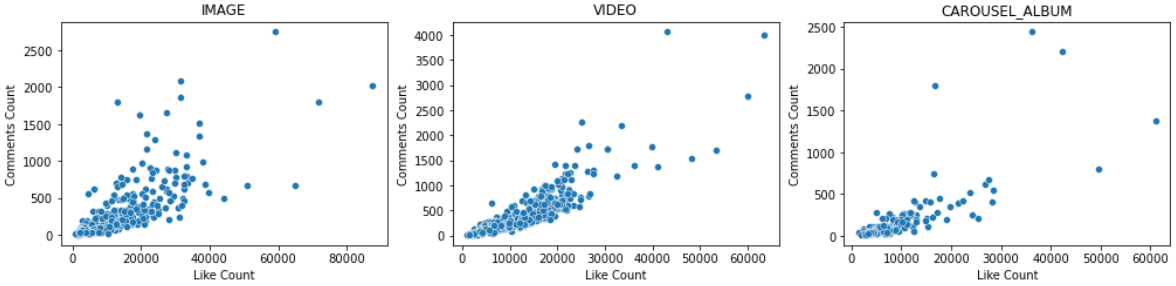
Exploración de bases de datos de Instagram:

Alberto Linero

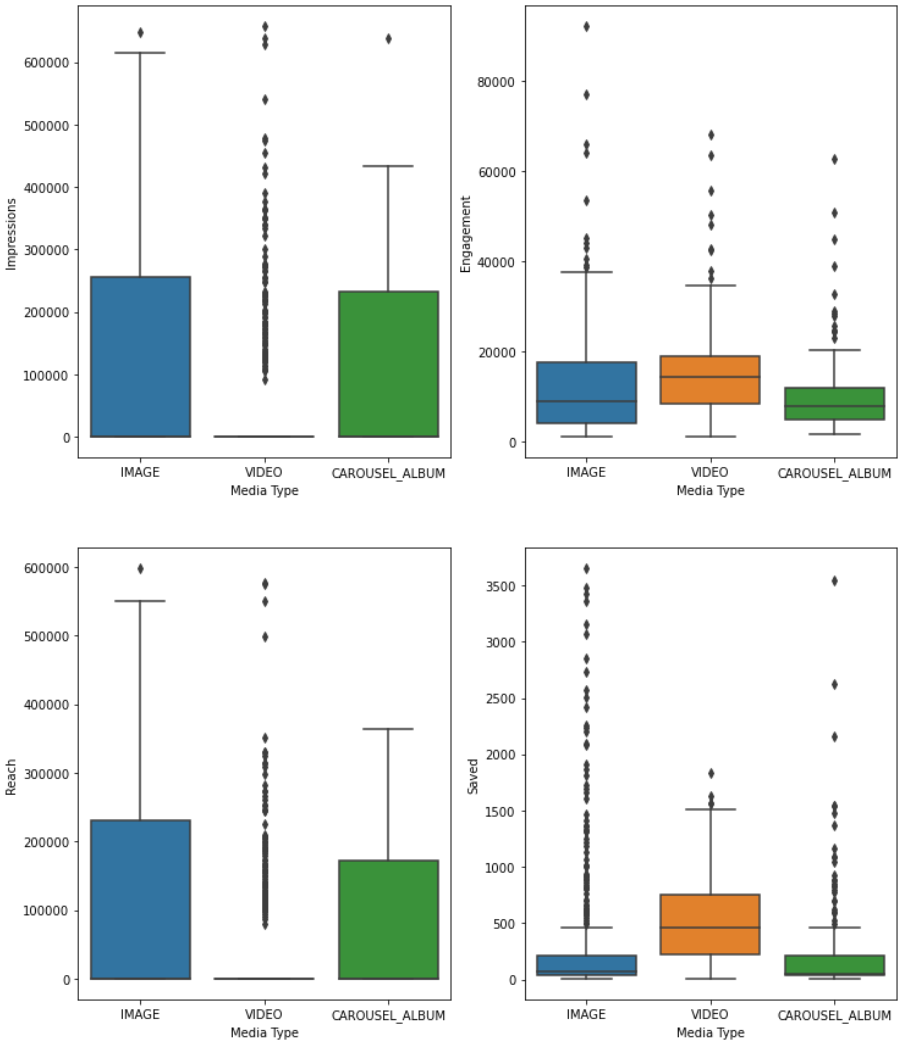
Distribución del conteo de Me gusta y Comentarios



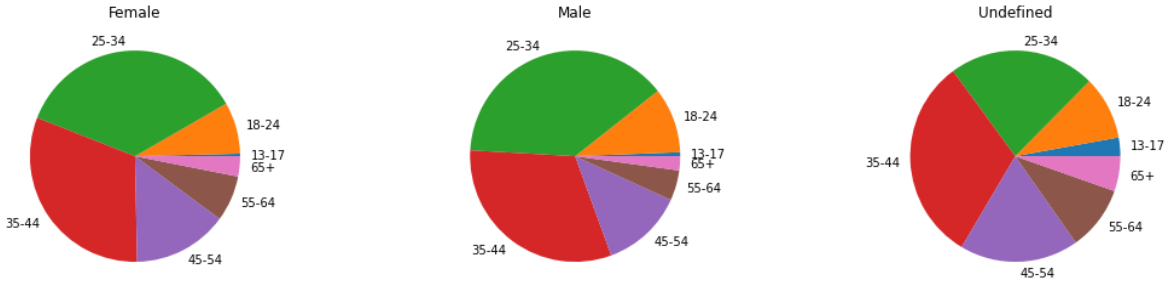
Es fácil notar en la Figura anterior cómo tanto el recuento de Me gusta como el de comentarios tienen una distribución muy sesgada hacia la derecha, con solo unos pocos registros con una cantidad extremadamente alta de comentarios y Me gusta.



Ahora, al analizar los diagramas de dispersión bivariados, está claro en la Figura anterior cómo los Me gusta y los comentarios cuentan, están ligeramente correlacionados. El siguiente paso para identificar la verdadera relación entre estas dos variables sería hacer un análisis de sentimiento para discriminar entre buenos y malos comentarios.

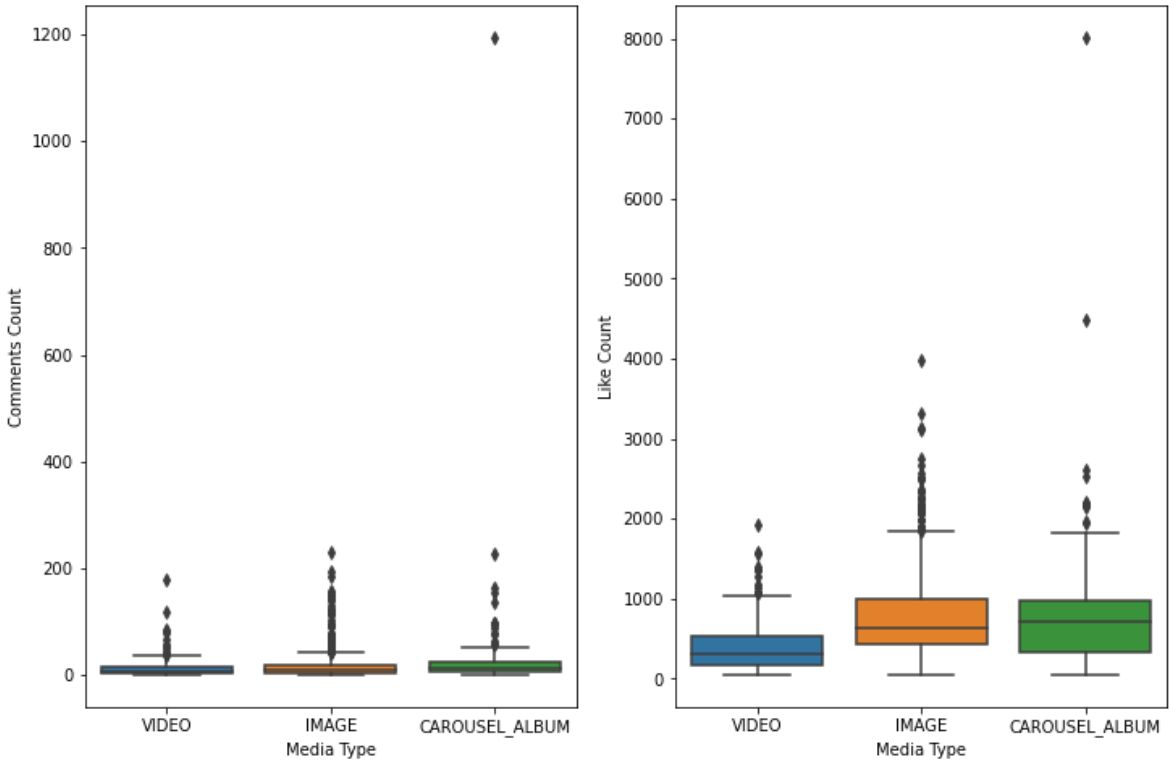


Con la figura anterior es interesante cómo las impresiones y el alcance tienen una distribución ligeramente similar, por lo tanto, se recomienda ahora hacer un análisis de correlación entre estas dos variables, para ver si realmente están correlacionadas o si es solo una coincidencia que ambas distribuciones son muy similar.

