UNIVERSIDAD CENTRAL DE VENEZUELA
FACULTAD DE CIENCIAS
ESCUELA DE MATEMÁTICA

Aplicación de estadística multivariada para caracterizar palabras claves de Adwords usadas por una compañía de ventas por internet

Trabajo Especial de Grado presentado ante la ilustre Universidad Central de Venezuela por el **Br. José Antonio Castillo** para optar al título de Licenciado en Matemática.

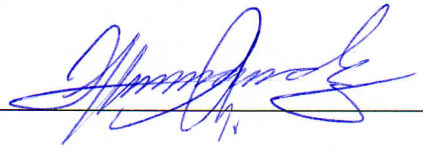
Tutor: Dra. Mairene Colina.

Caracas, Venezuela

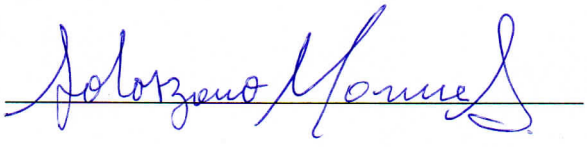
Julio 2018

Nosotros, los abajo firmantes, designados por la Universidad Central de Venezuela como integrantes del Jurado Examinador del Trabajo Especial de Grado titulado "**Aplicación de estadística multivariada para caracterizar palabras claves de Adwords usadas por una compañía de ventas por internet**", presentado por el **Br. José Antonio Castillo**, titular de la Cédula de Identidad **24.636.302**, certificamos que este trabajo cumple con los requisitos exigidos por nuestra Magna Casa de Estudios para optar al título de **Licenciado en Matemática**.

Dra. Mairene Colina
Tutor



Dr. José Benito Hernández
Jurado



Lic. Manuel Solórzano
Jurado

Dedicatoria

A Dios.

A mi abuela y mis padres

Agradecimiento

Ante todo le doy gracias a Dios por ser mi guía y darme todo lo necesario para lograr mis objetivos y metas.

A mi abuela Eumelia por su amor y apoyo incondicional en todo momento.

A mis padres por los valores que me inculcaron y su apoyo.

A mis amigos que me acompañaron durante toda la carrera en especial a Yoselin Arvelaiz por motivarme y siempre estar en los momentos más duros de la carrera.

A los profesores que me formaron durante la carrera, especialmente a mi tutora Mairene Colina quién me alentó a seguir cuando más dudas tuve. También agradecerle a la prof. Angie Pineda que con sus consejos me ayudaron en el transcurrir de la carrera.

Resumen

Google lanzó al mercado en el año 2000 **AdWords**, su sistema de publicidad en línea y principal fuente de ingresos. El sistema de publicidad **AdWords** permite orientar los anuncios de las empresas con palabras claves mediante un modelo de *pago por clic (PPC)*, las **AdWords** son un conjunto de palabras claves que usan los anunciantes para que sus anuncios aparezcan una vez que las mismas son colocadas en el buscador de Google. Conocer que palabras claves tuvieron mas impacto en el anuncio, saber cual fue el rendimiento de estas palabras, que tan eficaces fueron o que características presentan, son de un gran valor informativo para las empresas, pues con esto pueden tomar mejores decisiones con respecto a que palabras vale la pena apostar e invertir, modificar su presupuesto acorde a su propósitos, entre otros.

Este trabajo muestra la utilidad del análisis multivariado de datos pues en él se implementan técnicas de agrupamiento y reducción de dimensionalidad derivadas del aprendizaje no supervisado de la minería de datos que permiten segmentar a las palabras clave, de una cierta campaña publicitaria de una empresa de artículos de sombreros, en grupos los cuales son posteriormente caracterizados mediante estadística descriptiva.

Palabras claves: Cluster, Agrupación, Google, Componentes Principales, Minería de Datos, AdWords.

Índice general

Introducción

El enorme impacto que ha tenido el Internet en nuestra sociedad ha generado un cambio radical en la comunicación hoy en día. La publicidad ha encontrado en la red un espacio para atraer a los usuarios a las diferentes marcas. Así, muchas empresas han visto la idea de promocionarse en la web como una manera más económica, fácil y eficaz de posicionarse en sus respectivos mercados. Conforme las sociedades en masas comenzaban a generarse también comenzó la necesidad de acceder de manera masiva a la información, con la llegada del Internet el acceso a la información se hizo mucho más fácil generando así un medio de comunicación ideal para la publicidad. El Internet, como medio publicitario, tiene como valor más destacable el de permitir a todo tipo de anunciantes sin importar el tamaño de la empresa poder promocionarse, posibilitando de esta manera la competencia entre empresas nacionales o multinacionales en igualdad de condiciones, planificando sus campañas en línea. En este sentido **Google** lanzó al mercado en el año 2000 **AdWords**, su sistema de publicidad en línea y principal fuente de ingresos.

El sistema de publicidad **AdWords** permite orientar los anuncios de las empresas con palabras claves mediante un modelo de *pago por clic (PPC)*, las **AdWords** son un conjunto de palabras claves que usan los anunciantes para que sus anuncios aparezcan una vez que las mismas son colocadas en el buscador de Google.

Conocer que palabras claves tuvieron mas impacto en el anuncio, saber cual fue el rendimiento de estas palabras, que tan eficaces fueron o que características presentan, son de un gran valor informativo para las empresas, pues con esto pueden tomar mejores decisiones con respecto a que palabras vale la pena apostar e invertir, modificar su presupuesto acorde a su propósitos, entre otros.

Por esta razón en este trabajo se presenta un análisis multivariado de datos usando técnicas de agrupamiento y reducción de dimensionalidad derivadas del aprendizaje no supervisado de la minería de datos para poder caracterizar a los grupos resultantes mediante

estadística descriptiva a un conjunto de palabras claves de una cierta campaña publicitaria de una empresa de artículos de sombreros desde septiembre de 2017 hasta octubre de 2017, esto nos permitirá saber que grupo de palabras influyeron o no en la promoción de un anuncio.

El presente trabajo está estructurado en tres capítulos, para el primer capítulo se muestran los fundamentos de la minería de datos y características del sistema Adwords de Google. El segundo capítulo se dan conceptos y resultados asociados a las metodologías de las técnicas mencionadas en el capítulo 1. Finalmente el tercer capítulo mostraremos los resultados obtenidos a partir de los métodos explicados en el capítulo 2.

Capítulo 1

La minería de datos y el sistema AdWords

En este capítulo daremos los aspectos principales de la minería de datos y sus estrategias para generar conocimientos a partir de los datos, además se muestra un breve resumen del sistema Adwords y sus principales características.

1. Minería de datos y sus técnicas

La minería de datos es el proceso no trivial de identificar, a partir de datos, patrones válidos, novedosos, potencialmente útiles y comprensibles para poder generar conocimiento y realizar procesos que permitan un mejor acierto en las tomas de decisiones. La minería de datos está conformada por un arreglo de estrategias que nos permiten llevar a cabo estas acciones tales como lo son las tecnología de bases de datos, la visualización de datos, estadística, el aprendizaje automático, la inteligencia artificial, entre otras disciplinas.

Las técnicas más usadas en la minería de datos son las *técnicas predictivas* del *Aprendizaje Supervisado* y las *técnicas descriptivas* del *Aprendizaje No Supervisado*, los cuales vamos a ver con más detalle a continuación:

1.1. Aprendizaje Supervisado.

El *Aprendizaje Supervisado* tiene como objetivo hacer predicciones a futuro basadas en comportamientos o características que se observan en los datos almacenados, buscando patrones en estos para luego poder ajustar un modelo que nos permita hacer inferencia y poder así predecir con la mayor precisión posible.

Algunas de las técnicas predictivas más importantes y usadas son las siguientes:

- Máquinas de Soporte Vectorial.
- Arboles de clasificación o regresión.
- Redes Neuronales.

- K-vecinos más cercanos (KNN).
- Redes Bayesianas.
- Series Temporales.

1.2. Aprendizaje No Supervisado.

El *Aprendizaje No Supervisado* tiene como objetivo explorar y obtener información sobre las posibles relaciones entre los datos para así poder armar una estructura que nos permita etiquetar estas asociaciones entre los datos para luego obtener conocimiento de estos.

Algunas de las técnicas descriptivas más importantes y usadas son las siguientes:

- Análisis de agrupamiento.
- Detección de anomalías.
- Reglas de asociación.
- Análisis de Componentes Principales (ACP).

El presente trabajo se enfocará usando 2 técnicas del *Aprendizaje No Supervisado* como lo son el análisis de agrupamiento y el análisis de componentes principales a un conjunto de datos del sistema **AdWords** de Google, el cual vamos a ver con más detalles en la próxima sección.

2. Google AdWords

Google AdWords es el sistema de publicidad en línea de Google con la cual las empresas de cualquier tipo pueden promocionarse colocando sus anuncios en las páginas de busqueda de Google, así como tambien en varios sitios web que pertenezcan a la red publicitaria de Google. Este sistema permite orientar los anuncios de las empresas mediante palabras claves o *keywords* que no son más que un conjunto de palabras que usan los anunciantes para mostrar sus anuncios una vez que las mismas son colocadas en el buscador de Google.

El sistema **AdWords** funciona bajo un modelo de subasta donde los anunciantes compiten emitiendo pujas para posicionar sus anuncios en los primeros resultados de Google mediante las palabras claves, para así poder atraer a potenciales clientes hacia la pagina web del anunciante. Este sistema también usa el modelo de *pago por clic (PPC)* lo cual resulta ser bastante atractivo a las empresas y anunciantes pues estos solo pagarán cuando un usuario haya hecho clic a su anuncio siendo así una manera efectiva y económica de pagar por publicidad.

