Explorando sitios de comercio electrónico

Análisis Inteligente de Datos

Leonardo Aranda

Maestría en Explotación de Datos
y Descubrimiento del Conocimiento
Universidad de Buenos Aires

Índice

Introducción

El comercio electrónico viene creciendo sostenidamente desde hace varios años impulsado por factores como son el aumento del alcance de Internet, la disminución de costos de computadoras y la adopción de tecnología por parte de los usuarios finales, entre otros.

Muchos de los sitios de comercio electrónico que operan en la actualidad ejecutan diferentes estrategias de marketing digital para hacer crecer el volumen del negocio. Además de ello, desde el punto de vista del usuario final, algunos sitios generan mayor tracción que otros, que a largo plazo se transforman en visitas repetidas y un mayor tiempo de permanencia por cada visita.

En el presente trabajo se analizarán un conjunto de sitios de comercio electrónico para intentar descubrir características que presentan aquellas empresas que están mejor posicionadas a nivel mundial e identificar patrones basados en el comportamiento de los usuarios y acciones de marketing digital.

Conjunto de datos

Alexa Internet es una empresa perteneciente a Amazon.com que ofrece varios servicios para poder realizar investigación de mercado. Es un referente desde el punto de vista de las estimaciones de tráfico a sitios webs, que las realiza obteniendo datos a partir de instalaciones de diferentes toolbars que se agregan en navegadores web. Además de ello, SimilarWeb, es otra compañía muy reconocida en la industria que ofrece soluciones basadas en datos para entender el comportamiento del mercado en Internet.

Los datos del presente análisis se basan en el ranking de Alexa.com junto con información obtenida desde SimilarWeb.com. Algunos ejemplos de datos:

- https://www.similarweb.com/website/google.com.ar
- http://www.alexa.com/siteinfo/google.com.ar

Variables del análisis

Variable	Descripción
url	Dirección del sitio web
position	1 = Top 50.000 mundial, 0 = No es Top 50.000
bounce	Porcentaje de visitas con una sola página vista
ppv	Páginas vistas por visita
time	Tiempo promedio de la visita
paid_search	Tráfico pago
traffic_direct	Tráfico directo
traffic_display	Tráfico de redes de publicidad
traffic_search	Tráfico de buscadores
traffic_mail	Tráfico de correo electrónico
traffic_social	Tráfico de redes sociales
traffic_referrals	Tráfico referido desde otras páginas

Las variables fueron obtenidas directamente desde la fuente de datos. La variable *position* está construída utilizando el ranking a nivel mundial de Alexa. Algunas estadísticas de resumen:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
bounce	6.09	32.56	40.52	41.010	48.860	89.85
paid_search	0.00	0.00	9.11	14.870	24.900	84.28
ppv	1.12	3.08	4.34	4.786	5.880	19.17
time	8.00	118.00	177.00	190.300	249.000	480.00
traffic_direct	0.00	16.98	23.19	25.320	31.440	88.08
traffic_display	0.00	0.00	0.08	1.057	1.030	29.71
traffic_search	0.66	40.12	54.05	52.680	66.230	96.39
traffic_mail	0.00	0.47	2.57	4.420	5.945	42.99
traffic_social	0.00	0.54	1.50	3.268	3.722	75.95
traffic_referrals	0.00	7.98	11.73	13.250	16.030	91.49

El conjunto de datos cuenta con 1928 observaciones, que se distribuyen respecto a la variable *position* de la siguiente manera:

position	cantidad
0	1525
1	403

Análisis

El desarrollo del análisis se va a centrar en describir los sitios en función de la variable *position*, que indica si un sitio pertenece a aquellos con mayor nivel de tráfico.