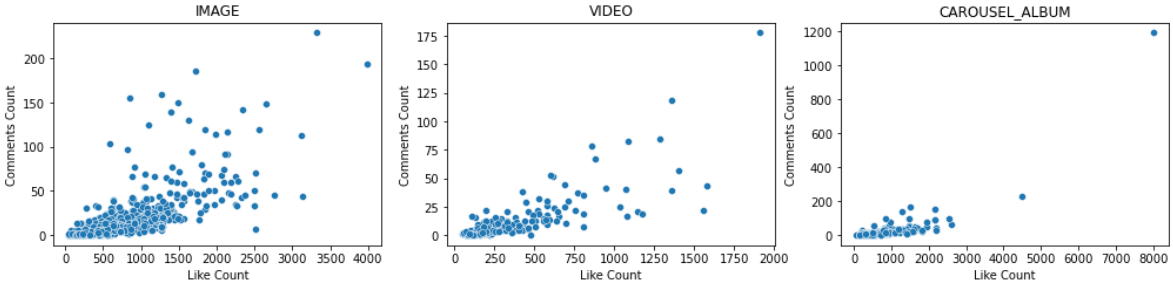


Finalmente, es bastante obvio cómo la mayoría de seguidores se encuentran entre 25 y 44, con el menor número entre 13 y 17. Es de esperar ya que el contenido de Alberto se centra principalmente en consejos espirituales y los adolescentes no suelen estar tan preocupados por este aspecto en sus vidas.

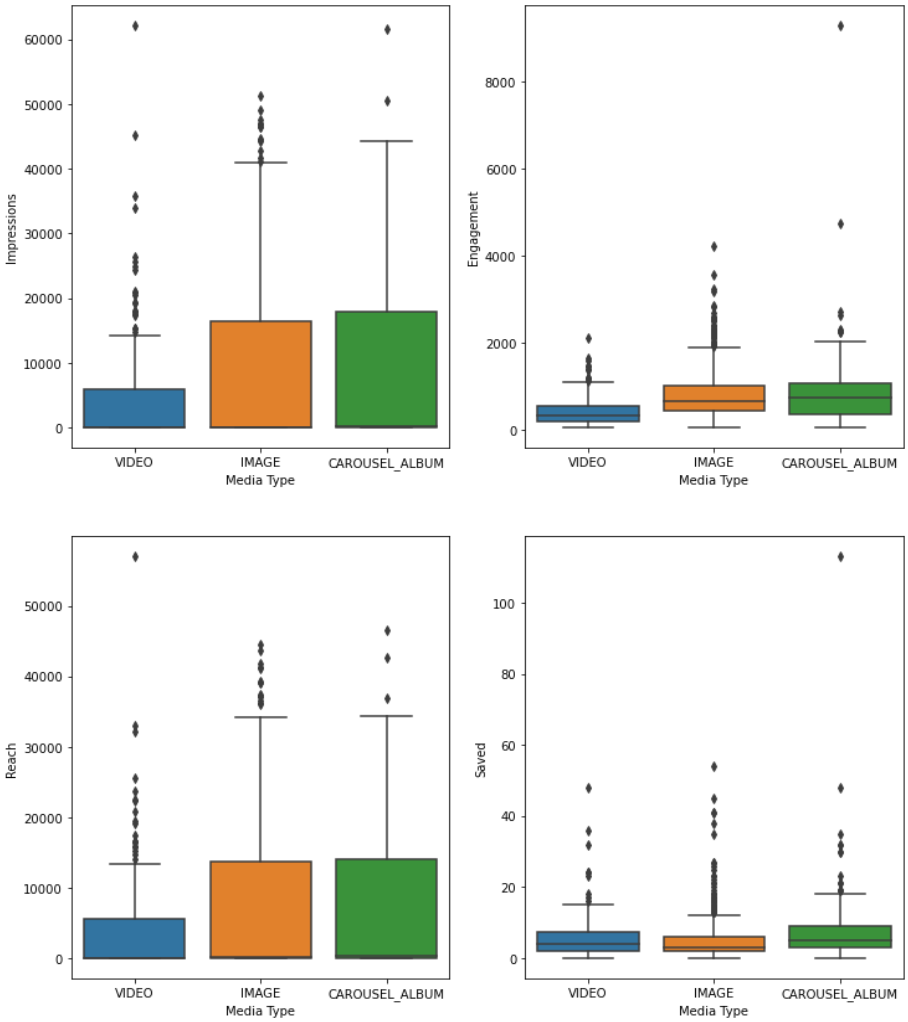
Juanita Kremer:



Con Juanita tenemos el mismo caso que con Alberto, que muestra una distribución muy sesgada hacia la derecha para los comentarios y los recuentos de me gusta, pero ahora tenemos un valor atípico extremo en un álbum de carrusel, esto es interesante y es importante analizar la causa de este caso raro.



En los diagramas de dispersión bivariados muestra el mismo comportamiento que el diagrama de Alberto, y no agrega información importante, más que la que dio el diagrama de Alberto.

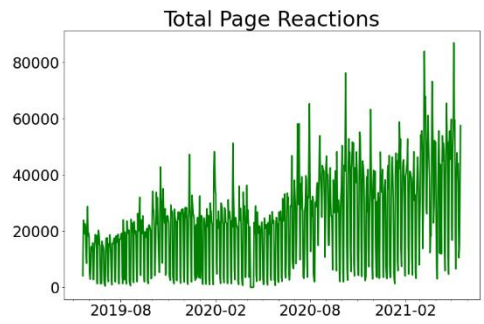


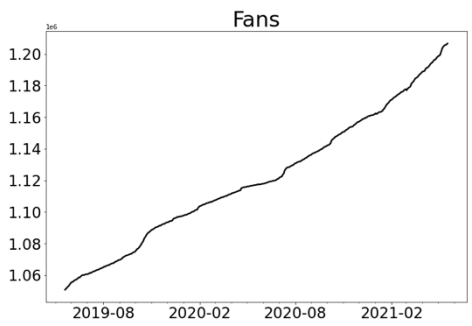
Y finalmente, en los diagramas de caja para Impresiones, guardados, alcanzados y Compromiso, es interesante cómo nuevamente dimos la misma distribución para Alcance e Impresiones, por lo tanto, tenemos más evidencia para suponer una alta correlación y dependencia entre estas dos variables.

Exploración de bases de datos de Facebook:

Alberto Linero – Cronología

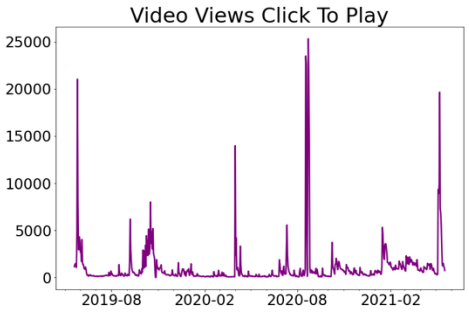
Recibimos tres dataframes, que contienen el comportamiento de varias variables en un período de tiempo de 3 años, desde mayo/2019 hasta mayo/2021. El marco de datos principal contiene 60 columnas, cada fila es un día diferente. Estas columnas tienen mucha información sobre la actividad en la página de fans de Facebook. Se muestran algunas tramas iniciales interesantes:

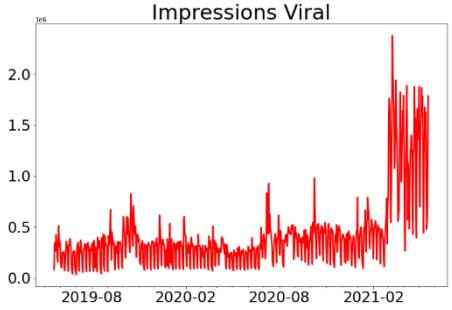




Los gráficos de arriba representan la cantidad de fanáticos (izquierda) en millones y la cantidad de reacciones diarias (derecha). Ha habido un crecimiento constante en el número, aunque se nota una gran fluctuación en el número de reacciones de un día a otro. Será interesante correlacionar estos números con la cantidad y el tipo de publicaciones publicadas en diferentes fechas, para comprender cómo la cantidad de reacciones puede verse empujada hacia el extremo superior de la distribución de valores. Sin embargo, la amplitud de las fluctuaciones es sospechosamente alta, por lo que será importante discutir más con los representantes de Maracuyá sobre el verdadero significado para descartar la posibilidad de que la fluctuación sea un artefacto debido a un factor que aún no se comprende.