Algunas características relevantes de **Google AdWords** son:

- No hay requisitos de inversión mínima.
- Establecer y controlar su presupuesto.
- Medir el impacto de su anuncio.
- Las cuentas de AdWords se administran en línea.
- Modificar el texto de sus anuncios.
- Elegir dónde aparecerá el anuncio.

Cada palabra clave o *keyword* tiene asociada variables que describen el rendimiento que desempeñaron en un período de tiempo, algunas de estas son:

- **Clic:** Cantidad de clics que recibió el anuncio.
- **Impresiones:** Número de veces que se muestra el anuncio en la red de Google.
- **Tasas de clics (CTR):** Proporción que muestra con que frecuencia los usuarios que ven el anuncio hacen clic en el, es decir, la proporción de clics entre las impresiones.

- **Campaña:** Grupo de palabras claves que comparten un presupuesto, una orientación geográfica y otros parámetros de configuración.
- **Nivel de calidad:** Estimación de la calidad de los anuncios y palabras claves en una escala del 1 al 10.
- **Conversión:** Acción que se cuenta cuando una persona interactúa con su anuncio y lleva a cabo una acción que el anunciante define como valiosa para su empresa .
- **Costo total:** Importe total por los clics que recibió el anuncio.
- **Costo promedio:** Importe promedio por cada clic que recibió el anuncio, es decir, el costo total entre la cantidad de clics.
- **Oferta máxima de costo por clic:** Es la oferta que establece el anunciante para determinar el importe más alto que está dispuesto a pagar por clic.
- **Posición promedio:** Determina el orden promedio de aparición de los anuncios en la página.

Capítulo 2

Técnicas del Análisis Multivariado de Datos No Supervisado

Como se mencionó en el Capítulo 1, las técnicas del análisis multivariado usadas en el presente trabajo fueron el análisis de agrupamiento y el análisis de componentes principales. En lo que sigue daremos los conceptos y resultados básicos asociados a estas metodologías.

1. Análisis de agrupamiento

El análisis de agrupamiento consiste en un conjunto de métodos númericos cuyo objetivo en común es encontrar o descrubrir grupos o *'clusters'* de un conjunto de datos que sean homogéneos entre sí y diferentes de los otros grupos. El propósito de dicho análisis es identificar patrones en los grupos resultantes.

Los métodos de agrupamiento principales son:

- Basados en particiones: Son algoritmos que dividen los datos en k clusters, donde el número entero k es especificado por el usuario.
- Agrupamiento jerárquico: Este método a su vez se subdivide en dos métodos: Los algoritmos *aglomerativos* y los *divisivos*. Los algortimos aglomerativos consisten en que cada observación forma su propio cluster, luego los clusters con mayor similitud se agrupan formando un nuevo cluster, este proceso se repite hasta que solo quede un solo cluster el cual contiene a todo el conjunto de datos. Por otro lado los algoritmos divisivos empiezan con el conjunto de datos como un solo cluster y a partir de ahi este se divide en varios clusters hasta que cada observación se convierta en su propio cluster.

1.1. Medidas de distancia o similitud.

Es de central importancia conocer que tan 'cerca' o 'lejos' están las observaciones entre sí para poder agruparlos, conocer esta medida de cercanía se le conoce como similitud (disimilitud) ó distancia. Dos observaciones son cercanas si tienen una disimilitud o distancia corta o una similitud grande.

La elección de la distancia es de gran peso para las agrupaciones pues dependiendo de la similitud entre las observaciones esta influirá en la conformación de los grupos. Las medidas de disimilitud están divididas entre *las medidas de distancia* y *las medidas de distancia basadas en la correlación.*

Algunas de las medidas de distancia más usadas son:

1. *Distancia Euclídea:*
$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}.$$
2. *Distancia Manhattan:*
$$d(x,y) = \sum_{i=1}^{n}|x_i - y_i|.$$
3. *Distancia Canberra:*
$$d(x,y) = \sum_{i=1}^{n}\frac{|x_i - y_i|}{|x_i| + |y_i|}.$$

Donde $x = (x_1, ..., x_n)$ e $y = (y_1, ..., y_n)$ son dos observaciones o vectores de tamaño n.

Asociadas a las medidas de distancia basadas en correlación tenemos:

1. *Distancia de correlación Pearson:*
$$d(x,y) = 1 - \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

2. *Distancia de correlación del Coseno Eisen:*

$$d(x,y) = 1 - \frac{\left|\sum_{i=1}^{n} x_i y_i\right|}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}}.$$

Con $x = (x_1, ..., x_n)$ e $y = (y_1, ..., y_n)$ dos observaciones. Otras medidas de distancias de este tipo son:

- *Distancia de correlación Kendall.*
- *Distancia de correlación Spearman.*

1.2. Estandarización.

En la mayoría de los casos, al aplicar el análisis de agrupamiento, las variables que describen a las observaciones que serán agrupadas no se encuentran en las mismas unidades (por ejemplo: kilometros, altura, peso, kilogramos, etc..), por lo que al aplicar cualquier medida de distancia a estas variables continuas las medidas obtenidas se verán severamente afectadas. La solución más popular para lidiar con éste problema de unidades diferentes es la **Estandarización**, así logramos que las variables puedan ser comparables y evitar la influencia de unidades que afecten en la similitud de los datos.

Cuando estandarizamos las variables, para cada observación x_i tenemos:

$$\frac{x_i - center(x)}{scale(x)}.$$

Donde *center(x)* es una medida de tendencia central de las variables, donde la más común es la media y *scale(x)* la medida de dispersión, donde la más común es la desviación estándar.

Una vez escogidas la manera de estandarizar y la medida de distancia, se procede a elegir el algoritmo (ó los algoritmos) a usar para el agrupamiento de los datos, ya sean basados en

particiones ó mediante agrupamiento jerárquico. Explicaremos algunos de estos algoritmos con más detalles en la siguiente sección.

2. Algoritmos Basados en Particiones

2.1. Agrupación K-medias.

El algoritmo K-medias busca particionar al conjunto de datos de $\{x_1, .., x_m\}$ observaciones en k grupos o clusters $(G_1, G_2, ..., G_k)$, donde $G_{i,1\leq i\leq k}$ denota el i-ésimo grupo formado por $\{x_j\}_{1\leq j\leq m}$ observaciones y k ya previamente especificado por el usuario.

El objetivo del agrupamiento por K-medias es definir una cantidad de clusters de manera tal que la variación total inter-cluster definida normalmente como la suma al cuadrado de la distancia euclidea de las observaciones del i-ésimo grupo al correspondiente centroide sea mínima. La varianza inter-cluster para un grupo es:

$$W(G_i) = \sum_{x_j \in G_i} (x_j - \mu_i)^2.$$

Donde x_j son las observaciones pertenecientes al i-ésimo grupo y μ_i el centroide correspondiente definido como la media de las observaciones pertenecientes al i-ésimo grupo, es decir, μ_i está definido por:

$$\mu_i = \frac{1}{|G_i|} \sum_{x_j \in G_i} x_j.$$

La variación total inter-cluster se define como:

$$\text{Variación Total Inter-Cluster} = \sum_{i=1}^{k} W(G_i).$$

De manera natural, si la suma total de la variación inter-cluster de los k grupos formados es lo más pequeña posible, obtendremos una medida de calidad óptima del agrupamiento.

El algoritmo empleado en K-medias se puede resumir de la siguiente forma:

1. Especificar el número de grupos (k).
2. Se seleccionan aleatoriamente k observaciones como centroides iniciales o medias.
3. Asignar a cada observación al centroide más cercano, basado en la distancia euclídea.

4. Para cada uno de los k grupos recalcular el centroide con la nueva media de las observaciones de cada grupo.
5. Repetir los pasos 3 y 4 hasta que las asignaciones no cambien ó se alcance el número máximo de iteraciones establecido.

Dado que el algoritmo de K-medias no evalúa todas las posibles distribuciones de las observaciones sino parte de ellas, los resultados dependerán de las asignaciones aleatorias realizadas en el paso 1, por esta razón se recomienda en gran medida ejecutar el algoritmo varias veces (30-50) con asignaciones aleatorias distintas y escoger de esta manera aquella agrupación que obtenga la variación total Inter-Cluster más pequeña.

2.2. Agrupación K-medoids ó PAM (Partitioning Around Medoids).

El algoritmo PAM es muy similar al algoritmo de K-medias, pues en ambos algortimos se agrupan las observaciones en K clusters, donde K es un valor preestablecido por el usuario. La diferencia es que en PAM, cada cluster está representado por una observación presente en el mismo (conocido como *medoid*), mientras que en el algoritmo de K-medias cada cluster está representado por su centroide que corresponde con el promedio de las observaciones de dicho cluster.

El objetivo de PAM es encontrar un subconjunto de observaciones (*medoids*) $\{m_1, ..., m_k\} \subset \{x_1, ..x_m\}$ del conjunto de datos tal que la suma total de la distancias de todas las observaciones a su *medoid* más cercano sea mínima, es decir, encontrar $\{m_1, ..., m_k\}$ tal que la expresión dada por:

$$\sum_{i=1}^{k} \sum_{x_j \in G_i} d(x_j, m_i).$$

sea mínima, donde $G_i = \{x_j : d(x_j, m_i) = \min_{i=1,..,k} d(x_j, m_i)\}$.

El algoritmo empleado en PAM se puede resumir de la siguiente manera:

1. Seleccionar K observaciones como medoids iniciales. También es posible identificarlas de forma específica.

2. Calcular la matriz de distancia entre todas las observaciones si esta no se ha calculado.
3. Asignar cada observación a su medoid más cercano.
4. Para cada uno de los clusters creados, comprobar si seleccionando otra observación como medoid se consigue reducir el coeficiente de distancia promedio de toda la agrupación (es decir, de todos los grupos), si esto ocurre seleccionar esa observación como nuevo medoid.
5. Si al menos un nuevo medoid es escogido en el paso 4, volver al paso 3, de lo contrario se termina el proceso.