Diferencias entre grupos según el tiempo de la visita

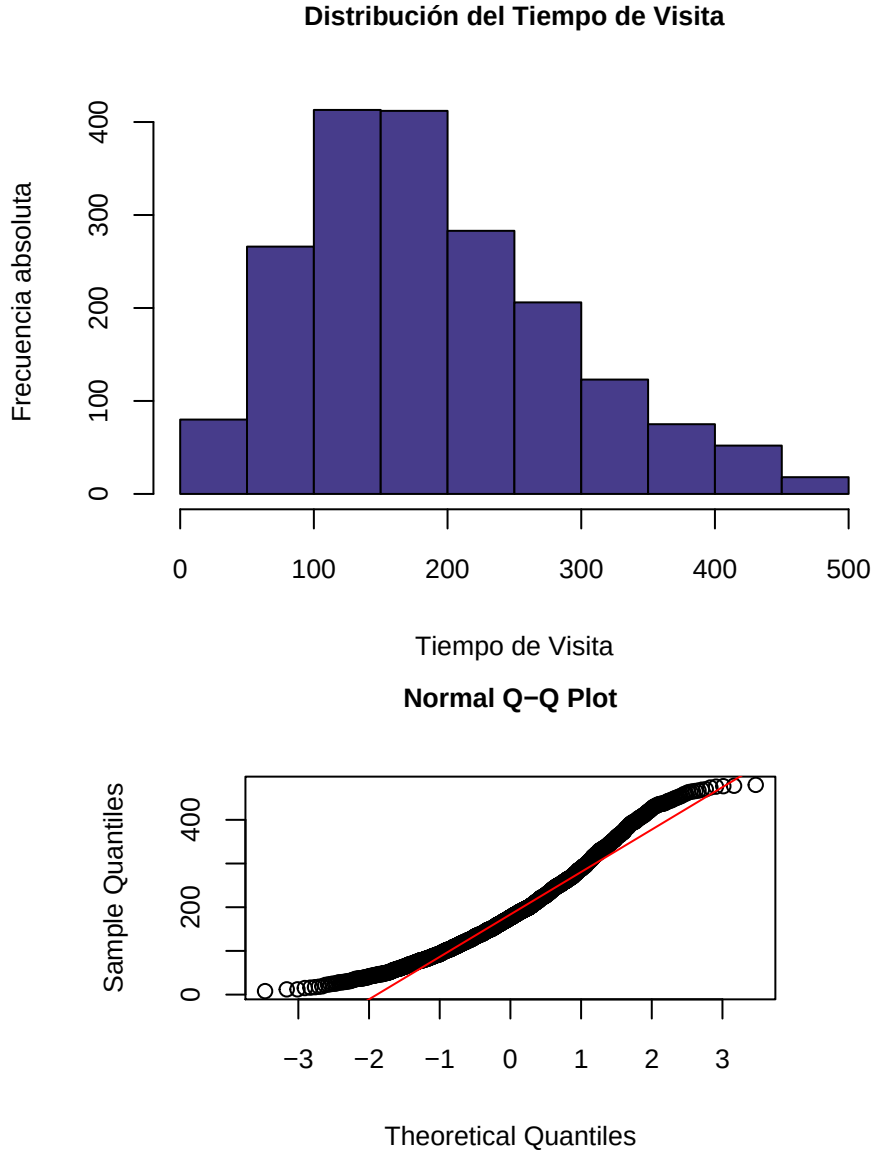
El vector de medias de las variables según la posición global de los sitios nos brinda información sobre cada uno de los grupos. Resulta de interés comprender si existen diferencias significativas entre ellos.

	V1	V2
Group.1	0	1
bounce	40.48016	43.02896
paid_search	13.92329	18.45615
ppv	4.517698	5.803474
time	172.9875	255.8759
traffic_direct	24.45626	28.58323
traffic_display	0.8779869	1.7365757
traffic_search	54.47384	45.90362
traffic_mail	4.214623	5.198238

	V1	V2
traffic_social	3.117430	3.837295
traffic_referrals	12.86008	14.74079

El Tiempo de Visita (time) es un indicador muy importante para medir el nivel de actividad de los usuarios en cada sitio de comercio electrónico. Se espera intuitivamente que los sitios con mayor tráfico en la web tengan un Tiempo de Visita elevado.

Debemos analizar la normalidad de los datos para comprender el tipo de prueba a realizar, es decir si será un test paramétrico o no paramétrico.



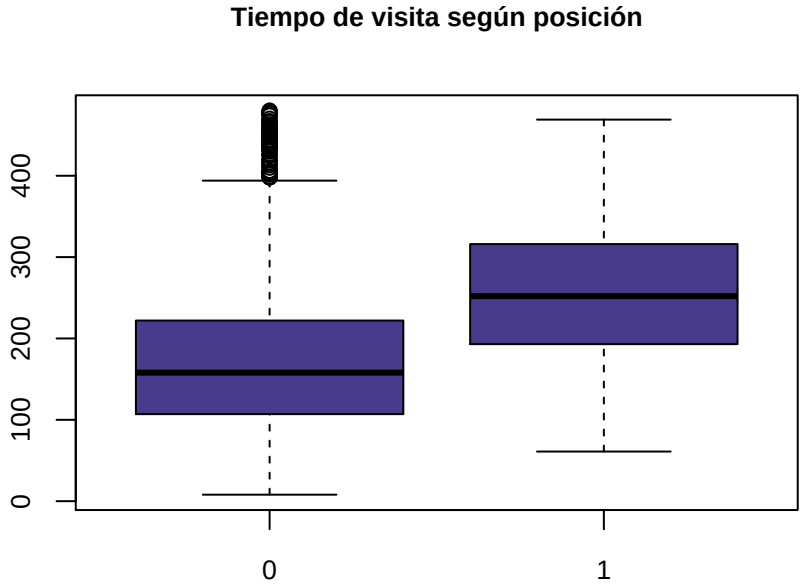
Gráficamente se visualiza que los datos no son normales. Además de ello, podemos verificarlo mediante un test de Shapiro-Wilk.