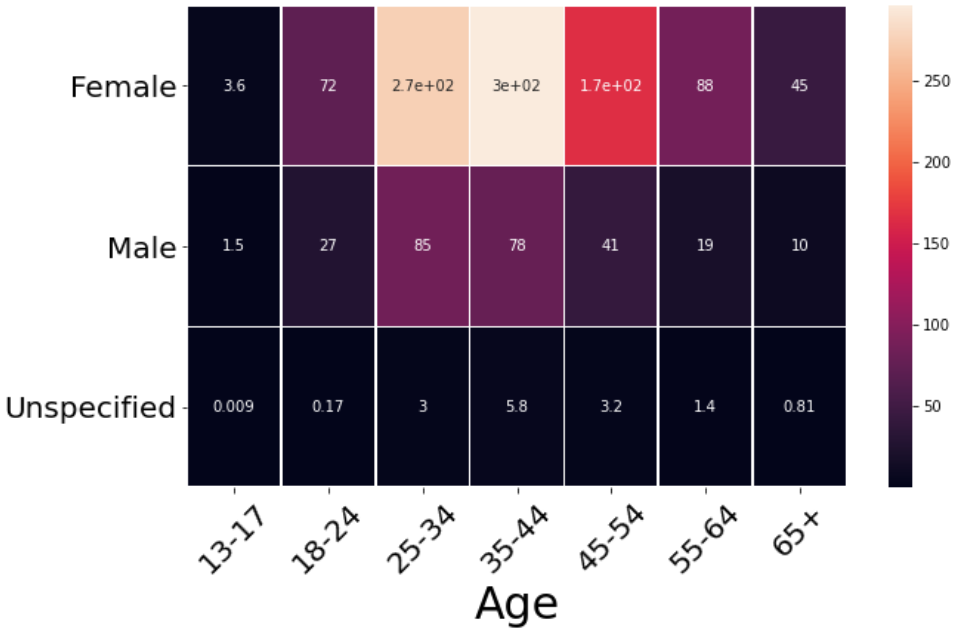
Ahora un par de ejemplos de otro tipo de tramas que parecen prometedoras:





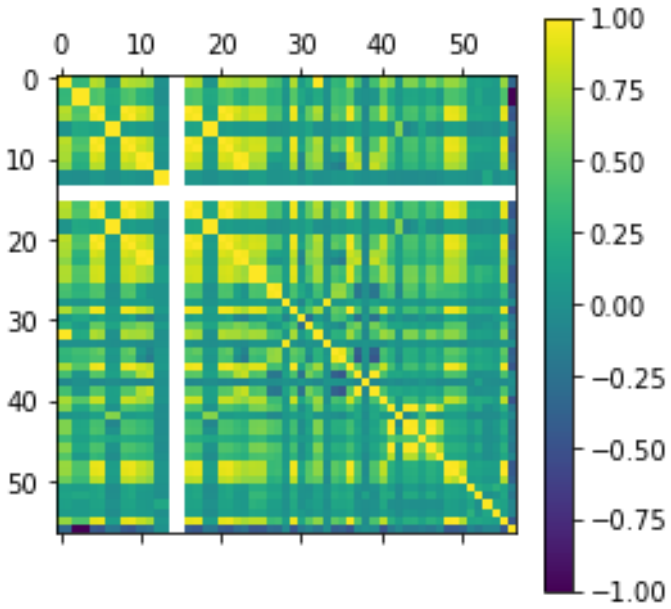
Al comprender qué factores hicieron que las publicaciones tardías de 2021 se volvieran virales (izquierda) y cuáles motivaron tantos clics para reproducir los videos (derecha) en las fechas en las que aparecen los picos, se pueden hacer recomendaciones para mejorar el factor viral de las publicaciones. y aumentar el compromiso con el público.

Los otros marcos de datos que recibimos contienen menos información, pero con un mayor grado de granularidad: están segmentados por género y por rango de edad. Para entender mejor a la audiencia, se creó un mapa de calor del número total de fans (en unidades de miles) por las diferentes categorías:



El grueso de la afición son mujeres de entre 25 y 54 años. La segunda contribución más importante se divide equitativamente en 4 categorías: mujeres en los contenedores vecinos de edad al grupo principal y hombres en los contenedores de edad [25-34] y [35-44].

También se exploró un mapa de calor que muestra el coeficiente de correlación lineal entre todas las variables numéricas (izquierda). Muestra que hay varias variables correlacionadas, lo que no es del todo una sorpresa (por ejemplo, el número de reacciones y el compromiso se correlacionan bien), pero de todos modos es interesante medirlo y comprenderlo en detalle.



City Only	Impressions Unique	Fans
Bogotá	213755	220381
Barranquilla	148349	94561
Medellín	59961	70433
Cartagena	81737	54068
Santiago de Cali	38847	44476
Bucaramanga	35971	40627
Montería	38087	24840
Cúcuta	21312	22406
Santa Marta	34101	21757
Villavicencio	15636	16298

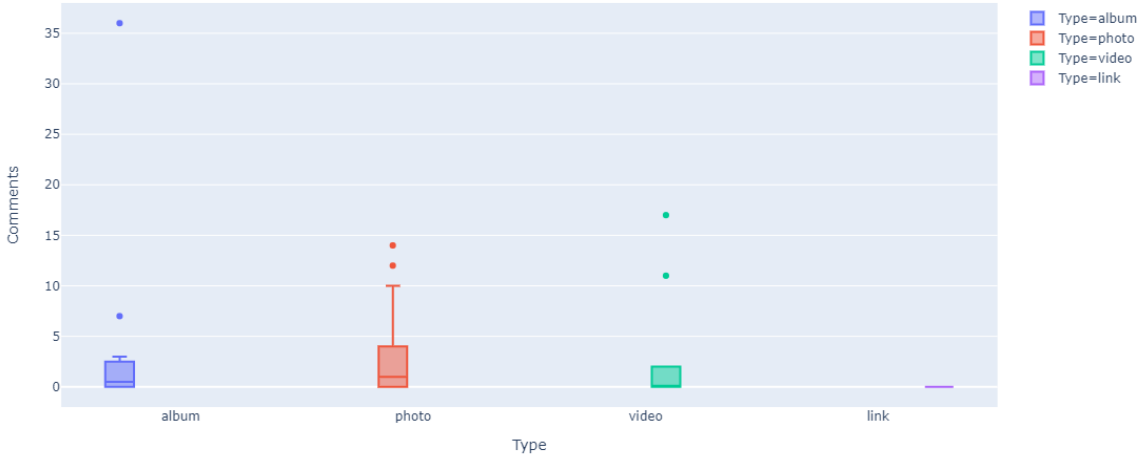
El tercer marco de datos proporciona información desagregada por referencia geográfica. Después de algunas manipulaciones, se determinaron las 10 ciudades con la mayor cantidad de fanáticos (derecha). Parecen ser las ciudades con la mayor cantidad de impresiones también, aunque curiosamente, la cantidad de impresiones no es estrictamente monótona en el mismo orden que la cantidad de fanáticos.

Juanita Kremer – Cronología

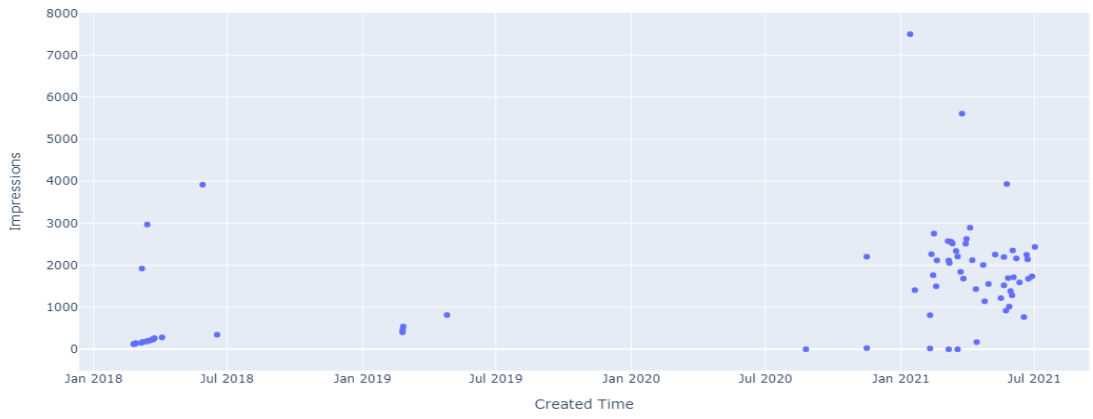
Se analizaron las bases de datos que ofrece la empresa para el talento Juanita Kremer, conformadas por tres marcos de datos que contienen variables entre enero de 2018 a julio de 2021. Los marcos de datos se constituyen con sus respectivos nombres en la siguiente tabla:

Filename	(Rows,Columns)
INSIGHTS_EDADES-_FACEBOOK_JUANITA_KREMER.csv	(4054, 7)
Facebook Juanita - Page .csv	(731, 58)
Facebook Insights Post - Juanita Kremer.xlsx	(78, 23)

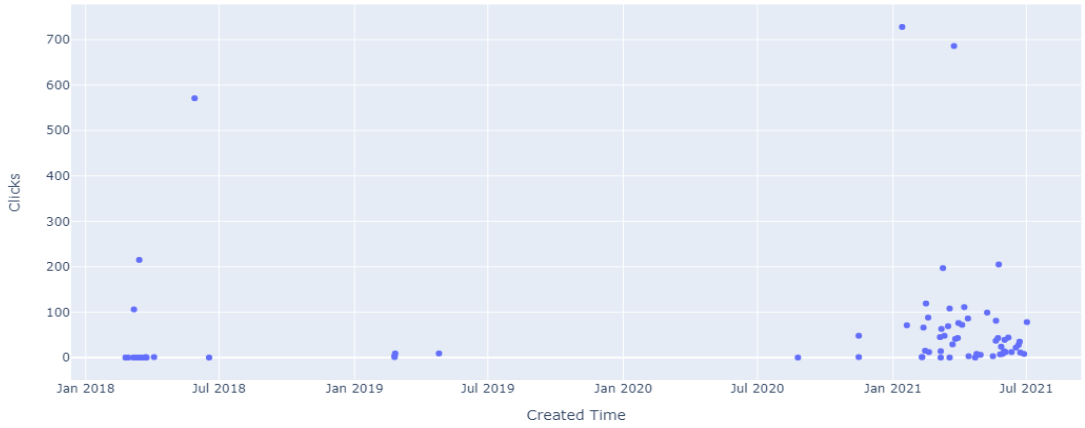
Cada uno de estos archivos contiene bastantes variables que describen la actividad en esta red social por parte de este talento. Así, para esta primera etapa, consideraremos algunos de ellos de los cuales haremos la interpretación correspondiente.