La agrupación por K-medoids es un método más robusto que la agrupación por K-medias, es decir, es menos sensible a valores atípicos por lo que se recomienda usar este método si se sospecha de la presencia de estos en el conjunto de datos.

Al igual que en la agrupación por K-medias, necesitamos que se especifique de antemano el número de grupos que se van a crear. Esto puede ser complicado de determinar si no se dispone de información adicional sobre los datos. Muchas de las estrategias empleadas en K-medias para identificar el número óptimo, pueden aplicarse en K-medoids.

2.3. Estimacón de K clusters y Validación.

Determinar el número óptimo de grupos en un conjunto de datos es un problema fundamental en las agrupaciones basadas en particiones. Lamentablemente, no hay una solución definitiva a esta interrogante, el número óptimo de grupos es relativo y depende del método usado para medir similitudes y los parametros usados en la partición. A su vez conocer y medir que tan bien son los grupos resultantes es de gran importancia pues la idea principal es agrupar las observaciones que sean similares a aquellas que se encuentran en el mismo cluster y distintas a las observaciones de los otros clusters. A continuación se presentan dos métodos que ayudarán a determinar estos K clusters y a su vez a la calidad/significancia de estos.

- **Método del Codo *(Elbow method)***

Recordemos que en la agrupación K-medias el objetivo es encontrar K clusters tal que la varianza total inter-cluster sea mínima. El método del Codo consiste en calcular la varianza

total inter-cluster en función del número de grupos y escoge como óptimo aquel valor a partir del cual la reducción de la varianza total deja de ser sustancial, este método se presenta de manera gráfica visualizando el punto óptimo donde se genera dicho proceso, a este punto se le conoce como 'codo' (*knee*).

El método sigue de la siguiente manera:

1. Calcular el algortimo de K-medias para diferente valores de K, usualmente variando K de uno a diez.
2. Para cada K, calcular la varianza total inter-cluster.
3. Graficar la curva de la varianza total inter-cluster acorde al número de clusters K.
4. Localizar el 'codo' en la gráfica, usualmente considerado como un indicador del número apropiado de clusters.

Veamos el siguiente ejemplo:

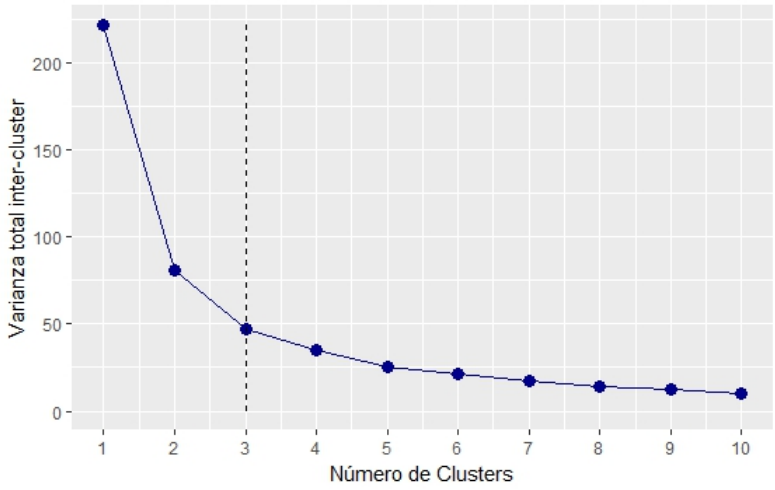


FIGURA 2.1. Método del Codo indicando $K = 3$ como valor óptimo

Obtener una varianza total inter-cluster lo más pequeña posible nos ayudará a conocer la calidad del cluster formado, pues al igual que el R^2 de los modelos de regresión, el agrupamiento por K-medias muestra el *radio de la suma de cuadrados ó Ratio_SS* que nos indica el porcentaje de varianza explicada por el modelo con respecto al total de la varianza de los datos, esto es:

$$\text{Ratio\_SS} = 1 - \frac{\text{Variación total inter-cluster}}{\text{Variación total}}.$$

- **Método de la Silueta *(Silhoutte method)***

El análisis de la Silueta mide que tan bién se agrupó una observación comparando su similitud con el resto de observaciones de su cluster frente a las de los otros clusters. Para conocer esta medida calculamos el índice de silueta $S_{(x_j)}$ para cada observación $x_j \in \{x_1, .., x_m\}$, donde $\{x_1, .., x_m\}$ son todas las observaciones del conjunto de datos. El índice de silueta $S_{(x_j)}$ se calcula de la siguiente manera:

1. Para la observación $x_j \in G_i$ definimos $a_{(x_j)}$ como:

$$a_{(x_j)} = \frac{1}{|G_i|} \sum_{\substack{x_j, x_l \in G_i \\ x_j \neq x_l}} d(x_j, x_l).$$

2. Para todos los clusters G_z con $z \in \{1, .., k\} \backslash \{i\}$ definimos:

$$\bar{d}(x_j, G_z) = \frac{1}{|G_z|} \sum_{x_p \in G_z} d(x_j, x_p).$$

3. Una vez calculado $\bar{d}(x_j, G_z)$ para todos los clusters definimos a $b_{(x_j)}$ como:

$$b_{(x_j)} = \min_{z \in \{1,..,k\} \backslash \{i\}} \bar{d}(x_j, G_z).$$

4. Finalmente el índice de silueta está dado por:

$$S_{(x_j)} = \frac{b_{(x_j)} - a_{(x_j)}}{max\{b_{(x_j)}, a_{(x_j)}\}}.$$

Podemos observar que el valor $S_{(x_j)}$ esta entre -1 y 1, siendo el valor 1 un indicativo que la observación se ha asignado al grupo correcto y -1 como una mala asignación.

El método de la silueta consiste en calcular el promedio de todos los índices de silueta $S_{(x_j)}$ de las observaciones y escoger el valor K tal que maximiza este promedio, es decir, escoger el valor K tal que la expresión dada por:

$$\bar{S} = \frac{1}{m} \sum_{i=1}^{m} S_{(x_j)}.$$

sea máxima. Al igual que el método del codo, este método se presenta de manera gráfica para así encontrar el valor K óptimo.

Podemos resumir el método de la siguiente manera:

1. Calcular el algoritmo de agrupación (por ejemplo PAM) para diferentes valores de K, usualmente variando K de uno a diez.
2. Para cada K, calcular el promedio de los índices de silueta ($\bar{S}$).
3. Graficar la curva de los valores $\bar{S}$ acorde al número de clusters K.
4. Localizar el valor K que maximiza $\bar{S}$ en la gráfica, usualmente considerado como un indicador del número apropiado de clusters.

Veamos el siguiente ejemplo:

FIGURA 2.2. Método de la Silueta indicando $K = 2$ como valor óptimo

Obtener un promedio $\bar{S}$ lo más grande posible nos indica una mejor calidad del agrupamiento formado, podemos interpretar el valor $\bar{S}$ de la siguiente forma:

Valor $\bar{S}$	**Interpretación**
0.71-1	Fuerte estructura
0.51-0.70	Estructura razonable
0.26-0.50	Estructura débil o superficial
<0.25	Estructura no encontrada

TABLA 2.1. Interpretación del valor $\bar{S}$

3. Algoritmos jerárquicos

En el agrupamiento jerárquico, los grupos se caracterizan por formar una jerarquía representados bajo una estructura de árbol conocido como *dendograma*.

A diferencia de las agrupaciones basadas en particiones que de acuerdo a la asignación de los centroides (o medoids) los grupos se van modificando hasta lograr la convergencia, el agrupamiento jerárquico efectúa una inspección para agrupar basado en las similitudes de las observaciones, de modo que los resultados de los grupos que se van formando (sean por aglomeración ó división) no podrán modificarse en los pasos sucesivos, solamente se irán anidando a otros grupos (en el caso aglomerativo) ó dividiendo (en el caso divisivo).

Este método no requiere pre-especificar de antemano el número de grupos que se crearán, la elección de los grupos se basa de acuerdo a la visualización del dendograma ó estructura de árbol.

En esta sección solo mencionaremos los algortimos jerárquicos más usados, estos son:

1. **Aglomeración por anidación ó AGNES (Agglomerative Nesting)**.
2. **Análisis divisivo ó DIANA (Divisive Analysis)**.
3. **Análisis monotético ó MONA (Monothetic Analysis)**.

4. Análisis de componentes principales

Grandes conjuntos de datos que contienen multiples observaciones y variables son recolectados cada día por investigadores en varios campos tales como la medicina, finanzas, páginas web, entre otros. Descrubir conocimiento de estos datos extraídos requiere técnicas específicas para analizar estos conjuntos de datos que contienen multiples variables. El análisis de componentes principales (ACP) nos permite resumir y visualizar la información de un conjunto de datos que contienen observaciones descritas por multiples inter-correlacionadas variables cuantitativas.

El análisis de componentes principales es un método de reducción de dimensionalidad que consiste en extraer la información de las variables (usualmente correlacionadas entre sí) de un conjunto multivariado de datos y expresar esta información en un conjunto nuevo de variables no correlacionadas conocidas como *componentes principales*. Estas nuevas variables corresponden a una combinación lineal de las variables originales. La cantidad de componentes principales es menor o igual a la cantidad de las variables originales.

El objetivo del ACP es representar en las primeras dos ó tres componentes principales la mayor cantidad de variabilidad de todo el conjunto de datos y poder así visualizarlas gráficamente obteniendo de esta manera una mínima perdida de información. El resto de las componentes principales representan el resto de la variabilidad de los datos.