##

```
##  Shapiro-Wilk normality test
##
## data:  df$time
## W = 0.96344, p-value < 2.2e-16
```

Nuestra hipótesis nula es que los datos siguen una distribución normal. El test rechaza dicha hipótesis, por lo cual los datos no presentan normalidad.

A continuación vamos a verificar la normalidad de los grupos por separado.



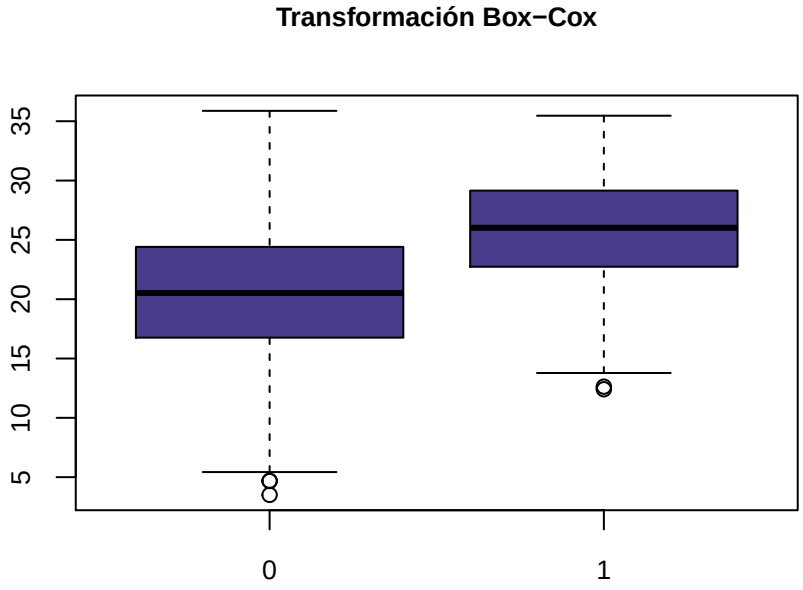
Gráficamente vemos que la población de la variable position con valor 0 presenta una asimetría muy marcada con cola pesada hacia la derecha. Al aplicar el test de Shapiro-Wilk sobre los grupos obtenemos lo siguiente:

position	p-value
0	0.0000000
1	0.0240839

El p-valor obtenido rechaza la hipótesis nula de normalidad.

Transformación Box-Cox

Podemos intentar realizar transformaciones Box-Cox sobre la variable de estudio para corregir la falta de normalidad. Luego de buscar varios parámetros lambda, se obtiene la siguiente mejora:



position	p-value
0	0.0002660
1	0.0019657

Se obtienen algunas diferencias pero se continúa rechazando la hipótesis nula de normalidad, por lo que procederemos a realizar un test no paramétrico para verificar diferencias entre las distribuciones.

Prueba de Mann-Whitney-Wilcoxon

Para proseguir y verificar si hay diferencias significativas entre los grupos respecto a la variable Tiempo de Visita, se realiza la prueba de rangos no paramétrica de Mann-Whitney-Wilcoxon.

Esto nos lleva a plantear las siguientes hipótesis:

- H_0: No hay diferencias significativas. Los grupos pertenecen a la misma distribución.
- H_1: Existen diferencias significativas entre los grupos.