Esta cifra describe el número de comentarios sobre las publicaciones y el tipo de publicaciones. Podemos ver claramente que en relación a Facebook para Juanita Kremer se observa variabilidad con respecto al promedio para publicaciones asociadas a fotos, a diferencia de álbumes y videos que presentan menor desviación. Se observa que los enlaces publicados en la página tienen un bajo impacto en las publicaciones.



Esta cifra describe el número de impresiones en publicaciones a lo largo del tiempo. Se observa que la actividad correspondiente a impresiones en Facebook entre los periodos de enero de 2018 a enero de 2021 es baja. Sin embargo, entre enero de 2021 a julio de 2021 presenta mayor actividad, lo que se muestra en el crecimiento de los datos en este intervalo de tiempo.



Esta cifra describe el número de clics en las publicaciones a lo largo del tiempo. Se observa que la actividad correspondiente al número de clics en Facebook entre los periodos de enero de 2018 a enero de 2021 es baja. Sin embargo, entre enero de 2021 a julio de 2021 presenta mayor actividad, lo que se muestra en el crecimiento de los datos en este intervalo de tiempo. Además, a partir de esto se puede determinar que las impresiones están relacionadas con la cantidad de interacciones asociadas con las publicaciones creadas por este talento.

En general, se puede concluir para esta primera etapa de análisis que la actividad de este talento con respecto a Facebook es baja, esto relacionado con que tanto el tipo de publicaciones, interacciones y número de clics es bajo, y que se observó un ligero crecimiento entre enero de 2021 y julio de 2021.

Exploración de la base de datos de Youtube:

Recibimos un conjunto de datos que contenía la cantidad de vistas, suscriptores e impresiones de todos los videos en el canal de Alberto Linero.

Alberto Linero

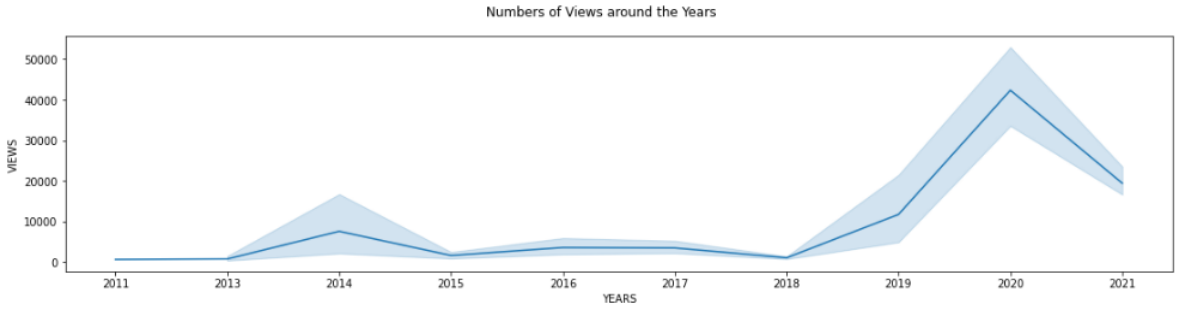
El período de tiempo es del 15/12/2011 al 16/08/2021. El conjunto de datos contiene 501 filas y 8 columnas, siendo cada fila un video diferente con sus estadísticas.

Se muestran algunas tramas iniciales interesantes:

	TITULO VIDEO	FECHA	VIEWS	Tiempo de visualización (horas)	Suscriptores	Ingresos Estimados	Impresiones	Porcentaje de clics de las impresiones (%)
0	Padre Alberto Linero - Bienaventuranzas - Jeru...	15/12/11	618	558.968	3	0.534	7690	5.19
1	Cuaresma: una nueva oportunidad para la vida -...	23/02/13	1491	3.665.091	5	3.969	18174	5.92
2	Católicos fariseos - Reflexión dominical - Alb...	21/11/13	384	861.849	3	1.019	4318	6.07
3	¿Qué es eso de la conversión? - Reflexión domi...	16/12/13	362	1.113.254	1	0.921	4506	5.28
4	Motivaciones para caminar 40 días - Catequesis...	26/03/14	1045	3.323.365	0	3.849	9392	7.70
...	...	...	...	...	...	...	...	...
496	Cómo vencer el estrés | Alberto Linero ft. Dr....	24/06/21	10960	16.708.047	35	23.449	112959	4.06
497	Fuertes, libres y nómadas | Alberto Linero ft....	01/07/21	11909	23.436.102	25	27.178	137729	3.38
498	Eres fuerte por su amor | Alberto Linero | Ref...	03/07/21	11435	27.049.687	14	41.874	106302	5.39
499	Vencer en nuestras batallas - Alberto Linero -...	16/08/21	23163	73.103.788	99	28.952	276854	5.08
500	NaN	NaN	10511313	20.247.721.801	134729	15.681.711	113858536	4.86

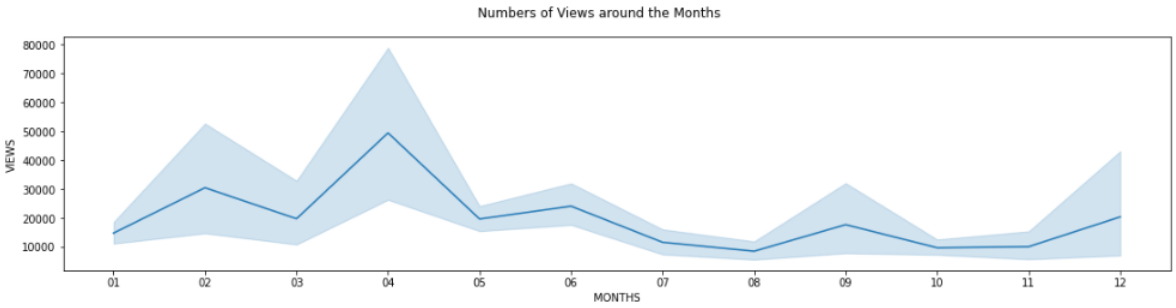
Vistas de Youtube por año

En el caso de las visualizaciones por año, podemos ver el trabajo que ha realizado el equipo para incrementar las visualizaciones en el canal, ya que se ha observado un aumento importante en los últimos dos años, cabe mencionar que esto también puede deberse a el efecto covid, que ha hecho que muchas personas tengan más tiempo para estar en casa y dedicar más tiempo a descansar.



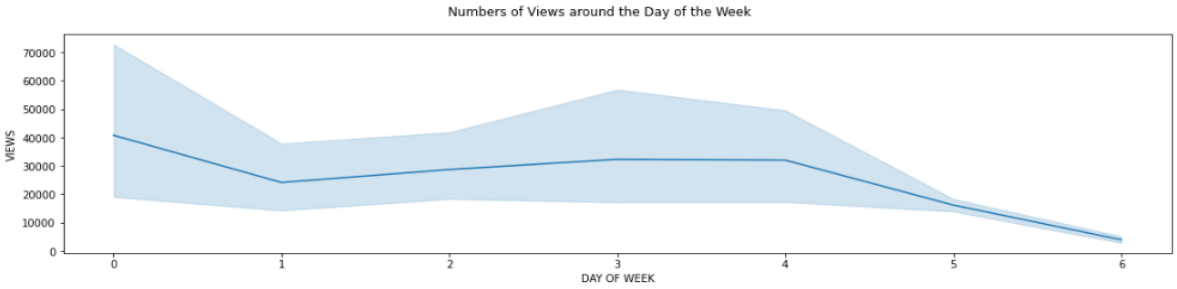
Vistas de Youtube por Meses

Alrededor de los Meses, es importante señalar que Alberto Linero tiene un repunte importante todos los años por Semana Santa, un aumento a fin de año y una tendencia para el mes de febrero. Es importante analizar ¿por qué estos meses Alberto Linero tiene un aumento en el número de vistas?



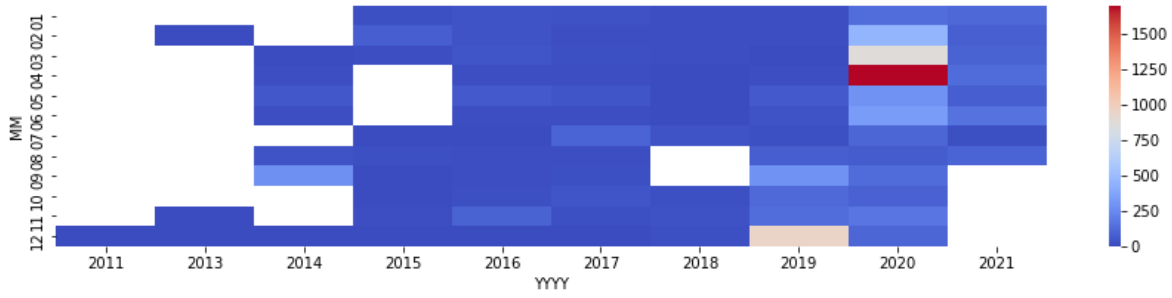
Vistas de Youtube por día de la semana

En este gráfico es fácil ver las tendencias de las vistas en el canal alrededor del día de la semana. Puede ver que las vistas más altas son el primer día de la semana (lunes) y cerca del final de la semana (jueves y viernes). Es importante ver que la menor vista es para el fin de semana posiblemente porque la gente que sigue a Alberto Linero está fuera de casa, tomando un descanso, y la segunda razón es porque el canal no sube un nuevo video para este período de tiempo.



Suscriptores de Youtube

El último gráfico importante son las tendencias de suscriptores alrededor de los meses y años para el canal Alberto Linero, se puede ver el crecimiento de suscriptores más importante solo para el inicio de la pandemia, un momento que claramente ha sido muy utilizado por el canal y que sin duda refleja la importancia del contenido y lo que significa para las personas que lo siguen.



