**ANÁLISIS MULTIVARIADO**

**TRABAJO PRÁCTICO 2**

**Tema: Descriptivo Multivariado**

**PROFESORA:**

**MARÍA INÉS STIMOLO**

**INTEGRANTES:**

**AGUSTIN GABRIEL PRETINI**

**SIGAL DANIELA ESQUENAZI**

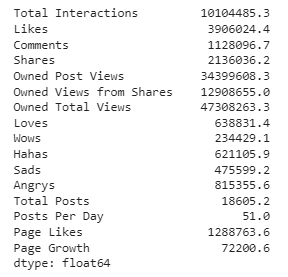
**ANÁLISIS DESCRIPTIVO**

Análisis sobre la estructura y relaciones de nuestras variables

Del análisis descriptivo obtuvimos una primera vista de las características centrales de nuestros datos.

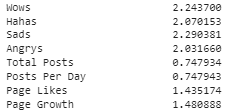
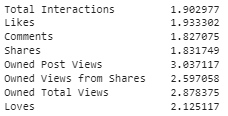
Procedimos a obtener las principales medidas descriptivas multivariadas, las cuales son:

* Vector de medias:



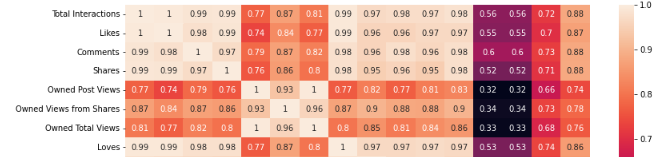
Aquí se puede observar el valor medio que asume cada una de las variables. Se observa que tenemos en general escalas de medida altas (en millones), con algunas excepciones siendo la más notoria ‘Posts Per Day’.

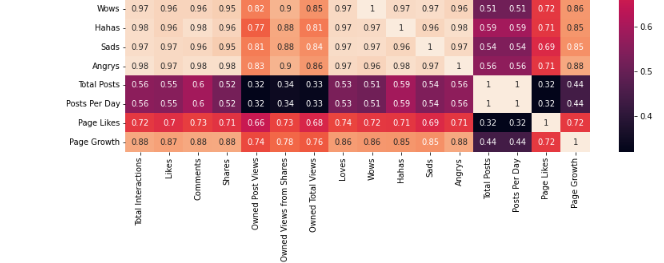
* Coeficientes de variación (son univariados pero nos pareció oportuno incluirlos)



Como parte de un primer acercamiento a nuestros datos, miramos los coeficientes de variación de cada variable, para observar la variabilidad relativa que presentan nuestras variables. Se observa que en la mayoría de los casos el mismo está entre 1,5 y 2,5 unidades; a excepción de ‘Owned Post views’ (3,037), que es la variable que tiene mayor variabilidad relativa, y ‘Total Posts’ y ‘Posts Per Day’ (0,747), que son las variables con menor variabilidad relativa.

* Matriz de correlación





Basaremos nuestro análisis de relaciones entre variables en la matriz de correlación, ya que esta nos permite ver el sentido y la intensidad de las asociaciones entre las variables (mientras que la matriz de varianzas y covarianzas únicamente permite ver el sentido) y además es más intuitiva su interpretación.

De la misma podemos concluir :

Todas las correlaciones son positivas, lo que indica que todas las variables de nuestra base tienen una relación directa (cuando una aumenta la otra también lo hace).

Primero que nada y aunque suene muy intuitivo, hay correlaciones en torno a 1 entre ‘Total interactions’ con ‘Likes’, ‘Shares’, ‘Loves’ ‘Angrys’, Sads’, ‘Comments,’Wows’, y ‘Hahas. Esto se debe a que ‘Total interactions’ se construye a partir de las variables mencionadas.

En este contexto nos planteamos la siguiente pregunta:

*¿Más posteos generan más crecimiento del medio periodístico?*

Analizando la correlación entre las variables, podemos ver que la misma es de 0,44; lo que implica que al realizar más posteos, aumenta el crecimiento del medio periodístico pero no tiene una intensidad tan alta. Por lo tanto seguiremos desglosando éste análisis a continuación, para intentar encontrar alguna variable que explique mejor el crecimiento del medio.

Para ellos nos planteamos otra pregunta:

*¿Son las interacciones responsables del crecimiento del medio periodístico?*

Podemos observar que el crecimiento de los medios periodísticos, está altamente correlacionado con la cantidad de interacciones que genera (corr=0,88). Por lo tanto, si un medio quiere crecer, necesita generar interacciones.