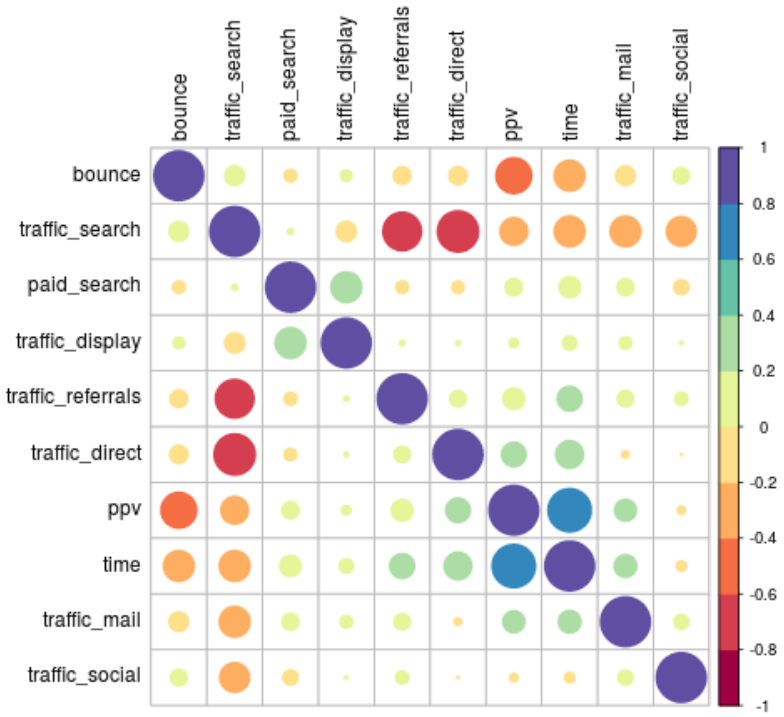
```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  time by position
## W = 148840, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

Se rechaza la hipótesis nula, por lo que dicha variable brinda información sobre diferencias entre sitios con poco y mucho tráfico.

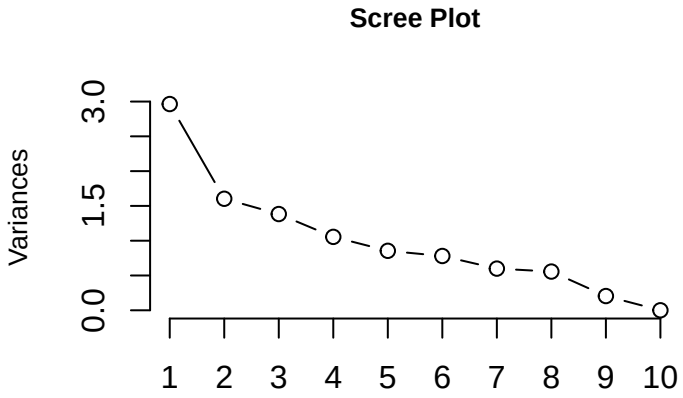
Componentes principales

El conjunto de datos en análisis posee una cantidad considerable de variables. Si quisiéramos visualizarlas todas al mismo tiempo sería prácticamente imposible o muy difícil de interpretar. Con componentes principales podemos hacer una reducción de la dimensionalidad y obtener mucha información sobre las variables y su correlación.

Componentes principales necesita que las variables en estudio estén correlacionadas. Para entender esto vamos a visualizar la siguiente matriz de correlaciones.



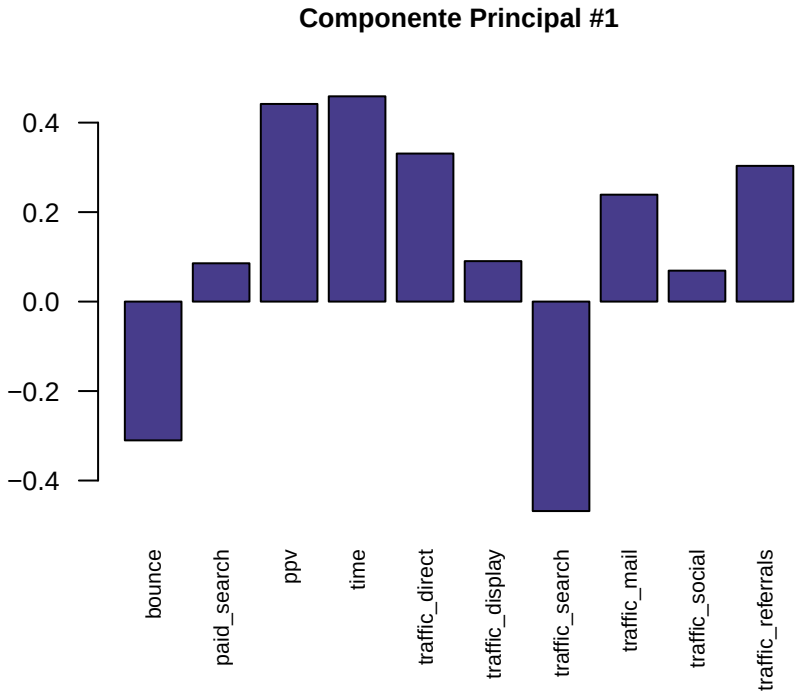
Observando el gráfico se pueden visualizar variables que tienen un nivel de correlación importante entre sí, sin embargo, esto no es extremadamente marcado.



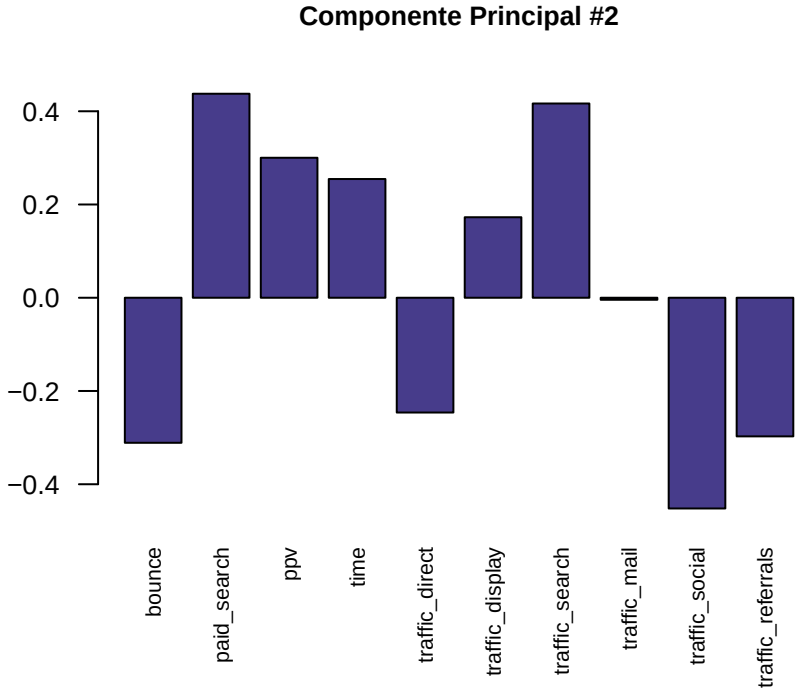
El gráfico nos da la pauta de que si bien las primeras componentes capturan una buena porción de la variabilidad, las variables no están altamente correlacionadas, por lo que para obtener un

porcentaje alto de explicación de los datos nos vemos obligados a seleccionar al menos 4 variables siguiendo el criterio de Kaiser. Para que el análisis se pueda interpretar mejor, continuaremos con 3 componentes, con los cuales explicamos aproximadamente el 60 % de la variabilidad de los datos.

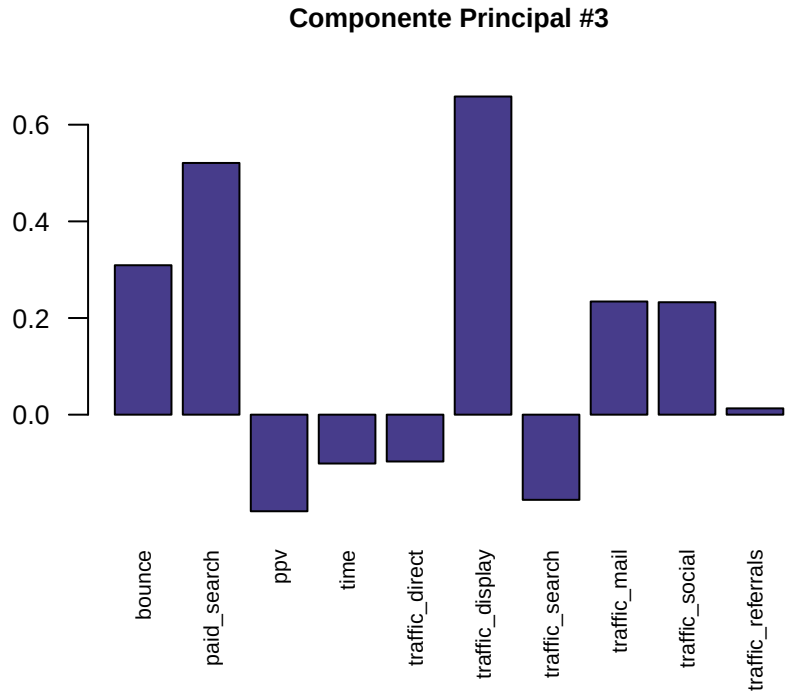
Análisis de las componentes



Los loadings de la componente #1 nos permite interpretar el engagement de los sitios. Es decir, son sitios donde los usuarios tienen alto nivel de actividad.

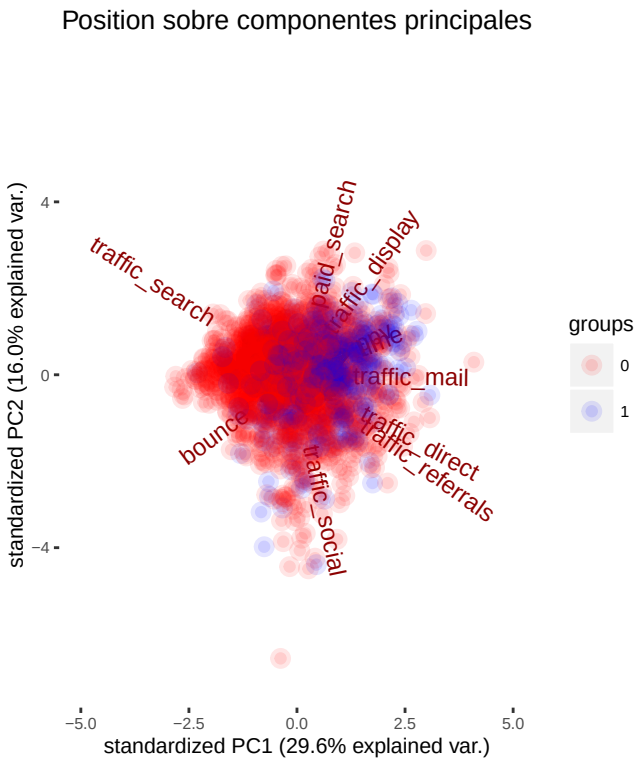