5. Modelamiento

5.1 Resultados de los modelos

Teniendo en cuenta que las necesidades comerciales están relacionadas con la comprensión de la audiencia, los nuevos territorios y el crecimiento de la comunidad, debemos considerar las mejores variables en los conjuntos de datos que se pueden modelar para responder a las expectativas de la agencia. La mayoría de los conjuntos de datos no tienen información estructurada para obtener información sobre la audiencia, pero la mayoría de las redes sociales como Facebook, Instagram, Twitter y YouTube tienen información enriquecida oculta en datos no estructurados (comentarios, tweets, descripciones de perfiles, títulos de videos), por lo que Encontramos en este tipo de datos una gran oportunidad para proporcionar hallazgos perspicaces que permitan a la agencia tomar las mejores medidas para mejorar el desempeño de sus talentos en las redes sociales.

Una de las mejores opciones para dar respuesta al problema empresarial a través de datos no estructurados es clasificar cada palabra de un texto diferente en varios grupos o clusters que aporten información relevante para definir las oportunidades en términos de audiencia, rendimiento y estrategias de crecimiento. Por esta razón, elegimos el modelo de agrupamiento no supervisado K-Means como la forma más óptima de clasificar palabras. Además, nos gustaría comprender las palabras con mejor desempeño, por lo que elegimos la regresión logística para clasificar las palabras en las de mejor o peor desempeño.

Modelo K-means: para procesar los datos de aprendizaje, el algoritmo K-means en la minería de datos comienza con un primer grupo de centroides seleccionados al azar, que se utilizan como puntos de inicio para cada grupo, y luego realiza cálculos iterativos (repetitivos) para optimizar las posiciones de los centroides

Regresión logística: la regresión logística es un proceso de modelado de la probabilidad de un resultado discreto dada una variable de entrada. La regresión logística más común modela un resultado binario; algo que puede tomar dos valores como verdadero/falso, sí/no, etc.

Proceso de modelado

Para entrenar los modelos descritos anteriormente es necesario preparar los datos para que los modeladores puedan entender los valores de entrada, ya que vimos que los datos no estructurados no son legibles en primera instancia. Por esta razón, necesitamos hacer un proceso de transformación antes de ajustar el modelo que se describe a continuación:

Texto de limpieza: El texto a modelar pasó previamente por un proceso de limpieza. Comenzó con la eliminación de todas las menciones en el texto, es fácil identificar las menciones debido a que todas ellas se identifican con el carácter "@". Después de eso, aplicamos otra regla de limpieza, donde eliminamos todo el ruido en el texto, como puntuación, números, caracteres especiales y URL. Finalmente, convertimos todos los textos a minúsculas. De esta manera obtuvimos un inventario de palabras limpio y normalizado. Este proceso fue posiblemente gracias a las herramientas NPL como el paquete Python regex.

Vectorización de texto: aunque el texto se limpia y se normaliza, sigue siendo ilegible para los modelos de aprendizaje automático debido a que el texto está en formato de cadena. Para entrenar estos datos, necesitamos convertir todo el texto a un formato legible. Por esta razón, utilizamos el método denominado Frecuencia de Término-Frecuencia de Documento Inversa (TF-IDF). Esto evalúa qué tan importante es una palabra para un documento dentro de una gran colección de documentos (es decir, corpus). La importancia aumenta proporcionalmente en función del número de veces que aparece una palabra en el documento, pero se compensa con la frecuencia de la palabra en el corpus. Este proceso devuelve una matriz utilizando la transformación de ajuste del método que está lista para el entrenamiento del modelo.

TFIDF

For a term i in document j:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

tf_{ij} = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

TFIDF

For a term i in document j:

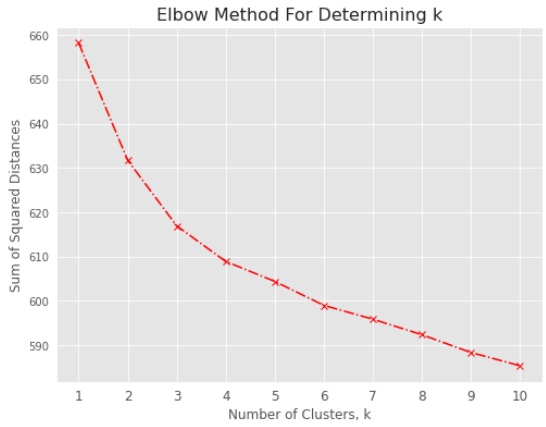
$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

tf_{ij} = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

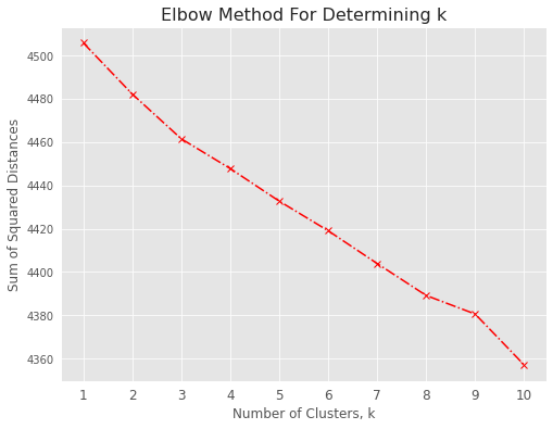
Dividir conjunto de datos: para ajustar los modelos K-means y Logistic Regression necesitamos dividir el conjunto de datos en subconjuntos de entrenamiento y prueba, para entrenar nuestro modelo trabajamos con el 80% de los datos y probamos los resultados con el 20% del conjunto de datos. Esto nos permite confirmar que los modelos son efectivos no solo para datos entrenados.

Obtenga el número de conglomerado óptimo para K-Means: una vez que tengamos las palabras en un formato legible, podemos entrenar varios modelos. En los modelos de K-medias, debemos elegir el número óptimo de conglomerados admitidos en "codo". Permite seleccionar el número óptimo de conglomerados ajustando el modelo con un rango de valores para K. Si el gráfico de líneas se parece a un brazo, entonces el "codo" (el punto de inflexión en la curva) es una buena indicación de que el modelo subyacente se ajusta mejor a ese punto. En el visualizador “codo” se anotará con una línea discontinua, en los gráficos a continuación podemos ver este proceso

Twitter - Alberto Linero Timeline



Instagram - Juanita Kremer Followers



Ajuste de modelos de K-Means: con el número óptimo de conglomerados, podemos entrenar los modelos. En este caso, trabajamos con la biblioteca Scikit-Learn y ajustamos los modelos con el método "k-means++” para la inicialización que selecciona los centros de clúster iniciales para el agrupamiento de k-mean de una manera inteligente para acelerar la convergencia y ajustamos el Número de veces que el El algoritmo k-means se ejecutará con diferentes semillas de centroide en 100. Este último método garantiza suficientes iteraciones para los modelos.

Ajuste de modelos de regresión logística: al igual que los modelos de K-means, en este caso entrenamos modelos de regresión logística a través de la biblioteca

Scikit-Learn y ajustamos los modelos con el solucionador de métodos que asignamos como "liblineal" considerando que los conjuntos de datos son lo suficientemente grandes, además asignamos el método de penalización a L1 como método de regularización. Teniendo en cuenta que la regresión logística necesita un objetivo binario. Dividimos el conjunto de datos en dos grupos, el primero con más de 400 retweets y más de 2000 me gusta para Alberto Linero Timeline. Este enfoque se utilizó para entrenar los otros conjuntos de datos.

Evaluación del modelo: para evaluar el rendimiento de los modelos de regresión logística, calculamos la puntuación y la precisión de F1, debajo de los resultados de los modelos ajustados:

Twitter Alberto Linero - Timeline

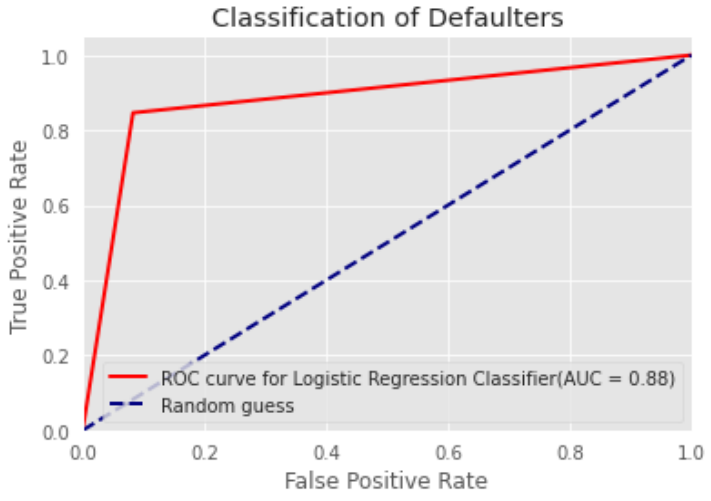
```
F1 score:  0.5945945945945945
Accuracy:  91.17647058823529
```