Sea $x = (x_1, .., x_p)$ un vector de P variables con una correlación ó covarianza de interés entre sí. La primera componente principal se define como:

$$y_1 = \alpha_1 x^T = \alpha_{11} x_1 + ... + \alpha_{1p} x_p = \sum_{i=1}^{p} \alpha_{1i} x_i.$$

donde $\alpha_1 = (\alpha_{11}, .., \alpha_{1p})$ es un vector de P constantes también conocidos como pesos. Buscamos que la varianza de y_1 (Var$[y_1]$, la cual se puede escribir en términos de la matriz de correlación\covarianza) sea máxima a partir de α_1, para esto se impone la siguiente restricción:

$$\alpha_1 \alpha_1^T = \sum_{i=1}^{p} \alpha_{1p}^2 = 1.$$

La técnica usual para maximizar la varianza de y_1 son los multiplicadores de Lagrange, el resultado de esto es que el vector α_1 que maximiza la varianza de y_1 es el autovector

correspondiente al autovalor (λ_1) mas grande de la matriz de correlación\covarianza de las x_p variables originales.

Análogamente para la segunda componente principal definida por:

$$y_2 = \alpha_2 x^T = \alpha_{21}x_1 + ... + \alpha_{2p}x_p = \sum_{i=1}^{p} \alpha_{2i}x_i.$$

donde $\alpha_2 = (\alpha_{21}, .., \alpha_{2p})$, usamos la misma restricción que se impuso a los pesos de la primera componente principal, es decir, $\alpha_2\alpha_2^T = \sum_{i=1}^{p} \alpha_{2p}^2 = 1$. Como las componentes principales no pueden estar correlacionadas entre sí añadimos una nueva restricción que esta dada por:

$$\alpha_2\alpha_1^T = \sum_{i=1}^{p} a_{2i}a_{1i} = 0.$$

Esta condición nos garantiza que y_1 y y_2 no estén correlacionadas, aplicando los multiplicadores de Lagrange como se realizó para y_1, el vector α_2 que maximiza la varianza de y_2 es el autovector asociado al segundo autovalor (λ_2) más grande de la matriz de correlación\covarianza.

Este proceso continúa para todas las componentes principales $y_j = \alpha_j x^T$, siempre sujetas a las condiciones $\alpha_j\alpha_j^T = 1$ y $\alpha_j\alpha_i^T = 0 (i \neq j)$. La aplicación de la técnica de los multiplicadores de Lagrange demuestra que el vector de pesos α_j de la j-ésima componente principal, es el autovector asociado al j-ésimo autovalor (λ_j) más grande de la matriz de correlación\covarianza y además que la varianza de la j-ésima componente principal (Var[y_j]) está dada por λ_j.

Finalmente la elección de la matriz de correlación o de covarianza dependerá de la distribución de los datos. En caso de que los datos sean homogéneos entre sí se procede a usar la matriz de covarianza de lo contrario se eligirá en su lugar a la matriz de correlación.

Capítulo 3

Caracterización de las Palabras Claves.

En este capítulo mostraremos los resultados obtenidos al aplicar las técnicas desarrolladas en el Capítulo 2 aplicadas al conjunto de datos del sistema Adwords.

1. Origen de los Datos

Los datos usados en el presente trabajo provienen de una cierta campaña publicitaria de una empresa de artículos de sombreros, dichos datos comprende un registro de observaciones por hora desde septiembre de 2017 hasta octubre de 2017 y son resultados de búsqueda por parte de los usuarios del sistema Adwords desde un computador. Estos datos constan de 16.306 observaciones y 11 variables del sistema AdWords las cuales fueron descritas en el Capítulo 1.

2. Aplicación ACP al conjunto de datos

Se realizó un análisis de componentes principales al conjunto de datos debido a la alta dimensionalidad de variables que estos presentan, para esto se usaron solamente las variables de tipo numérica, quedandonos asi con 9 de las 11 variables mencionadas al principio del capítulo. Las variables usadas fueron:

- Clics.
- Tasas de clic (CTR).
- Oferta máxima de costo por clic.
- Impresiones.
- Costo total.
- Costo promedio.
- Posición promedio.
- Calidad del anuncio.
- Conversiones.

Al ser estas variables heterogéneas entre sí, el cálculo de las componentes principales se realiza a partir de la matriz de correlación.

Variables	Oferta maxima	Impresiones	Clics	Tasas de clics	Costo total	Promedio costo	Posicion Promedio	Calidad anuncio	Conversiones
Oferta maxima	1	0.50	0.32	0.01	0.44	0.36	-0.22	-0.22	0.004
Impresiones	0.50	1	0.55	-0.04	0.67	0.44	-0.03	-0.13	0.01
Clics	0.32	0.55	1	0.59	0.89	0.78	-0.08	0.02	0.11
Tasas de clics	0.01	-0.04	0.59	1	0.33	0.61	-0.09	0.15	0.15
Costo total	0.44	0.67	0.89	0.33	1	0.77	-0.08	-0.03	0.07
Promedio costo	0.36	0.44	0.78	0.61	0.77	1	-0.11	0.02	0.13
Posicion promedio	-0.22	-0.03	-0.08	-0.09	-0.08	-0.11	1	-0.03	-0.01
Calidad anuncio	-0.22	-0.13	0.02	0.15	-0.03	0.02	-0.03	1	0.01
Conversiones	0.004	0.01	0.11	0.15	0.07	0.13	-0.01	0.01	1

TABLA 3.1. Matriz de correlación.

En la Tabla 3.1 podemos observar que las variables conversiones, posición promedio y calidad del anuncio son aquellas que, en general, presentan una menor correlación con el resto de las variables, en cambio clics, costo total y costo promedio son las que tienen una mayor correlación con el conjunto de variables observadas.

Los autovectores (o los vectores pesos) que maximizan la varianza de las componentes principales que se calculan a partir de la Tabla 3.1 están dados en la Tabla 3.2. En dicha Tabla 3.2 se pueden observar no solo los vectores pesos asociados a cada componente principal sino también a cuales variables les corresponden cada valor de los pesos en las respectivas componentes principales. Podemos notar que para cuatro componentes principales, las variables que aportan más información son clics, tasas de clics, posición promedio y conversiones, pues poseen los pesos en valor absoluto más altos en cada componente principal. Vale la pena

mencionar que dos de estas cuatro variables son de las que tienen una menor correlación con respecto a las nueve variables (a saber, posición promedio y conversiones) y hay una de las que presentan mayor correlación con el resto como lo es la variable clic. Esto resulta lógico pues recordemos que el objetivo del ACP es reducir la dimensionalidad para obtener nuevas variables no correlacionadas con la menor pérdida de información posible.

Variables	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7	Comp 8	Comp 9
Oferta maxima	-0.288	0.449	-0.264	0.065	-0.059	0.782	-0.038	-0.156	0.031
Impresiones	-0.36	0.398	0.158	-0.052	-0.365	-0.257	0.665	0.202	-0.038
Clics	-0.488	-0.134	0.107	-0.066	0.05	-0.184	-0.07	-0.492	0.667
Tasa de clics	-0.286	-0.544	-0.063	0.012	0.424	0.231	0.527	-0.118	-0.307
Costo	-0.486	0.066	0.122	-0.086	-0.098	-0.22	-0.439	-0.246	-0.652
Promedio costo	-0.465	-0.169	0.028	-0.012	0.132	0.082	-0.283	0.785	0.183
Posicion promedio	0.093	-0.03	0.928	0.046	-0.02	0.354	-0.017	-0.023	-0.002
Calidad anuncio	0.023	-0.488	-0.102	-0.418	-0.723	0.233	-0.011	-0.003	0.008
Conversiones	-0.08	-0.234	-0.046	0.897	-0.362	-0.018	-0.021	-0.025	-0.007

TABLA 3.2. Autovectores de la matriz de correlación.

La varianza (autovalores) asociada a cada componente principal, su porcentaje de varianza explicada (P.V.E) y el porcentaje de la varianza explicada acumulada (P.V.E.A) se muestran en la Tabla 3.3. En dicha tabla observamos que las dos primeras componentes principales explican un 57.1 % de la variabilidad del conjunto de datos, para tres componentes se tiene un 68.7 % y así sucesivamente. En general la elección de la cantidad de componentes principales a ser utilizadas se hace considerando el P.V.E.A más significativo, para el presente trabajo se consideró el uso de dos y cuatro componentes principales las cuales explican el 57 % y 79.5 % de variabilidad de los datos respectivamente.

	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7	Comp 8	Comp 9
Varianza	3.62	1.51	1.04	0.97	0.82	0.50	0.24	0.21	0.04
P.V.E	40.2 %	16.8 %	11.6 %	10.8 %	9.1 %	5.6 %	2.7 %	2.3 %	0.5 %
P.V.E.A	40.2 %	57.1 %	68.7 %	79.5 %	88.7 %	94.3 %	97.0 %	99.4 %	100 %

TABLA 3.3. Varianza de las componentes.

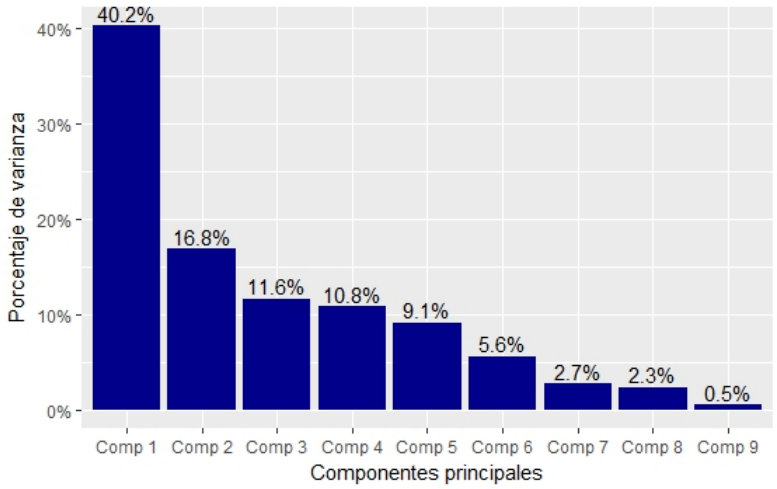


FIGURA 3.1. Porcentaje de varianza explicada por cada componente principal.