Los loadings de la componente #2 la podemos resumir como Awareness, es decir, demuestra el nivel de conocimiento que se tiene sobre la marca.



En cuanto la componente #3 está relacionada con el impacto que tienen las fuentes de tráfico pago en los sitios.

Biplot

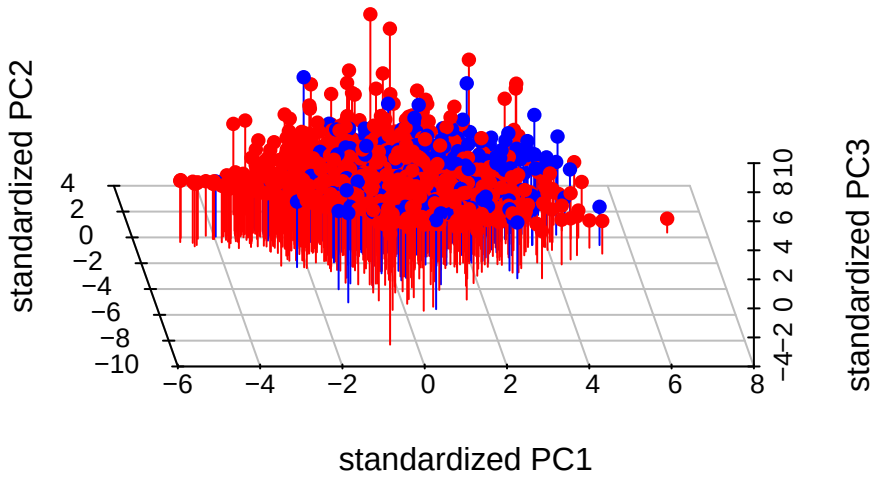


Al graficar las dos primeras componentes en un biplot y asignándole a cada sitio un color de acuerdo a si pertenecen al top 50.000, comenzamos a ver algunos patrones. Los sitios que tienen mayor tráfico se concentran a la derecha. De acuerdo a lo que veíamos en los loadings, sobre la primer componente se está representando el engagement. Por otro lado, sobre el biplot podemos resaltar algunas cuestiones:

- Correlación positiva entre el tráfico directo y el tráfico referido.
- Correlación negativa entre la tasa de rebote y las páginas vistas.

La componente 1 posee un peso importante sobre el tráfico directo y el tráfico referido. En la WWW los sitios se suelen clasificar como Hub o Authority. Un sitio es un Authority si está siendo linkeado desde otros sitios, y por otro lado, un sitio es un Hub si generalmente su comportamiento es linkear hacia otros sitios. En el Biplot estamos observando que los sitios con mayor tráfico proyectan sobre valores positivos de la componente 1, que ademaś de tener altos niveles de actividad, sus fuentes de tráfico guardan relación con los de Authority.

Sobre el siguiente gráfico se representan las 3 primeras componentes y se puede apreciar con mayor detalle la diferencia en las proyecciones de ambos grupos.

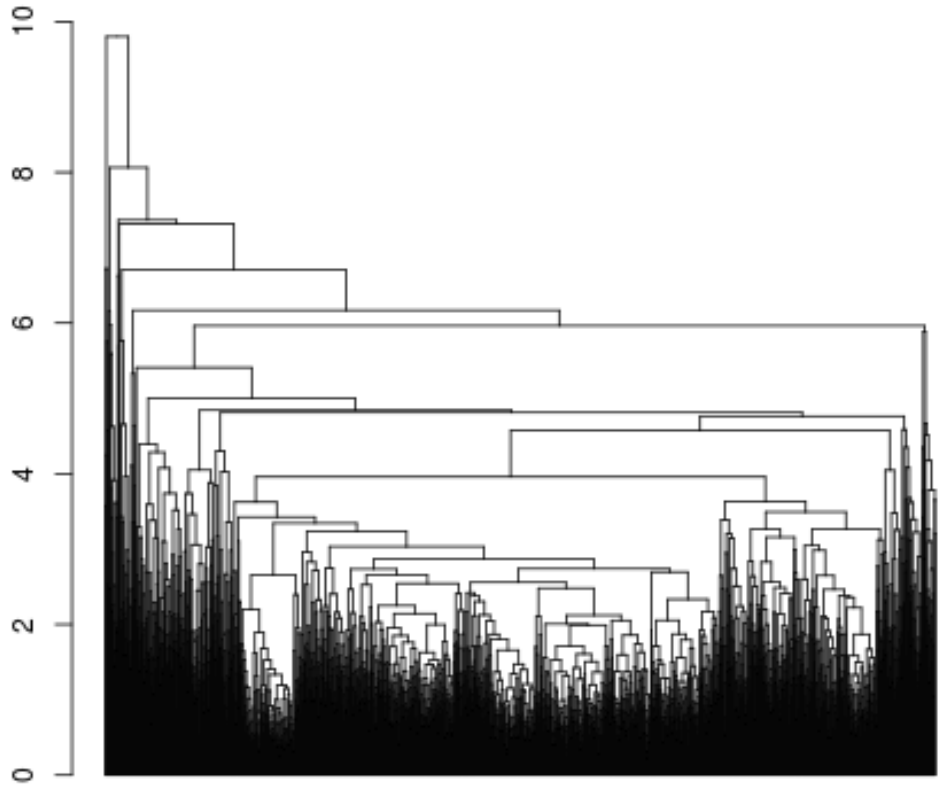


Clustering

Jerárquico

El coeficiente de correlación cofenético brinda información sobre la capacidad que tiene un agrupamiento jerárquico de responder a la matriz de distancias de las variables analizadas.

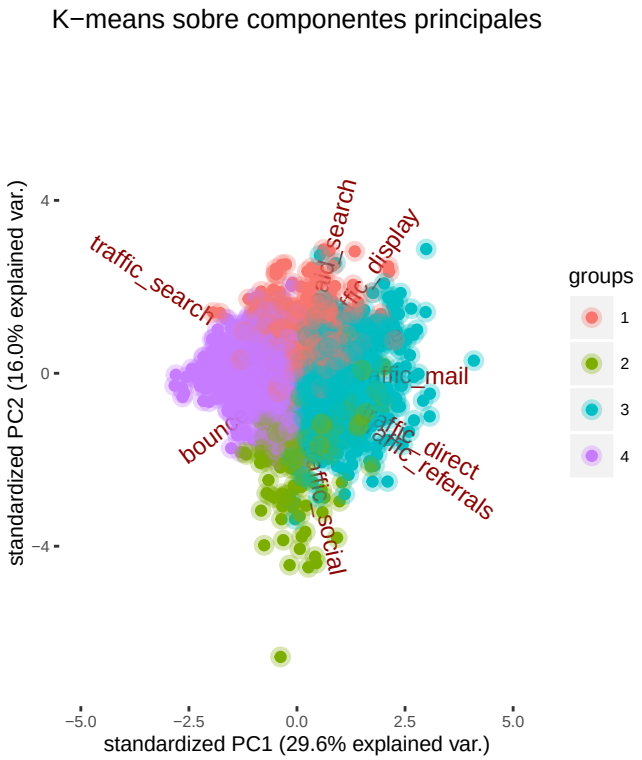