Instagram Juanita Kremer – Followers

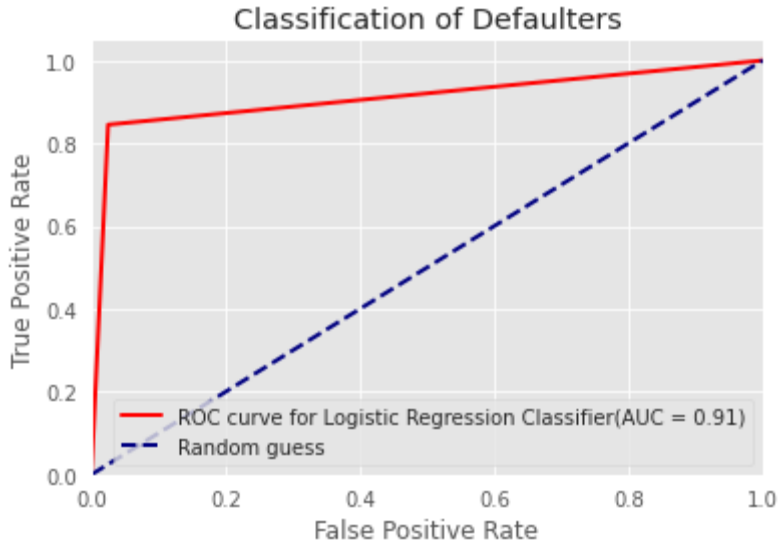
```
F1 score:  0.8697478991596639
Accuracy:  94.86754966887418
```

Además del cálculo de precisión, generamos la curva ROC para estimar el área bajo la curva - AUC y determinar qué tan bueno es el modelo para la clasificación en términos de especificidad y sensibilidad. Debajo de las curvas ROC para modelos entrenados

Curva ROC para el modelo de clasificación de Alberto Linero Timeline (Twitter)



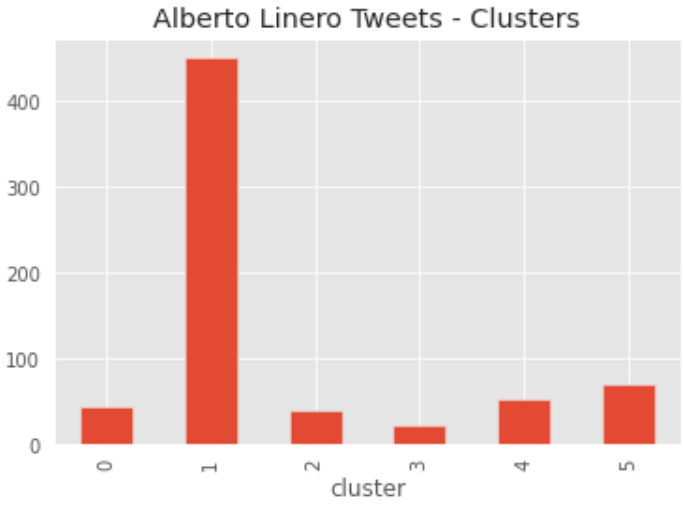
ROC curve for Junita Kremer classification model (Instagram)



Resultados del modelado

Resultados de K-means para la línea de tiempo de Alberto Linero – Twitter

Número de Clusters: 6
Tamaño de los Clusters: hay un segmento que concentra más parte del conjunto de datos. sin embargo, los otros son grupos interesantes que pueden ofrecer información valiosa

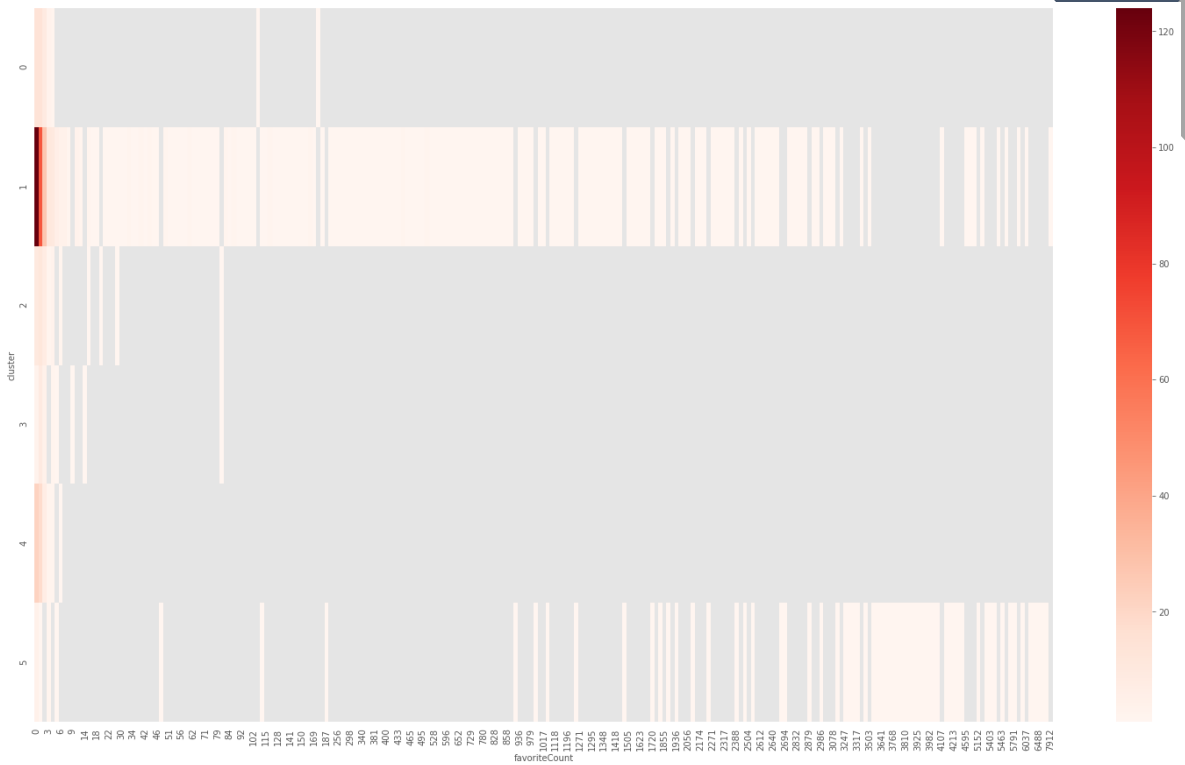


Principales palabras por grupo: cuando convertimos la matriz vectorizada en cadenas de palabras por grupo, podemos encontrar algunas cosas interesantes, podemos ver que hay algunos temas diferentes de comunicación con la audiencia,

por ejemplo, el grupo No. 5 tiene palabras relacionadas con orar. El texto a continuación representa las 50 palabras principales por cada grupo

```
0 : bendiciones, toda, gracias, oro, nixon, catano, esm, casi, red, amn, hermano, dale, pabon, comentar, seora, bot
ero, recuerdo, alegr, saludarlos, oros, oramos, saludos, total, bro, l, enjuiciar, encontramos, encontrar, encontra
ras, enemigo, enfermedad, enfrentamos, engaar, engaarse, estrs, engaas, espacio, encontr, ense, ensear, enseas, ent
ender, entenderlo, estn, entendernos, entenderte, entendido, encontrado, encarnada
1 : lee, gracias, tarea, bien, m, respeto, mejor, linero, bro, acuerdo, est, deseo, amor, dice, uff, siempre, ah, q
u, hace, ser, total, t, usted, columna, amable, a, nunca, creo, felicitaciones, hoy, s, duelen, mvm, mal, gusta, tr
iste, nadie, amaresganarlotodo, da, violencia, slo, favor, cree, tan, persona, katyusha, berna, vida, verdad
2 : bueno, bendiciones, saludos, bro, alegra, nimo, lvaro, bienvenida, aprue, bienvenido, jajajajajaja, felicitacio
nes, ah, evangelista, ehhh, leerte, gracias, jajajajajajaja, herrera, entendido, entiende, empatar, entienden, enti
endes, emociones, emocionante, entiendo, emocionan, entonces, encontrado, entrada, entrar, entreg, encanta, entrega
r, entregarnos, entrego, entrevista, entristece, emocional, entendiste, entenderte, encontramos, empticamente, enca
nt, encontrar, encontraras, enemigo, enfermedad
3 : jajajaja, saludos, total, bendiciones, nico, primo, tal, bueno, sta, bro, bello, dios, invento, retir, recuerdo
, joy, suero, federer, mo, compa, presente, rangel, ah, malo, menos, hablo, claro, parece, equipo, pecado, irle, gu
illermo, s, bara, ningn, apoyo, jhonja, tipo, jajajajaja, hermano, creo, dice, usted, violencia, nunca, respeto, em
ociones, espritu, engaar
4 : saludos, jajajajaja, gracias, respeto, opinin, m, creo, entiendo, a, humm, hago, hablo, bien, bro, agradezco, t
rata, comentario, hombeeeeee, afecta, encontrado, halvar, cayeye, imponer, defensa, puedo, comentarme, muchas, habl
ar, comparto, futebol, salud, bueno, comprobado, t, quiero, mopri, hacer, sorprendes, satisfaga, correccin, ortogra
fa, noviolencia, jodidos, plata, tolis, bicarbonato, ojo, lamento, jajaja
5 : dios, seor, fuerza, corazn, bendiga, gracias, cada, da, danos, seguir, paz, podamos, noche, ay, permtenos, amor
, adelante, l, vivir, ser, vida, claridad, salir, aydanos, nuevo, descansar, maana, fuerzas, padre, permites, nimo,
necesidades, despertar, toda, semana, permita, permitas, frente, duele, necesitamos, ver, poner, mejores, momento,
t, refugio, hoy, oportunidades, salmo
```