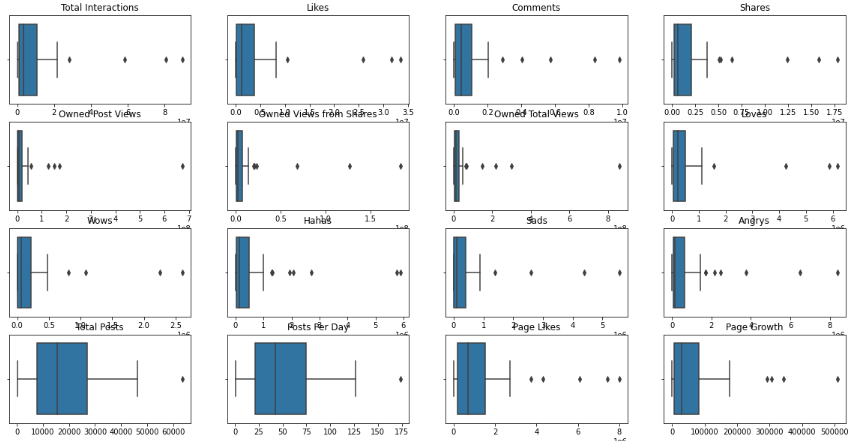
Para seguir profundizando el análisis, pasamos a mirar las correlaciones entre el total de posteos con la cantidad de interacciones; y de la cantidad de vistas de un posteo con el total de interacciones. Esto con el objetivo de entender cómo hace un medio para generar interacciones y por lo tanto para crecer. Del análisis mencionado, se observa lo siguiente:

1. El total de posteos que hacen los medios periodísticos, tiene una correlación de 0,56 contra la cantidad de interacciones de esos posts. De ésta manera, podemos inferir que la cantidad de posteos es importante si un medio quiere generar atención, pero no es lo único. También denota una posible importancia en el contenido de esos posteos.
2. Si observamos la correlación entre cantidad de vistas de un posteo y cantidad de interacciones, vemos que la misma es de 0,81, lo cual estaría indicando que casi todos los usuarios que ven una publicación, luego interactúan con ella.
3. Como conclusión a esta cuestión sobre los posteos, y comparando las dos correlaciones descritas, deducimos lo siguiente:

Casi todos los usuarios que ven un post interactúan con el (corr=0,81), pero no casi todos los posteos tienen interacciones (corr=0,56). Por lo tanto llegamos a la conclusión, de que hay muchos posteos que no están siendo vistos por los usuarios. Recomendamos a las compañias que revisen esta situación, y vean por ejemplo el horario de las publicaciones (podría estar ocurriendo que hay una importante cantidad de posteos que se publiquen en horarios en los que la gente usa poco las redes).

Análisis de la distribución de probabilidad

Para finalizar el análisis descriptivo procedemos a analizar la distribución de probabilidad de nuestra base de datos.

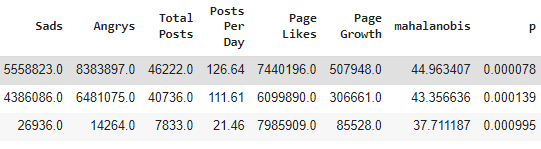
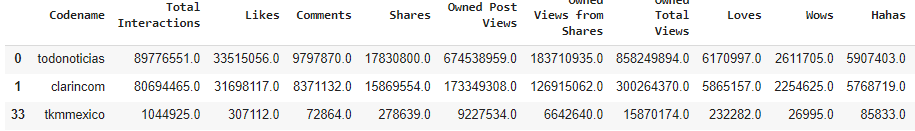
Primero realizamos boxplots univariados, los cuales se ven a continuación:

Se observa de los mismos, que todas las variables presentan asimetría derecha con outliers univariados hacia la derecha.

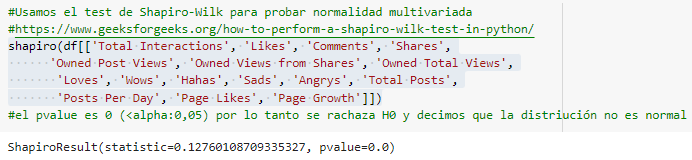
A continuación analizamos si hay o no outliers multivariados. Para ello se procedió al cálculo de la distancia de Mahalanobis para cada observación y al correspondiente p-valor para dichas distancias, teniendo en cuenta que la distancia mencionada se distribuye chi2.

Clasificaremos como outliers a aquellas observaciones con un p<0.001.

Así llegamos a la conclusión de que tenemos 3 outliers multivariados:



Por último aplicamos el test de Shapiro-Wilks para determinar si tenemos una distribución normal multivariante. El resultado fue negativo, ya que se observa un p-value de 0 (menor a alfa de 0,05) para la prueba.