Luego de algunas pruebas de diferentes métodos, se observa que el método *average* brinda el mejor resultado, obteniendo una correlación cofenética de 0.77, que es bastante aceptable.



El gráfico sugiere que la cantidad de grupos existentes es elevada, parecerían existir diferentes

ploblaciones. Para lograr que los resultados se puedan interpretar con claridad, vamos a definir 4 grupos que se buscarán mediante k-means. Los grupos se grafican sobre las componentes que habían sido identificadas previamente.

K-means



No parecerían existir grupos claramente identificados. Sin embargo, si se realiza una matriz de contingencias del cluster con la variable *position*, se visualiza que algunos grupos poseen una proporción mayor de sitios con tráfico elevado.

cluster	0	1
1	259	106
2	122	21
3	440	187
4	704	89

En base al agrupamiento obtenido por k-means se visualiza que en el grupo 3 es donde se concentran la mayor cantidad de sitios con elevado tráfico. Podríamos suponer que dentro del grupo 3 se encuentran los sitios con mejor desempeño en cuanto al nivel de interacción y calidad de las fuentes de tráfico, en donde dicho comportamiento es independiente a la cantidad de visitas que reciben. Esto es una hipótesis para explorar y que podemos hacer una primera aproximación mirando el vector de medias de cada uno de los grupos:

	V1	V2	V3	V4
Group.1	1	2	3	4
bounce	40.45811	46.58832	34.98708	45.02733
paid_search	38.133342	7.369371	11.379027	8.277062
ppv	5.019808	3.825105	6.575694	3.437718
time	212.4137	142.6154	263.6587	130.7503
traffic_direct	22.65986	21.16154	34.77341	19.81710
traffic_display	3.6406575	0.5546154	0.6288038	0.2980580
traffic_search	54.04164	33.90420	37.90638	67.12605
traffic_mail	5.600247	12.359231	5.123636	1.889294
traffic_social	2.165479	15.557063	3.006268	1.766103
traffic_referrals	11.891781	16.463007	18.562041	9.103468

A primera vista el cluster 3 muestra mejores valores para las variables analizadas:

- Bajo nivel de rebote