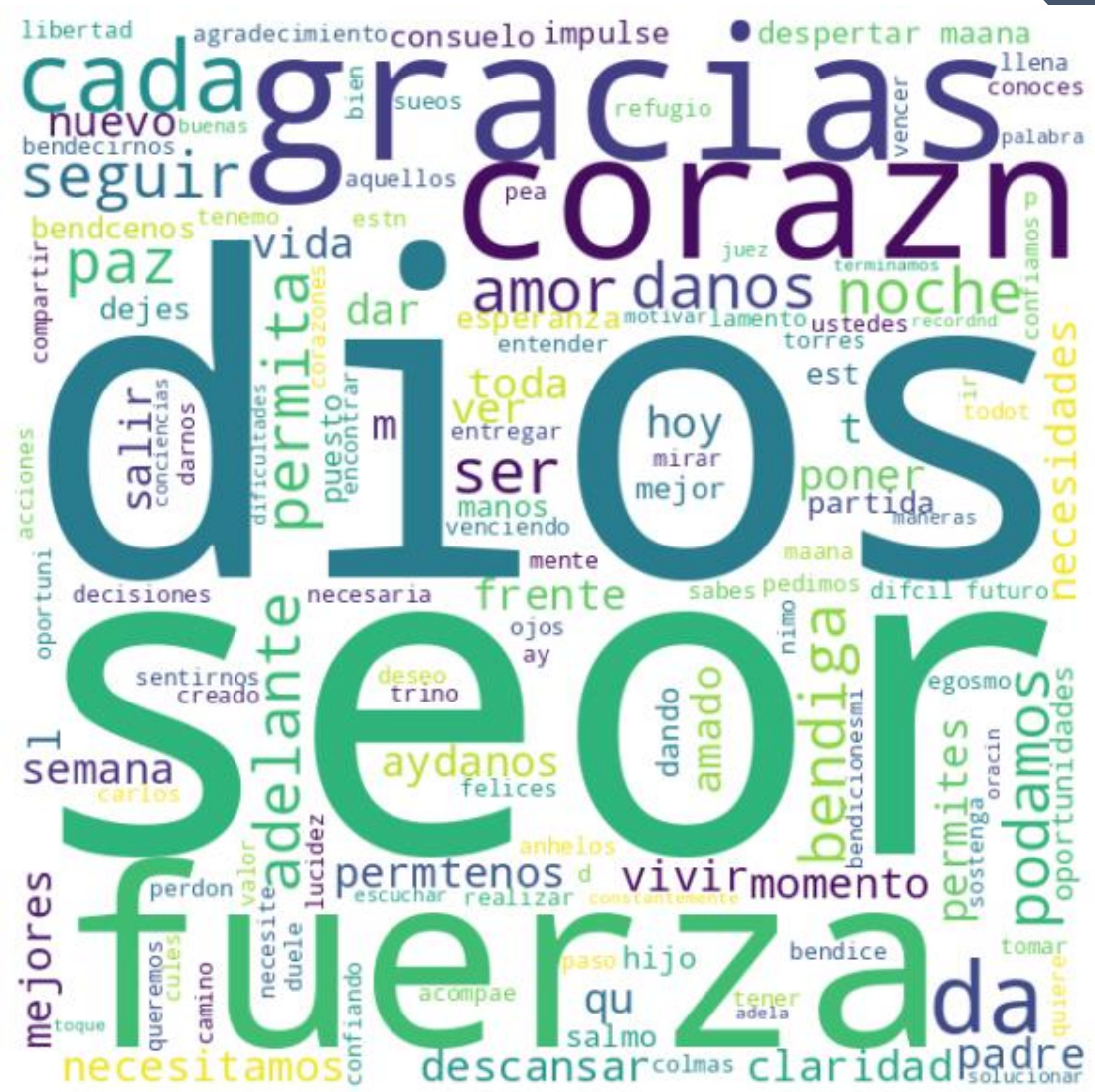
Distribución de likes por cluster: Para entender las diferencias entre cada cluster en términos de negocio cómo likes o favoritos generamos un mapa de calor para identificar la distribución de cada cluster. Como podemos ver, el grupo No. 5 vuelve a ser interesante debido a su gran inclinación a ser el de mayor rendimiento en términos de me gusta por tweet. Es muy diferente cuando se compara con los otros segmentos. A pesar de que el grupo No 1 (el más frecuente) tiene tweets con un alto rendimiento de me gusta, tiene más tweets antes de los 10 me gusta.



Nubes de palabras para grupos más relevantes

Grupo No 5: Oraciones a Dios

La mayoría de las palabras en este grupo están relacionadas con una actividad de oración, como podemos ver en el mapa de calor anterior, es el que más se desempeña en términos de me gusta, llamamos a este grupo Ora a Dios debido a algunas palabras como "danos, permitenos, podemos , padre" por lo general puede estar relacionado con un acto de oración.



Clúster No 4: Interacción con la audiencia

Este cluster es uno de los de menor desempeño en términos de likes, podemos encontrar algunas palabras como “Saludos, bro, gracias” donde el talento tiene una interacción con su audiencia, sin embargo, el resultado no es tan viral como el segmento anterior. Esto crea un compromiso directo con las personas a las que el talento responde o menciona, pero no es masivo.



Resultados de la regresión logística para la línea de tiempo de Alberto Linero – Twitter

Una vez que se ajustó la regresión lineal, extrajimos las palabras más y menos importantes en orden de relevancia para predecir un buen o mal desempeño en términos de retweets,

- Más importante: cómo podemos ver en la siguiente tabla, las palabras relacionadas con una oración como “Dios nuestro, señor e Isaías son las que más se desempeñan para estimar un tweet con alto nivel de retweets. Además, algunas palabras como “tarea” generan una alta respuesta de la

audiencia, esto significa que los tweets donde el talento pública un Llamado a la acción para la audiencia generará un buen desempeño.

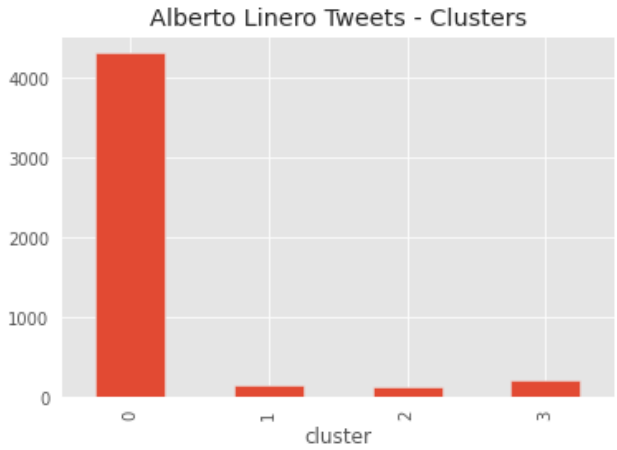
	words	weights
4023	dios nuestro	34.493047
14617	seor	33.928281
15475	tarea	25.889713
7655	isaas	24.001168
6158	ey	23.750266
9167	luchas	18.293416
3978	dios	18.160337
1560	cada	17.837121
7056	hay	16.810454
11095	nuestro	15.808094

Menos importante: Por otro lado, los tuits donde el talento usa las palabras más comunes como “Bendiciones, Dios te bendiga, saludos” suelen tener un bajo nivel de respuesta por parte de su audiencia. Este es un hallazgo muy interesante porque el talento puede generar algunas estrategias con el conocimiento de las palabras con mejor y peor desempeño.

	words	weights
1261	bendiciones	-2.498490
4065	dios te bendiga	-2.598034
9429	me	-2.897761
12310	por	-3.039721
15576	te bendiga	-3.605739
8727	lo	-3.873297
7103	hay que	-4.467825
112	a los	-4.562955
1133	ay dios	-6.250280
14258	saludos	-6.804467

Resultados K-means para seguidores de Juanita Kremer – Instagram

- Número de grupos: 4
- Tamaño de los clústeres: hay un segmento que concentra más parte del conjunto de datos. sin embargo, los otros son grupos interesantes que pueden ofrecer información valiosa



- **Palabras principales por cluster:** cuando convertimos la matriz vectorizada en palabras de cadena por grupo, podemos encontrar algunas cosas interesantes, podemos ver que hay un perfil de seguidores muy diferente. Por ejemplo, el grupo número 2 se refiere a profesiones técnicas o de ingeniería como ingenieros industriales, ingenieros civiles, desarrolladores electricistas, mientras que el segmento número 3 está enfocado a ser más apasionados y disfrutadores de la vida. A pesar de que pueden representar una pequeña porción. Representan nichos de segmento muy interesantes

```
0 : gusta, colombia, social, ser, periodista, abogado, amante, msica, m, antiuribista, and, familia, estudiante, co
municador, mundo, siempre, persona, ftbol, ing, universidad, derecho, i, amor, ver, solo, opiniones, colombiano, am
o, ingeniera, the, administrador, profesional, mejor, rock, empresas, trabajo, alegre, especialista, vivir, bogot,
in, futbol, digital, pa, poltica, personales, love, buen, naturaleza
1 : dios, familia, gracias, amor, hijos, vida, rodillas, doy, gloria, da, creyente, siempre, mano, primero, mundo,
amo, ma, bendiga, hijo, patria, apasionado, hizo, honra, confo, pie, puede, mejor, posible, temro, mujer, social, u
niversidad, cambio, conmigo, nico, temeroso, madre, esposo, perfecto, solo, protejenos, m, jess, uribista, medicina
, grasias, seguidor, agradecida, mio
2 : ingeniero, civil, electrnico, industrial, sistemas, ambiental, mecnico, telecomunicaciones, gerencia, proyectos
, gestin, electricista, universidad, amante, ftbol, electronico, entusiasta, m, vias, alimentos, mecanico, quimico,
familia, emprendedor, especialista, soador, apasionado, colombia, hincha, gran, industriallasallistaempresario, bar
ranquilla, millos, innovacin, tecnologa, productor, transformacin, msica, distrital, ejecutante, acorden, profesin,
logstica, agrnomo, antiuribista, dgcd, digital, magister, siempre
3 : vida, feliz, vive, solo, amor, ser, apasionado, mejor, familia, disfrutar, m, amante, vivir, enamorada, mam, da
, amo, cosas, disfruta, bella, a, encanta, sonre, madre, escritor, momentos, eterna, sonreir, arrepentimientos, ama
r, creo, colombiano, felicidad, regalo, sola, persona, aprender, disfrutando, diego, ma, biologo, difratar, saber,
gran, amorpor, msica, tiempo, viajar, rock
```

- **Nubes de palabras para grupos más relevantes**

Clúster No 2: Profesionales técnicos

Este grupo está compuesto por personas que están enfocadas en sus carreras y expresan su pasión por eso. Podemos encontrar profesiones como ingeniero eléctrico, ingeniero civil, gerentes, gerentes de proyectos, personas que trabajan en el sector de TI, etc.