La proyección de los datos estandarizados (con medida de tendencia central la media y medida de dispersión la desviación estándar) en las dos primeras componentes principales se muestran en la siguiente gráfica.

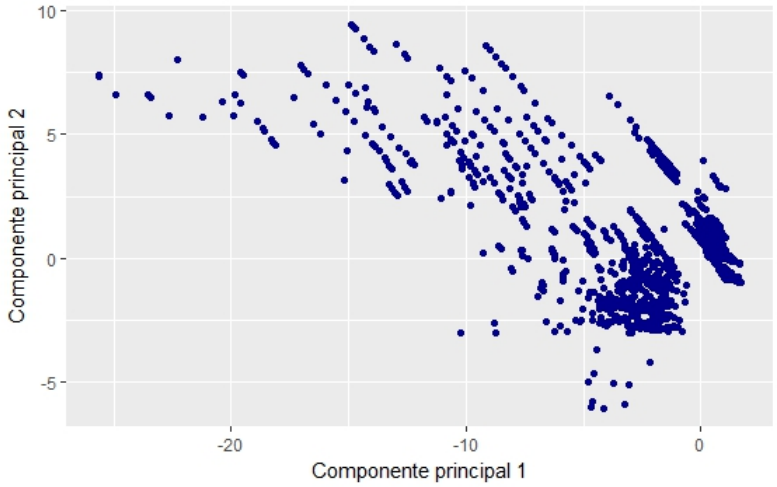


FIGURA 3.2. Proyección de los datos en las dos primeras componentes principales.

3. Agrupaciones de las palabras claves

Los algoritmos usados para el análisis de agrupamiento fueron:

- AGNES,
- K-medias,
- PAM,

siendo este último el que obtuvo un mejor desempeño. Las variables usadas para las agrupaciones fueron las que más información aportaban (los pesos en valor absoluto más altos) en las primeras dos y cuatro componentes principales respectivamente, éstas son:

- Clics,
- Tasas de clics,

para las dos primeras componentes principales y para las primeras cuatro componentes:

- Clics,
- Tasa de clics,
- Posición promedio,
- Conversiones.

En esta sección presentaremos los resultados obtenidos por los agrupamientos hechos mediante el algoritmo de K-medoids ó PAM.

3.1. Agrupación 1: Clics, Tasa de clics.

La agrupación se realizó con los datos estandarizados con medida de tendencia central la media y medida de dispersión la desviación estándar, la medida de distancia usada fue la distancia Euclídea.

- **Determinación de los K grupos**

Se usó el método del codo (*Elbow method*) para determinar el número óptimo de clusters, veamos esto en la siguiente gráfica.

FIGURA 3.3. Método del codo para la agrupación 1.

En la Figura 3.3 observamos que se recomienda usar entre tres o cuatro grupos, para esta agrupación se consideró $K = 4$.

La distribución de las palabras claves en los 4 grupos formados por el algoritmo PAM quedó de la siguiente manera.

Grupo 1	Grupo 2	Grupo 3	Grupo 4
13.617	1.054	157	1.478

TABLA 3.4. Distribución de las palabras claves.

La segmentación de las variables clics y tasas de clics con respecto a los cuatro grupos formados la observamos en la siguiente gráfica.

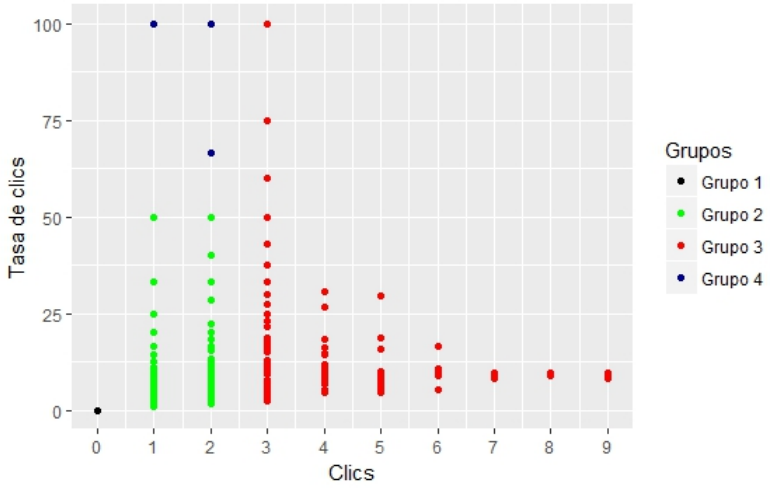


FIGURA 3.4. Segmentación de las variables clics y tasas de clics en los 4 grupos.

La proyección de los datos en las dos primeras componentes principales segmentadas por estos cuatro grupos la observamos en la Figura 3.5.

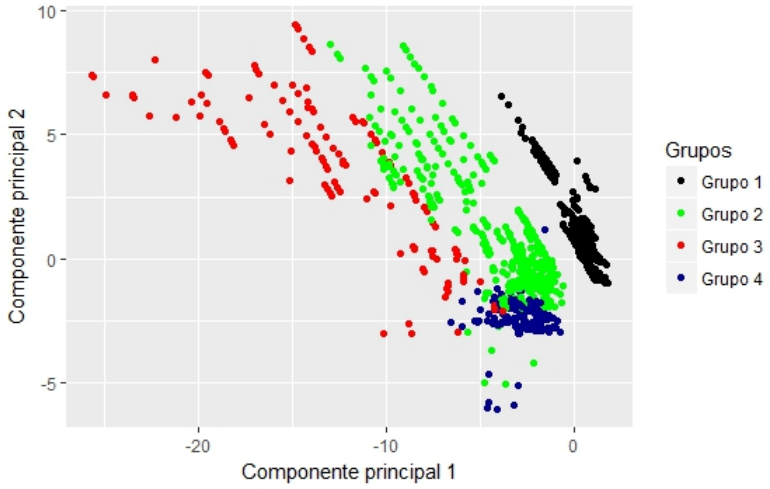


FIGURA 3.5. Proyección de los datos originales segmentados por los cuatro grupos.

Tanto en la Figura 3.4 como en la Figura 3.5 observamos que existe una buena segmentación de los datos con respecto a los cuatro grupos. Veamos las características de las palabras claves en los cuatro grupos formados.

- **Grupo 1**

Este grupo está formado por 13.617 palabras, en la siguiente tabla se muestra un resumen de las variables númericas de estos datos.

	Oferta maxima	Impresiones	Clics	Tasa de clics	Costo total	Costo promedio	Posicion promedio	Calidad anuncio	Conversiones
Min	0.13	1	0	0	0	0	1	0	0
Mediana	1	1	0	0	0	0	1	8	0
Media	1.07	1.75	0	0	0	0	1.43	8.69	0
Max	5	60	0	0	0	0	7	10	0

TABLA 3.5. Resumen del grupo 1.

Observamos en la Tabla 3.5 que la característica principal de este grupo de palabras es que ninguna obtuvo un clic, por ende estas palabras no generaron costos ni conversiones. Otra característica de este grupo es que la mayoría de estas palabras solo tuvieron una impresión, específicamente de las 13.617 palabras 9.533 tuvieron una impresión, esto representa el 70 % de los datos en este grupo.

- **Grupo 2**

Este grupo consta de 1.054 palabras, veamos el siguiente resumen para este grupo de palabras.

	Oferta maxima	Impresiones	Clics	Tasa de clics	Costo total	Costo promedio	Posicion promedio	Calidad anuncio	Conversiones
Min	0.34	2	1	0.92	0.01	0.01	1	5	0
Mediana	1.14	4	1	33.33	0.92	0.88	1	8	0
Media	1.56	10.02	1.1	29.19	1.05	0.92	1.31	8.76	0.02
Max	5	117	2	50	4.46	2.54	3.5	10	1

TABLA 3.6. Resumen del grupo 2.

En la Tabla 3.6 notamos que las 1.054 palabras obtuvieron al menos un clic y se mostraron por lo menos dos veces esto a diferencia del grupo 1 que no obtuvieron clics y que el 70 % de las palabras se mostraron una vez. También observamos que la tasa de clics está comprendida entre un 0.92 % hasta 50 %, es decir, las palabras en este grupo obtuvieron más impresiones

que clics. Otra característica importante para este grupo es que se obtuvieron conversiones, específicamente 22 palabras de las 1.054 en total.

- **Grupo 3**

Para el grupo 3 se tienen 157 palabras, el resumen está dado por la Tabla 3.7.

	Oferta maxima	Impresiones	Clics	Tasa de clics	Costo total	Costo promedio	Posicion promedio	Calidad anuncio	Conversiones
Min	0.59	3	3	2.27	0.42	0.14	1	7	0
Mediana	5	34	3	10.81	4.4	1.39	1	8	0
Media	3.41	43.41	3.99	21.55	5.43	1.32	1.13	8.03	0.01
Max	5	132	9	100	13.94	2.38	2.3	10	1

Tabla 3.7. Resumen del grupo 3.

Para este conjunto de palabras, observamos que todas recibieron al menos tres clics y tres impresiones, este grupo representa las palabras que más obtuvieron clics en todo el conjunto de datos, sin embargo la tasa de clics tiene un promedio de 21.5 %, esto significa que cada palabra obtuvo una mayor cantidad de impresiones que de clics. Otra característica a resaltar son las conversiones, de las 157 palabras que obtuvieron la mayor cantidad de clics solamente dos convirtieron.

- **Grupo 4**

Para el último grupo se obtuvo un total de 1.478 palabras, veamos el resumen para estos datos.