- Elevado tiempo de visita
- El tráfico directo es importante.

Y por el contrario, el cluster 4 agrupa sitios con resultados pobres desde el punto de vista de la interacción de los usuarios con el sitio.

El agrupamiento por k-means nos brinda una separación interesante de sitios de acuerdo a su comportamiento. Para facilitar la interpretación podríamos asignarles un nombre a cada grupo:

- Cluster 1: Sitios con bajo nivel de interacción.
- Cluster 2: Sitios con tráfico pago.
- Cluster 3: Sitios con alto nivel de interacción.
- Cluster 4: Sitios con acciones en motores de búsqueda.

Análisis discriminante

Continuando con el análisis de nuestra variable respuesta *position* que indica si un sitio está en las posiciones más altas respecto a tráfico web, vamos a realizar un análisis discriminante, que nos permitirá encontrar una función que separe ambos grupos. Procederemos a aplicar el análisis discriminante cuadrático ante la falta de normalidad de las variables en análisis.

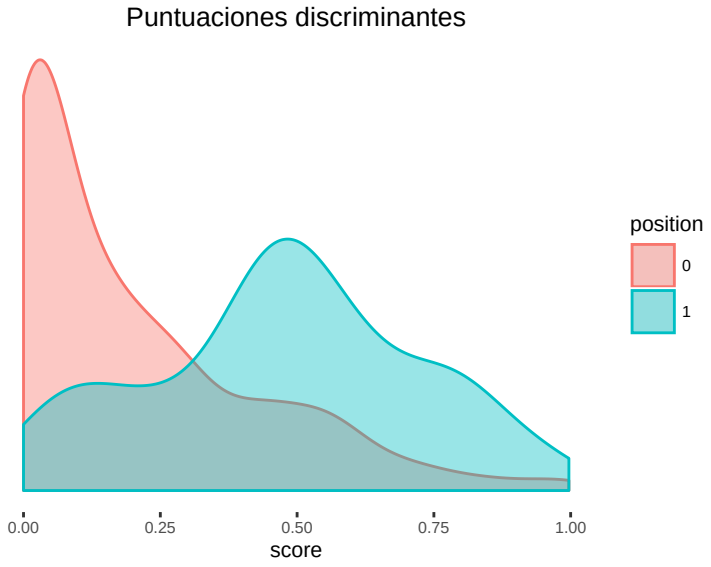
Para realizar el modelo, se separó el conjunto de datos en entrenamiento y test, haciendo un split de 70/30. Los resultados obtenidos que se muestran a continuación corresponden al conjunto de test.

Matriz de confusión

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 333  57
```

```
##          1  48  44
##
##                Accuracy : 0.7822
##                  95% CI : (0.7426, 0.8182)
##     No Information Rate : 0.7905
##     P-Value [Acc > NIR] : 0.6956
##
##                   Kappa : 0.3201
##  Mcnemar's Test P-Value : 0.4350
##
##             Sensitivity : 0.43564
##             Specificity : 0.87402
##          Pos Pred Value : 0.47826
##          Neg Pred Value : 0.85385
##              Prevalence : 0.20954
##          Detection Rate : 0.09129
##    Detection Prevalence : 0.19087
##       Balanced Accuracy : 0.65483
##
##        'Positive' Class : 1
##
```