**ANÁLISIS DE COMPONENTES PRINCIPALES**

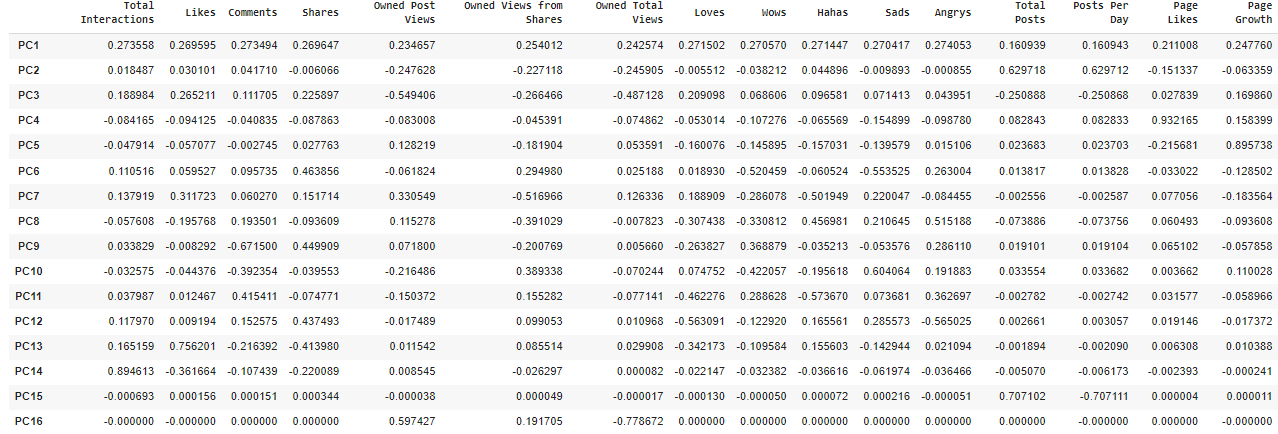
El análisis de componentes principales (PCA) permite transformar el conjunto de variables en un nuevo conjunto de variables ordenadas según su importancia en relación a la información que contienen del conjunto de datos original, incorreladas entre sí (corr=0).

Éstas nuevas variables que mencionamos se denominan componentes principales, y son combinaciones lineales de las variables originales, las cuales se proyectan hacia la dirección en la cual los datos presentan mayor variabilidad. Ésto se debe a que cada componente intenta maximizar la varianza que explica.

Es importante destacar que el máximo número de componentes principales que se pueden obtener es igual al número de variables que tenemos en la base de datos. También mencionar que cada componente tiene asociado un autovector (vector de coeficientes por los cuales se multiplican a las variables originales para realizar la combinación lineal) y un autovalor (varianza de la componente principal). Por último decir que la suma de los autovalores es igual a la varianza total de la base de datos original.

Como paso previo a aplicar el PCA, normalizamos nuestros datos para eliminar el efecto de las diferencias de escala en las variables. Para ello utilizamos Standard Scaler (restar a cada observación su media y dividir por su desviación estándar).

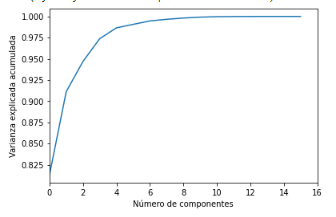
Como en la base de datos teníamos 16 variables numéricas, obtuvimos 16 componentes principales, de las cuales a continuación se observan sus autovectores:

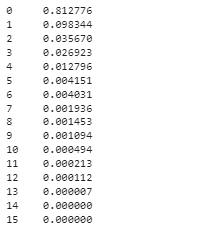


Como se puede observar, en la primera componente (PC1) las variables que representan interacciones (‘Total Interactions’, ‘Likes’, ‘Comments’, ‘Shares’, ‘Loves’, ‘Wows’, ‘Hahas’, ‘Sads’, ‘Angrys’) son aquellas que mayor peso tienen en la misma, con un coeficiente en el autovector que tiene leves variaciones en torno a 0,27, para las mencionadas variables.

Esto era de esperarse, ya que todas estas variables originales tienen correlación casi perfecta.

En la segunda componente (PC2), las variables con mayor influencia son aquellas referidas a la cantidad de posteos (*‘Total posts’, “post per day”),* con coeficientes en el autovector en torno a 0,63.

Una cuestión central en el análisis de componentes principales, son los autovalores. Pasaremos a analizarlos:



La tabla de la izquierda muestra la tasa de variabilidad que explica cada componente (autovalor de esa PC/varianza total), mientras que el gráfico de la derecha muestra la tasa de variabilidad explicada acumulada al sumar una nueva componente.

Las componentes principales siempre se ordenan en función de la variabilidad total que explican. En éste caso la primera, recoge el 81,27% de variabilidad, es decir casi la totalidad de la varianza. Ésto se debe a que, como ya se mencionó, nuestra base de datos tiene muchas variables altamente correlacionadas, que son aquellas que más pesan en la componente 1, y son las que se refieren a los diferentes tipos de interacciones de las publicaciones.