	Oferta maxima	Impresiones	Clics	Tasa de clics	Costo total	Costo promedio	Posicion promedio	Calidad anuncio	Conversiones
Min	0.13	1	1	66.67	0.01	0.01	1	0	0
Mediana	1	1	1	100	0.66	0.64	1	10	0
Media	1.12	1.05	1.05	99.8	0.73	0.67	1.17	9.33	0.04
Max	2.8	3	2	100	3.87	2.16	4	10	1

Tabla 3.8. Resumen del grupo 4.

En la Tabla 3.8 observamos que la principal características de este grupo son las tasas de clics con un promedio del 99.8 %, esto significa que la cantidad de clics que obtuvo una palabra tiene practicamente la misma cantidad de impresiones, mostrando así que este conjunto de palabras poseen una gran efectividad entre estas dos variables. Otras características importantes son que estos datos obtienen en promedio la mejor calidad de anuncio y a su vez tienen el resto de las conversiones, específicamente 57 palabras convirtieron.

Podemos notar que a diferencia del grupo 3 que obtuvo las palabras con mayores clics e impresiones, este grupo obtuvo un mejor rendimiento con tan solo un promedio de clics e impresiones de uno.

- **Validación**

Se usó el método de la silueta (Silhouette method) para conocer la calidad del agrupamiento formado, el valor $\bar{S}$ se ilustra en la siguiente gráfica.

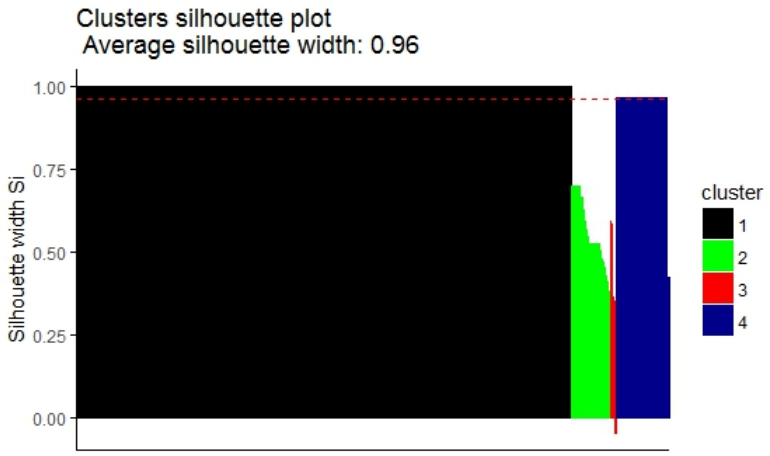


FIGURA 3.6. Promedio de los índices de silueta.

En la Figura 3.6 nos muestra que el valor $\bar{S}$ es de 0.96, según lo expuesto en la Tabla 2.1 este valor indica una fuerte estructura en la agrupación, es decir, que cada índice de silueta de cada observación obtuvo en promedio un valor de 0.96, lo que significa que cada registro se agrupó de manera correcta en su respectivo cluster.

3.2. Agrupación 2: Clics, Tasa de clics, Posición promedio, Conversiones.

Al igual que la Agrupación 1, se usó como medida de tendencia central la media, medida de dispersión la desviación estándar y medida de disimilitud la distancia Euclídea.

- **Determinación de los K grupos**

Análogamente a la Agrupación 1, se realizó el método del codo para obtener el número óptimo de grupos.

FIGURA 3.7. Método del codo para la agrupación 2.

En la Figura 3.7 observamos que se recomienda usar entre cinco o seis grupos, para esta agrupación se consideró $K = 6$.

La distribución de las palabras claves en los 6 grupos formados quedó de la siguiente manera.

Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6
9.600	1.112	2.660	1.420	1.433	81

TABLA 3.9. Distribución de las palabras claves de la agrupación 2.

Para visualizar la segmentación de las cuatro variables con respecto a los seis grupos formados, proyectamos estos datos en sus dos primeras componentes principales obteniendo la siguiente gráfica.

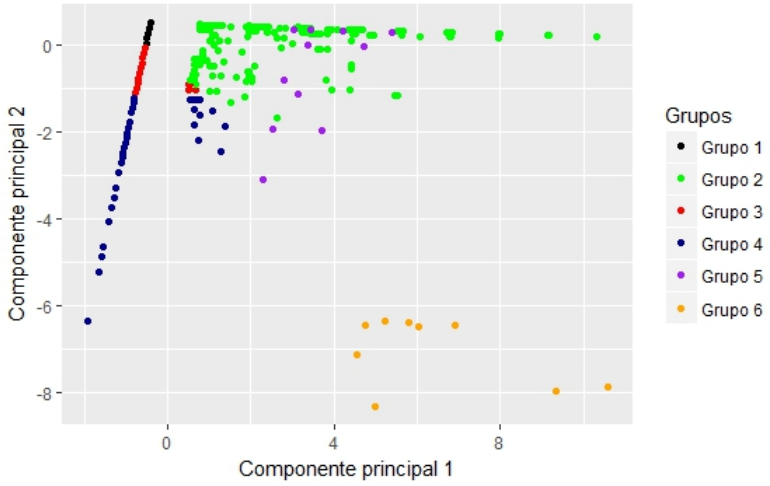


FIGURA 3.8. Segmentación de las cuatro variables en sus dos primeras componentes principales por los 6 grupos.

La proyección de los datos originales segmentadas por estos seis grupos la observamos en la Figura 3.9.

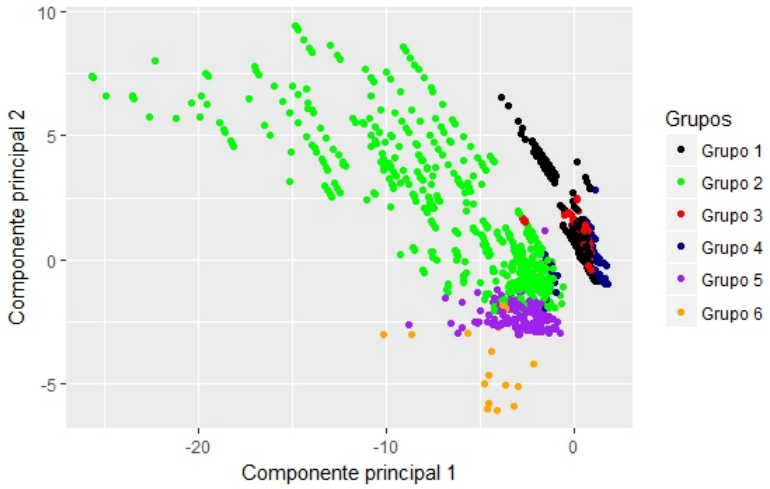


FIGURA 3.9. Proyección de los datos originales segmentados por los seis grupos.

Podemos observar una buena agrupación en la Figura 3.8, en cambio para la Figura 3.9 notamos un solapamiento o confusión entre los grupos uno, tres y cuatro. Veamos en detalle las características de cada grupo.

- **Grupo 1**

Este grupo está formado por 9.600 palabras, veamos un resumen de las variables.

	Oferta maxima	Impresiones	Clics	Tasa de clics	Costo total	Costo promedio	Posicion promedio	Calidad anuncio	Conversiones
Min	0.13	1	0	0	0	0	1	0	0
Mediana	1	1	0	0	0	0	1	8	0
Media	1.14	1.68	0	0	0	0	1.01	8.74	0
Max	5	60	0	0	0	0	1.4	10	0

TABLA 3.10. Resumen del grupo 1.

Observamos en la Tabla 3.10 que la características principal de este grupo es que las palabras no recibieron clics pero mantuvieron una posición promedio de entre 1 y 1.4, es decir, este grupo representa a las palabras mejor posicionadas que no obtuvieron clics.

- **Grupo 2**

El grupo 2 consta de 1.112 palabras, el resumen viene dado por la siguiente tabla.

	Oferta maxima	Impresiones	Clics	Tasa de clics	Costo total	Costo promedio	Posicion promedio	Calidad anuncio	Conversiones
Min	0.34	2	1	0.92	0.01	0.01	1	5	0
Mediana	1.31	4	1	25	1.08	0.98	1	8	0
Media	1.89	14.98	1.49	27.46	1.67	1	1.2	8.64	0
Max	5	132	9	60	13.94	2.54	2.8	10	0

TABLA 3.11. Resumen del grupo 2.

A diferencia del grupo 1, este grupo de palabras obtuvo al menos un clic. La principal característica de este grupo son las tasas de clics que se encuentra en un rango de 0.92 % hasta un 60 %, lo que implica que las palabras obtuvieron más impresiones que clics.

- **Grupo 3**

Para este grupo, se obtuvieron 2.660 observaciones. La Tabla 3.12 nos muestra el resumen de éstas.

	Oferta maxima	Impresiones	Clics	Tasa de clics	Costo total	Costo promedio	Posicion promedio	Calidad anuncio	Conversiones
Min	0.25	1	0	0	0	0	1.5	5	0
Mediana	1	1	0	0	0	0	2	8	0
Media	1.02	2.06	0.002	0.01	0.001	0.001	1.93	8.51	0
Max	2.44	37	1	11.11	0.63	0.63	2.4	10	0

TABLA 3.12. Resumen del grupo 3.

Para este conjunto de palabras, solamente 7 recibieron clics y sus tasas de clics estuvieron en un rango menor al 60 %, esto sugiere que estas palabras deberían pertenecer al grupo 2. Sin embargo, la variable que más influye en este grupo es la posición promedio que tiene un rango de 1.5 a 2.4, por lo que estas palabras se ven más influenciadas por esta característica.

- **Grupo 4**

Este grupo consta de 1.420 palabras, observemos el resumen de las variables.

	Oferta maxima	Impresiones	Clics	Tasa de clics	Costo total	Costo promedio	Posicion promedio	Calidad anuncio	Conversiones
Min	0.24	1	0	0	0	0	2.5	0	0
Mediana	0.63	1	0	0	0	0	3	8	0
Media	0.7	1.93	0.04	1.31	0.02	0.02	3.37	8.68	0
Max	2.44	25	1	50	0.63	0.63	7	10	0

TABLA 3.13. Resumen del grupo 4.