Si bien tenemos algo de capacidad para clasificar observaciones, los resultados podrían ser mejores. Al visualizar la distribución de las puntuaciones obtenidas por la función discriminante se puede comprender mejor que el modelo está logrando identificar los grupos con un desempeño medianamente aceptable.



Nuestro modelo resultante del análisis discriminante cuadrático puede ser empleado para diagnosticar un sitio y comprender si sus características se asemejan a las de los mejores sitios webs a nivel global.

Software

Para la preparación de los datos de desarrollaron algunos scripts en Python trabajando con notebooks interactivos en Jupyter por la necesidad de procesamiento y flexibilidad. En cuanto al análisis de los datos se utilizó R mediante la aplicación RStudio, tanto para gestionar el proyecto, como para redactar la presente documentación. El código siempre se mantuvo actualizado en un repositorio del servicio GitHub.

Conclusiones

Luego del análisis de las variables, se puede concluir que existen diferencias significativas entre sitios con poco y elevado tráfico, tanto desde el punto de vista del comportamiento de parte de los usuarios, como así también en cuanto a medios de generación de tráfico.

Los sitios con elevado tráfico poseen un mayor promedio de tiempo de visita. Además de ello, se ha podido observar con componentes principales, que los sitios con elevado tráfico poseen valores altos proyectados sobre la primer componente, que representa el nivel de interacción y se asemejan a la definición de un sitio web clasificado como Authority de acuerdo a las fuentes de tráfico.

Al realizar el agrupamiento de los sitios utilizando todas las variables del conjunto de datos, se pudieron generar cuatro grupos que tienen características claramente definidas.

Mediante el análisis discriminante cuadrático es posible clasificar con un desempeño medianamente aceptable a los sitios de acuerdo a su nivel de tráfico y posicionamiento global. Esto se basa en las diferencias significativas que existen en los vectores de medias para cada grupo.

Las variables con las que se trabajó no logran tener un alto nivel explicativo de la variable respuesta *position*. Hay que tener en cuenta que los datos son obtenidos en base a estimaciones, y una posible interpretación de esto es que se esté introduciendo algo de ruido que suavicen las diferencias.

Bibliografía

- Box-Cox Transformations: An Overview. Department of Statistics, University of Connecticut.
- Introduction to Information Retrieval. Cambridge University Press.
- Hubs and Authorities on the Internet. The Mathematics of Web Search.