Uno de los objetivos de PCA es reducir la dimensionalidad de los datos, para poder realizar un análisis de una manera más sencilla, sin perder mucha información.

En este caso al seleccionar solo las dos primeras componentes, la variabilidad explicada alcanza una tasa acumulada de 0,91.De esta forma podríamos trabajar en dos dimensiones (lo cual es sin dudas más cómodo que trabajar con 16) , perdiendo solo el 9% de la información.

Otro criterio para seleccionar con cuántas componentes trabajaremos, es el presentado en el segundo gráfico a la derecha. Este método sugiere seleccionar el número de componentes en el cual la pendiente empieza a aplanarse, ya que al agregar una nueva componente al análisis es poca la varianza explicada que se incorpora.

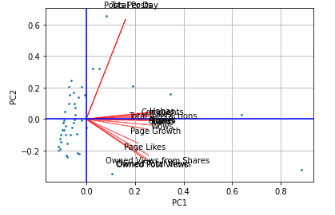
Según éste criterio deberíamos trabajar con 4 componentes, pero considerando que con 2 ya se explica el 91% de la varianza, nos inclinamos por éste criterio.

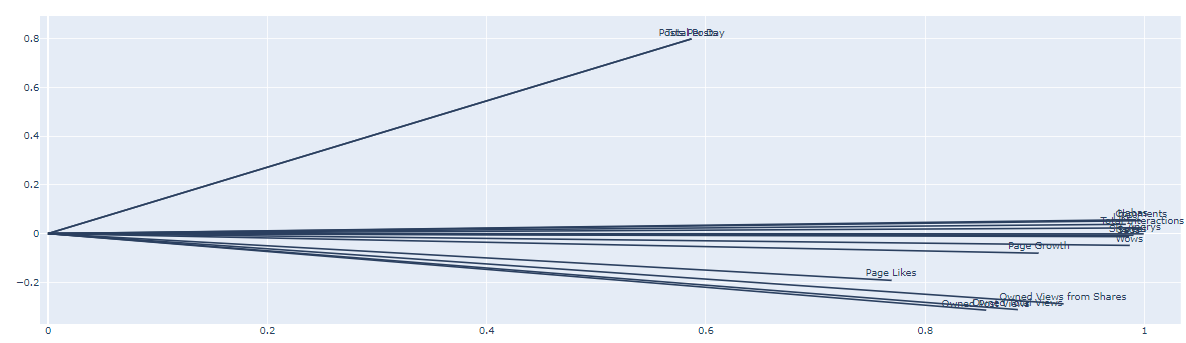
Biplot

El biplot es un gráfico en el cual los ejes son las 2 primeras componentes.

En el mismo se pueden ver las observaciones proyectadas sobre las 2 componentes mencionadas, lo que se denomina puntuación en las componentes principales. Éstas puntuaciones se obtienen reemplazando los valores de las variables originales en la ecuación (formada con el autovector asociado) de la respectiva componente.

Además en el gráfico se observan las variables originales proyectadas en el espacio de las componentes, en forma de vectores.





Se puede ver en el biplot que la mayoría de las observaciones presentan valores negativos en la componente 1 (expresados en desviaciones estándar en torno a la media), mientras que asumen de manera prácticamente uniforme valores positivos y negativos en la componente 2. También se observan algunas pocas observaciones con valores positivos en la componente 1, entre las cuales se encuentran los 3 outliers multivariados obtenidos con la distancia de Mahalanobis.

Para concluir, decimos que no se observan grupos diferenciados entre las observaciones, si no que casi todas presentan características similares (ya que en la componente 1, que es la que más variabilidad explica, la mayoría de las observaciones tiene valores similares).

**CONCLUSIÓN**

Para concluir nuestro análisis haremos un breve repaso de los insights obtenidos:

1. No se observa una marcada división de grupos entre las publicaciones, presentando la mayoría de ellas características similares, salvo 3 outliers (todonoticias, clarincom, tkmmexico).
2. La cantidad de interacciones de las publicaciones está altamente asociada con el crecimiento del medio periodístico.
3. Sugerimos a los medios periodísticos revisar el horario de las publicaciones, haciendo énfasis en que se realicen en horarios en los que haya muchos usuarios activos. Ya que esto determinará el alcance de las mismas, lo que muy probablemente aumente las interacciones y genere crecimiento a la página periodística.

Colab:

[https://colab.research.google.com/drive/1RDgBcORzQPAscI89aIV2eWMAYB3f9QEw?usp=sharing](https://colab.research.google.com/drive/1RDgBcORzQPAscI89aIV2eWMAYB3f9QEw?usp=sharing&authuser=1)