En la Tabla 3.13 observamos que la posición promedio de estos datos tiene un rango de 2.5 a 7 y una media de clics de casi cero, por lo que este grupo representa a las palabras que no obtuvieron clics con las más bajas posiciones. Sin embargo al igual que el grupo 3, en este grupo se tienen 56 palabras que si obtuvieron clics y una tasa de clics menor al 60 %, lo que sugiere una incorrecta agrupación para estas 56 observaciones.

- **Grupo 5**

El grupo 5 cuenta con 1.433 observaciones. En la siguiente tabla podemos observar el resumen de las variables.

	Oferta maxima	Impresiones	Clics	Tasa de clics	Costo total	Costo promedio	Posicion promedio	Calidad anuncio	Conversiones
Min	0.13	1	1	66.67	0.01	0.01	1	0	0
Mediana	1	1	1	100	0.65	0.63	1	10	0
Media	1.12	1.07	1.06	99.77	0.75	0.66	1.17	9.33	0
Max	2.8	4	3	100	5.6	2.16	4	10	0

TABLA 3.14. Resumen del grupo 5.

La característica más importante en este grupo son las tasas de clics, con un mínimo que supera el 60 % y un promedio del 99.7 %, lo que implica una gran efectividad de este conjunto de palabras con respecto a las impresiones y clics.

- **Grupo 6**

Para el último grupo se obtuvieron un total de 81 observaciones. Veamos el resumen para estos datos.

	Oferta maxima	Impresiones	Clics	Tasa de clics	Costo total	Costo promedio	Posicion promedio	Calidad anuncio	Conversiones
Min	0.44	1	1	3.33	0.18	0.18	1	7	1
Mediana	1.2	1	1	100	0.93	0.93	1	10	1
Media	1.18	4.31	1.12	77.07	0.9	0.83	1.2	9.11	1
Max	2	36	6	100	3.32	1.4	2.7	10	1

TABLA 3.15. Resumen del grupo 6.

Podemos observar en la Tabla 3.15 que la principal característica de estas 81 palabras es que todas obtuvieron una conversión, por lo que de las 16.306 palabras del conjunto de datos solamente 81 palabras tuvieron conversiones.

- **Validación**

Para la validación de esta agrupación, se usó el método de la silueta al igual que la agrupación 1, el valor $\bar{S}$ se muestra en la siguiente gráfica.

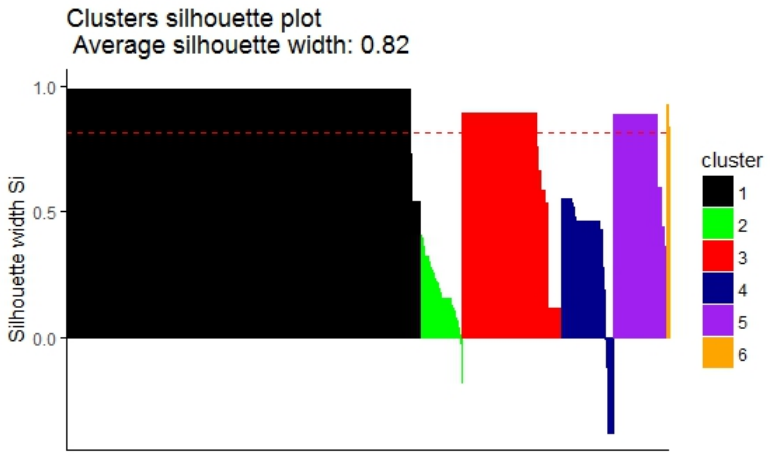


FIGURA 3.10. Promedio de los índices de silueta.

En la Figura 3.10 observamos que el valor $\bar{S}$ es de 0.82, la Tabla 2.1 nos indica que el valor obtenido de $\bar{S}$ representa una fuerte estructura en la agrupación. Podemos observar que para el grupo 4, existen observaciones con índices de silueta con valor negativo, esto se debe a las 56 observaciones que se confunden y por ende quedaron mal clasificadas.

A diferencia de la Agrupación 1 la cual presenta un valor $\bar{S}$ más alto y por lo tanto una estructura más fuerte que la Agrupación 2, ésta última involucra una mayor cantidad de variables lo cual permite dar una mejor explicación del comportamiento de los datos originales.

Hay que destacar que cada agrupación ofrece de forma individual caracterizaciones muy ricas del conjunto de datos y es recomendable detallar ambas estructuras pues una complementa a la otra, generando así un conocimiento más profundo que nos pueda ser de utilidad a la hora de tomar decisiones.

4. Modelos de regresión

Con la intención de dar una caracterización un poco más fuerte a los grupos de cada agrupación se buscaron patrones de dependencia o estructuras entre las variables con la finalidad de establecer modelos de regresión que pudieran servir para predecir valores futuros para determinadas variables de cada grupo. En esta sección mostraremos los modelos más significativos que se obtuvieron.

4.1. Agrupación 1.

- **Grupo 1**

Para el grupo 1 de ésta agrupación no se obtuvieron clics, por lo que se realizó el modelo de regresión comparando el total de las impresiones por cada posición promedio dada, el gráfico de esta comparación es el siguiente.

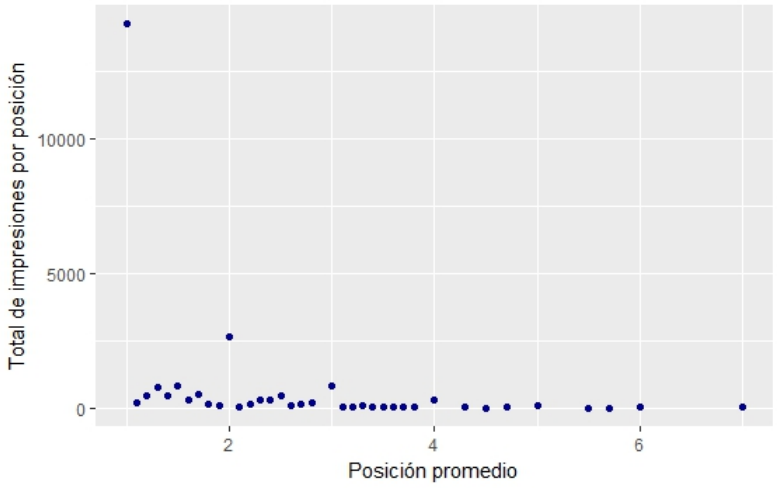


FIGURA 3.11. Total de impresiones por posición.

Para la gráfica mostrada en la Figura 3.11 se ajustó un modelo de tipo exponencial obteniendo un R^2 de 0.83, con una desviación estándar de los errores de 968.1.

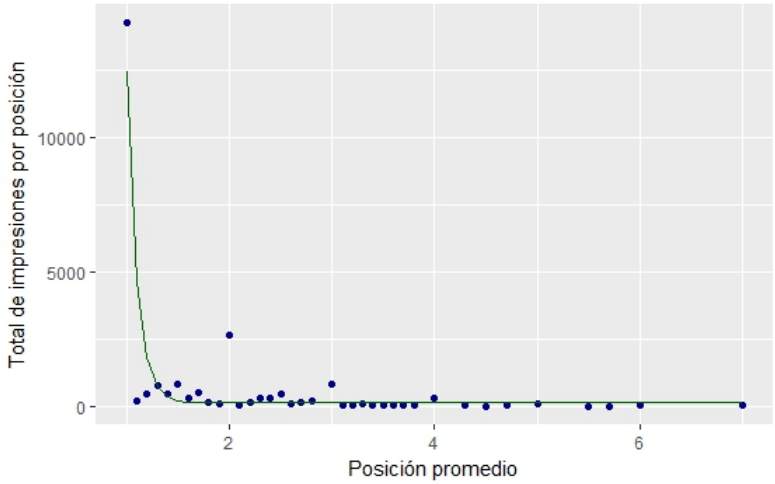


FIGURA 3.12. Modelo ajustado.

Notemos en la Figura 3.11 que el valor más alto que alcanza ésta gráfica corresponde a más de 14.000 impresiones en total para las palabras que tienen una posición promedio de 1. La cantidad de impresiones en este grupo fue de 23.903, por lo que las palabras con posición promedio de 1 obtienen el 60 % del total de las impresiones de este grupo.

- **Grupo 3**

El grupo 3 representa a las palabras que más recibieron clics, sin embargo, el rendimiento de estas palabras no fue el mejor pues obtuvieron en promedio una baja tasa de clics y ninguna conversión. Para este modelo de regresión comparamos las impresiones de las palabras con el promedio de tasas de clics, es decir, por cada impresión cual fue su promedio de tasas de clics.

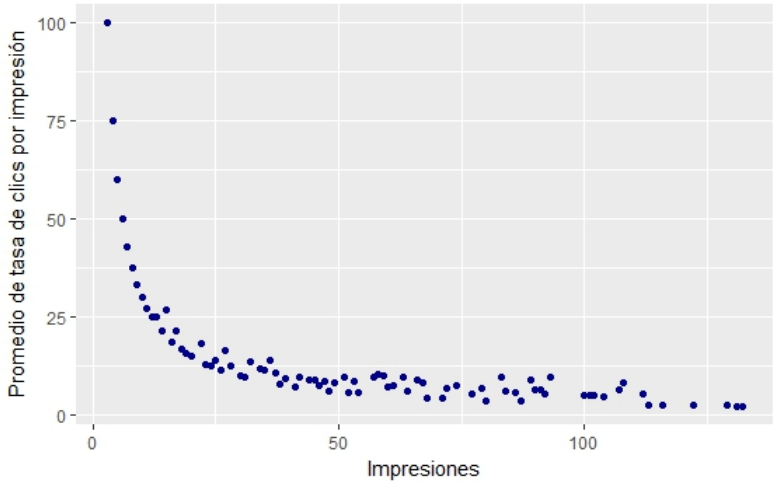


FIGURA 3.13. Promedio tasas de clics por impresión.