Clúster No 1: Seguidores familiares

Los seguidores de este segmento se pueden definir como centrados en la familia. Son personas que anteponen su relación familiar y mencionan a Dios y la religión como núcleo de su vida. Por lo general, las personas de este segmento tienen inclinación al partido político colombiano debido a palabras como “uribista”.



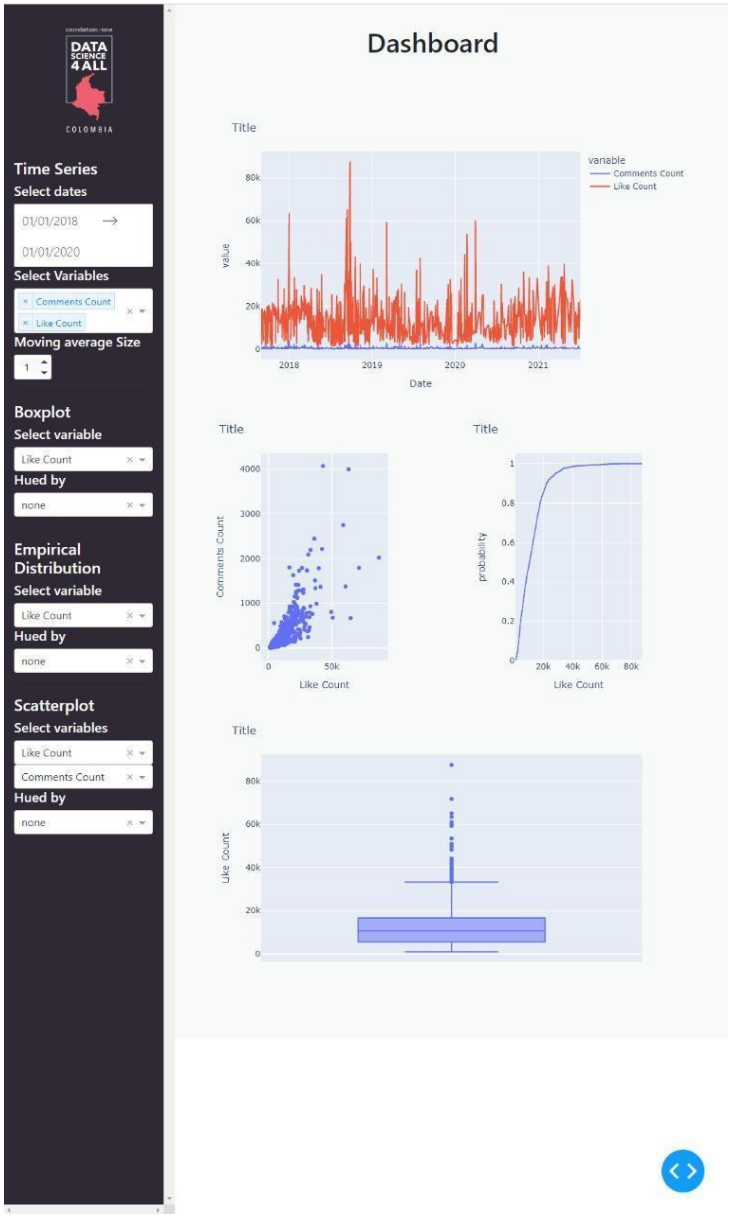
Clúster No 3: Seguidores apasionados

Los seguidores de este segmento son disfrutadores de la vida, expresan felicidad con su vida, son personas que dan y reciben amor de los demás. Expresar pasión por estas actividades y disfrutar al máximo cada momento. Amantes de la música y las experiencias


ser
amante
mi
amor
vida
feliz
mejor
solo
enamorada
madre
eterna
justicia
creo
tiempo
derecho
amo
apasionado
da
hacer
siento
leer
sola
mas
toda
diego
mam
rer
poder
sonre
colombiano
persona
universidad
disfrutar
juntos
encanta
buena
rock
estudiante
cosa
posible
vive
libertad
mara
momentos
proceso
nacional
deporte
gran
satisfaccion
compaa
dar
disfruta
aprender
familia
vivir
msica
saber
tecnologa
humanos
as
peor
monopolios
pasiones
real
escritor
antiuribista
lectura
travs
animal
muerte
arrepentimientos
hermoso
pueblo
parte
rodriguez
hermosa
dara
rolo
pierde
oportunidad
intentar
reciclaje
logrars
aqui
calienamorado
defender
diseo
alegre
padre
vivo
sof
entienden
mercadlogo
gastemos
orgulloso
contadora
miedo
camino
ahora
mala
abejas
tomar
huella
twitterme
apieloco
cielos
profundo
cambiar
dejan
opinin
universo
sonrer
vvela
gastronoma
mismo
juan
apasionada
enamorado
regalo
dedicado
ciego
musico
colombia
valor
preocupada
resistencia
dado
naturaleza
pertenecer
sino
puede
patriaami
creador
viajaras
instituciones
dolos
nave
sent
siempre
gastarse
fabricacin
moneda
bella
cada
dream
sonre
mujer
amorpor
loco
respeto
llegar
respiramos
buen
velocidad
nctar
competencia
tierra
love
mide
invito
maravillosa
disfrutando
maquinaria
lmite
sonrer
opinin
dejan
profundo
cielos
universo
vvela
tomar
huella
twitterme
apieloco
musico
colombia
valor
preocupada
resistencia
dado
naturaleza
pertenecer
sino
puede
patriaami
creador
viajaras
instituciones
dolos
nave
sent
siempre
gastarse
fabricacin
moneda
bella
cada
dream
sonre
mujer
amorpor
loco
respeto
llegar
respiramos
buen
velocidad
nctar
competencia
tierra
love
mide
invito
maravillosa
disfrutando
maquinaria
universidad
persona
colombiano
amo
derecho
mejor
creo
tiempo
dedicado
gastronoma
mismo
juan
apasionada
enamorado
regalo
contadora
miedo
camino
ahora
mala
abejas
tomar
huella
twitterme
apieloco
cielos
profundo
cambiar
dejan
opinin
universo
sonrer
vvela
justicia
madre
eterna
ser
amante
mi
amor
vida
feliz
mejor
solo
enamorada
madre
eterna
justicia
creo
tiempo
derecho
amo
apasionado
da
hacer
siento
leer
sola
mas
toda
diego
mam
rer
poder
sonre
colombiano
persona
universidad
disfrutar
juntos
encanta
buena
rock
estudiante
cosa
posible
vive
libertad
mara
momentos
proceso
nacional
deporte
gran
satisfaccion
compaa
dar
disfruta
aprender
familia
vivir
msica
saber
tecnologa
humanos
as
peor
monopolios
pasiones
real
escritor
antiuribista
lectura
travs
animal
muerte
arrepentimientos
hermoso
pueblo
parte
rodriguez
hermosa
dara
rolo
pierde
oportunidad
intentar
reciclaje
logrars
aqui
calienamorado
defender
diseo
alegre
padre
vivo
sof
entienden
mercadlogo
gastemos
orgulloso

6. Desarrollo de la aplicación

Tenemos un tablero inicial que puede mostrar diferentes gráficos: series de tiempo, diagrama de caja, distribución empírica y diagrama de dispersión. El gráfico de serie temporal tiene una opción de ventana móvil para seleccionar el intervalo de tiempo según las necesidades y la opción para seleccionar variables para mostrar en el gráfico. En el diagrama de caja, podemos elegir la variable y el tono para mostrar las cajas, es lo mismo para la distribución empírica y el diagrama de scartterplot (este último caso para dos variables).




DATA SCIENCE 4 ALL
COLOMBIA
Time Series
Select dates
01/01/2018
01/01/2020
Select Variables
Comments Count
Like Count
Moving average Size
1
Boxplot
Select variable
Like Count
Hued by
Month
Empirical Distribution
Select variable
Engagement
Hued by
none
Scatterplot
Select variables
Impressions
Reach
Hued by
Media Type
Dashboard
Title
variable
Comments Count
Like Count
value
Date
Titulo
Media Type
CAROUSEL_ALBUM
VIDEO
IMAGE
Reach
Impressions
Title
probability
Engagement
Title
Like Count
Month

7. Referencias Bibliográficas

- https://www.sciencedirect.com/topics/computer-science/logistic-regression
- https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1
- https://www.scikit-yb.org/en/latest/api/cluster/elbow.html
- https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
- https://www.maracuyaagencia.com/