Al comportamiento mostrado en la Figura 3.13 se le ajustó un modelo del tipo $\frac{1}{x}$ cuyo R^2 obtuvo un valor de 0.98 y una desviación estándar de los residuales de 1.87.

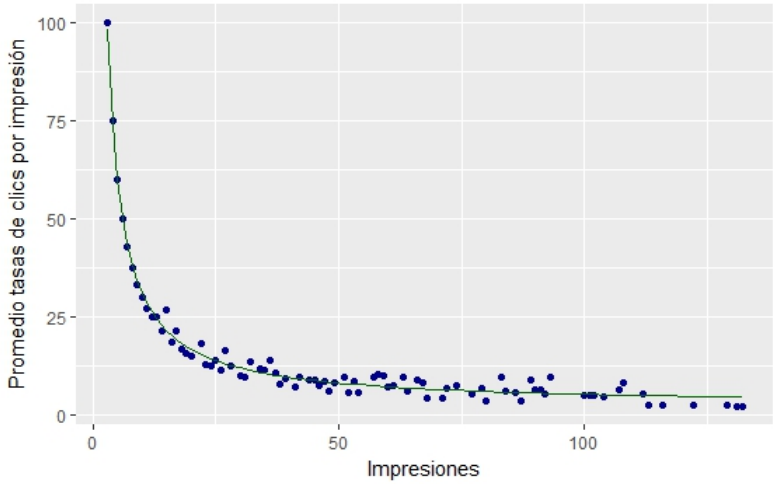


FIGURA 3.14. Modelo ajustado.

Podemos notar en la Figura 3.13 que a menor cantidad de impresiones la tasa de clics promedio incrementa y a mayor cantidad de impresiones la tasa promedio de clics disminuye.

4.2. Agrupación 2.

- **Grupo 2**

La principal característica de este grupo es la tasa de clics la cual no supera el 60 %, también este grupo obtiene en promedio el mayor costo de todos los grupos formados en la Agrupación 2. Para este modelo de regresión comparamos la cantidad total de costo por cada posición promedio dada, el gráfico es el siguiente.

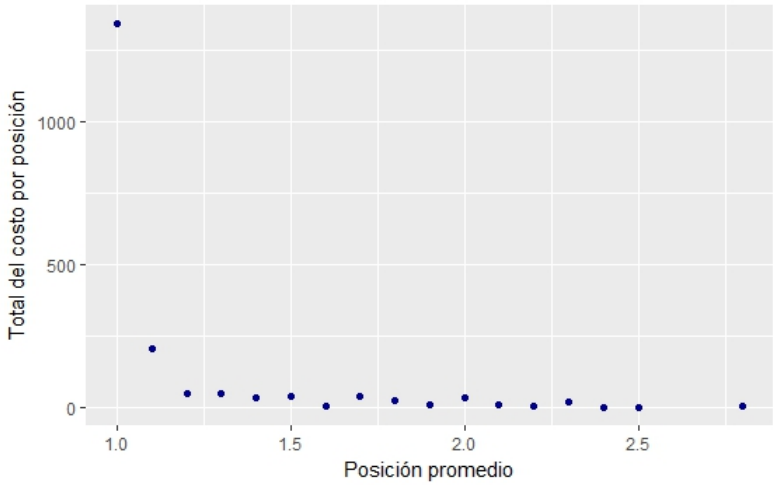


FIGURA 3.15. Total de costo por posición.

Se ajustó un modelo de tipo exponencial obteniendo un R^2 de 0.94, con una desviación estándar de los errores de 77.93.

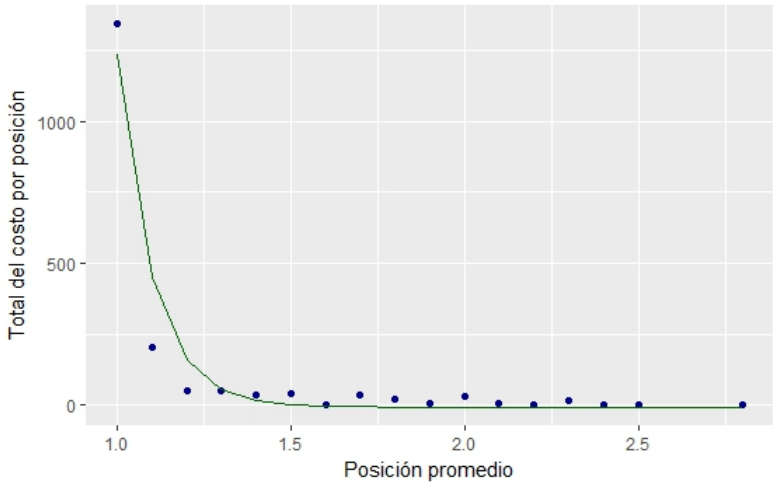


FIGURA 3.16. Modelo ajustado.

- **Grupo 5**

Este grupo está formado por las palabras que obtuvieron una tasa de clic mayor al 60 %. Para este grupo se ajustó un modelo comparando la oferta máxima por clic de cada palabra y su media del promedio de costo por clic. La gráfica es la siguiente.

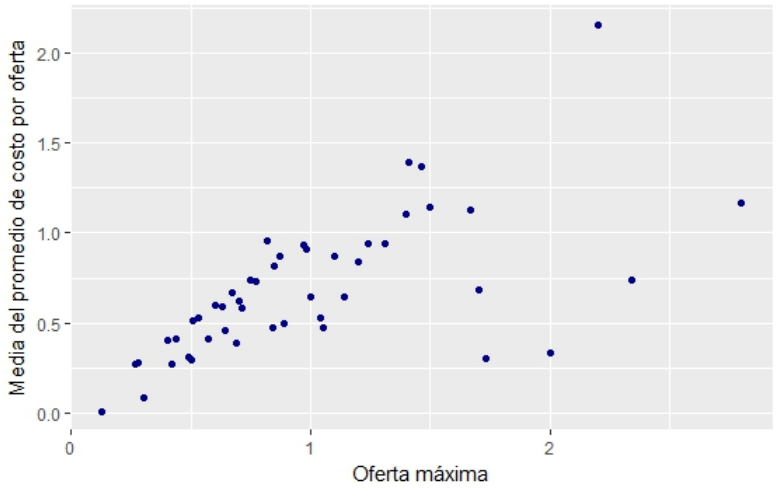


FIGURA 3.17. Media del promedio de costo por oferta.

Se ajustó un modelo del tipo logarítmico, resultando un R^2 de 0.47 y una desviación estándar de los residuos de 0.28.

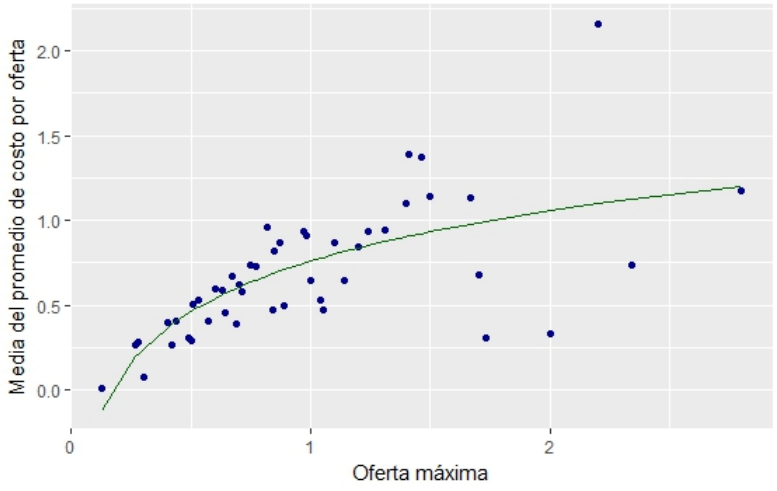


FIGURA 3.18. Modelo ajustado.

En general encontramos estructuras y patrones muy interesante entre las variables del conjunto de datos, entre las cuales destacan las variables impresiones y tasas de clics, para

estas dos variables observamos un comportamiento de tipo exponencial o tipo cociente. También observamos que las palabras con posición promedio exactamente igual a 1 tienen un mejor rendimiento que las demás.

Conclusiones

Para las dos agrupaciones hechas pudimos notar una buena segmentación de las palabras y una calidad óptima de agrupación lo cual se corroboró con el método de la silueta, obteniendo características muy marcadas e importantes en cada grupo formado, permitiendo observar grupos de palabras que tuvieron un mejor rendimiento que otras, obteniendo grupos en los cuales habían clics y no conversiones, grupos sin clics, grupos con solo las palabras que convirtieron, entre otros. La aplicación del análisis de componentes principales fue de gran utilidad, pues gracias a este se consiguió reducir la dimensionalidad del espacio de variables con lo cual fue posible caracterizar y clasificar al conjunto de palabras. Otro resultado a resaltar son los modelos de regresiones, donde se encontraron buenos ajustes para el comportamiento de las variables en los diferentes grupos.

Bibliografía

[1] B. Everitt, T. Hothorn (2011). An Introduction to Applied Multivariate Analysis with R.

[2] G. Dunteman (1989). Principal Components Analysis.

[3] A. Kassambara (2017). Practical Guide To Cluster Analysis in R.

[4] I.T Jolliffe (2002). Principal Component Analysis, second edition.

[5] L. Kaufman, P. Rousseeuw (1990). Finding Groups in Data: An introduction to cluster Analysis

[6] Cluster Analysis https://www.stat.berkeley.edu/~s133/Cluster2a.html

[7] H. Nuñez. Talle Minería de datos UCV.

[8] Google AdWords. https://support.google.com/adwords/answer/6319?hl=es-419