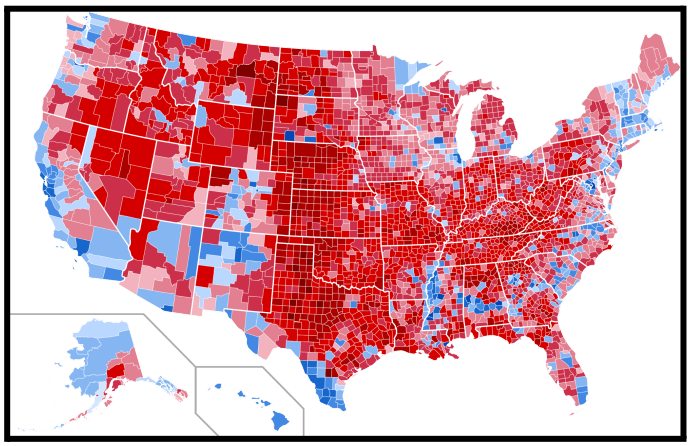
Machine Learning III: **TRABAJO FINAL**

Análisis político de los segmentos de votantes estadounidenses



Luís Mielgo Larriba

Francisco Javier Vázquez Visos

INTRODUCCIÓN

Para el caso que se busca examinar en el presente informe nos hemos puesto en los zapatos de un analista de datos enfocado en la política estadounidense. Básicamente, tenemos la tarea de realizar un análisis de clustering sobre como es el electorado político americano en función de diversas variables. De esta forma, podremos utilizar este análisis para poder entender mejor los distintos segmentos de la sociedad estadounidense y saber qué discursos tenemos que hacer para conseguir votos en cada segmento de votantes.

DATOS: FUENTES DE DATOS, DESCRIPCIÓN, PREPROCESAMIENTO

Los datos que estamos utilizando los hemos obtenido de la página web del censo de EE.UU., donde se registran las estadísticas para cada uno de los más de tres mil condados que hay en el país. Este dataset lo hemos juntado con los resultados de las elecciones de 2016 para esos mismos condados.

El dataset cuenta con un total de 3109 observaciones y 15 variables, las cuales explicaremos a continuación:

NOMBRE DE LA VARIABLE	TIPO	DESCRIPCIÓN
County Name	chr	Nombre de cada uno de los condados.
State Abbreviation	chr	Abreviación del estado al que pertenecen.
Fips	num	Identificador de cada condado.
Pctg. DEM	num	Porcentaje de voto demócrata.
Pctg. GOP	num	Porcentaje de voto republicano.
Votes DEM	num	Cantidad de votos demócratas.
Votes GOP	num	Cantidad de votos republicanos.
Black Pctg.	num	Porcentaje de población afroamericana.
Hispanic or Latino Pctg.	num	Porcentaje de población hispana o latina.
White alone Pctg.	num	Porcentaje de población caucásica.
Foreign born persons, pctg.	num	Porcentaje de extranjeros.
Bachelor's degree or higher, percent of persons age 25+	num	Porcentaje de ciudadanos mayores de 25 que poseen un título de bachillerato o superior.
Median house income	num	Mediana del ingreso familiar.
GDP per capita	num	PIB per cápita.
Persons below poverty level, pctg	num	Porcentaje de ciudadanos pobres.

Si ejecutamos el comando **`head(datos)`** podremos obtener una breve idea de que aspecto tiene el dataset:

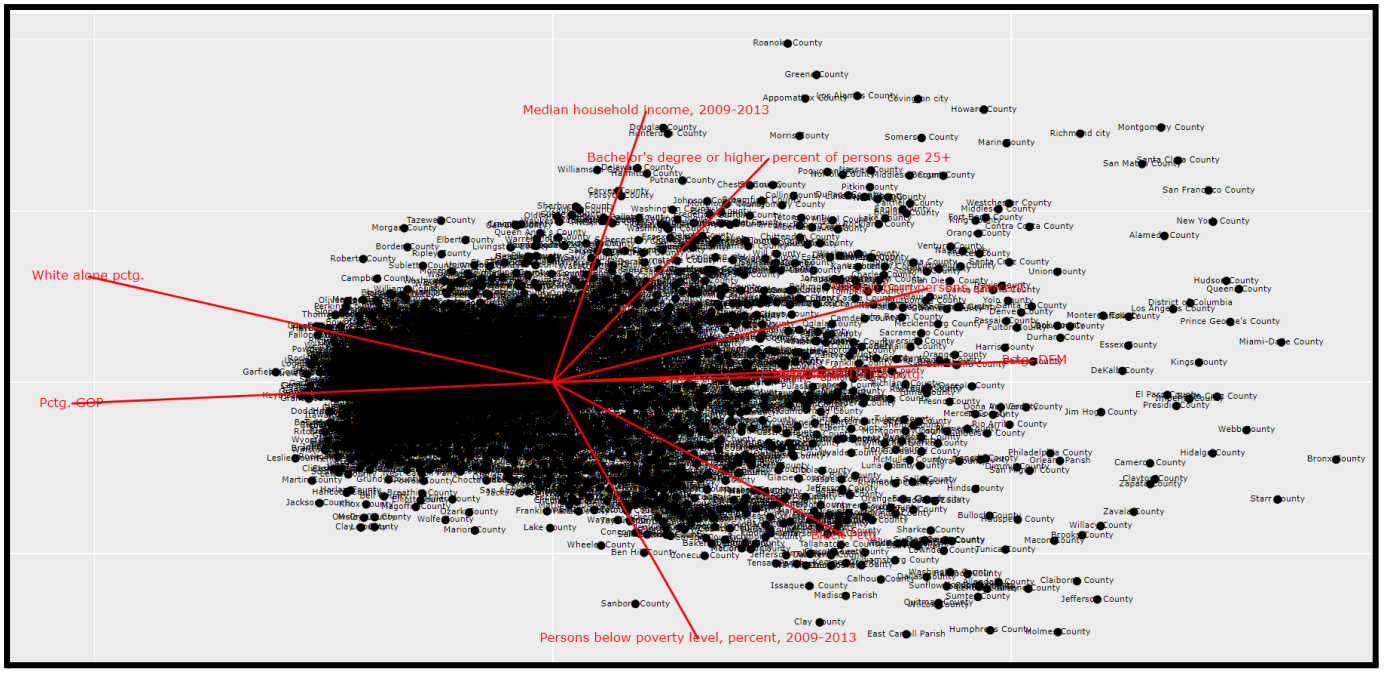
	`County Name`	`State Abreviatio~`	Fips	`Pctg. DEM`	`Pctg. GOP`	`Black Pctg.`	`Hispanic or Latino p~`
	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	Autauga County	AL	1001	0.246	0.754	18.7	2.7
2	Baldwin County	AL	1003	0.202	0.798	9.6	4.6
3	Barbour County	AL	1005	0.472	0.528	47.6	4.5
4	Bibb County	AL	1007	0.218	0.782	22.1	2.1
5	Blount County	AL	1009	0.0861	0.914	1.8	8.7
6	Bullock County	AL	1011	0.756	0.244	70.1	7.5

No hemos encontrado ningún valor NA.

De todas estas 14 variables hemos decido utilizar únicamente 8. Hemos excluido Votes DEM y Votes GOP porque en su lugar utilizaremos el porcentaje, puesto que a fin de cuentas dan la misma información. También hemos excluido el GDP per capita puesto que para información económica también tenemos la mediana de los ingresos familiares, escogiendo la segunda debido a que ciertos condados, como puede ser en las zonas de Los Ángeles o Nueva York, tienen un PIB per cápita desorbitado por unas unidades concretas de usuarios, no dando una imagen fiel a la realidad. Finalmente, y como es obvio, no se incluirá ni County Name, ni State Abbreviation ni Fips ya que no aportan información real en el análisis.

ANÁLISIS CLUSTER Y RESULTADOS E INTERPRETACIÓN

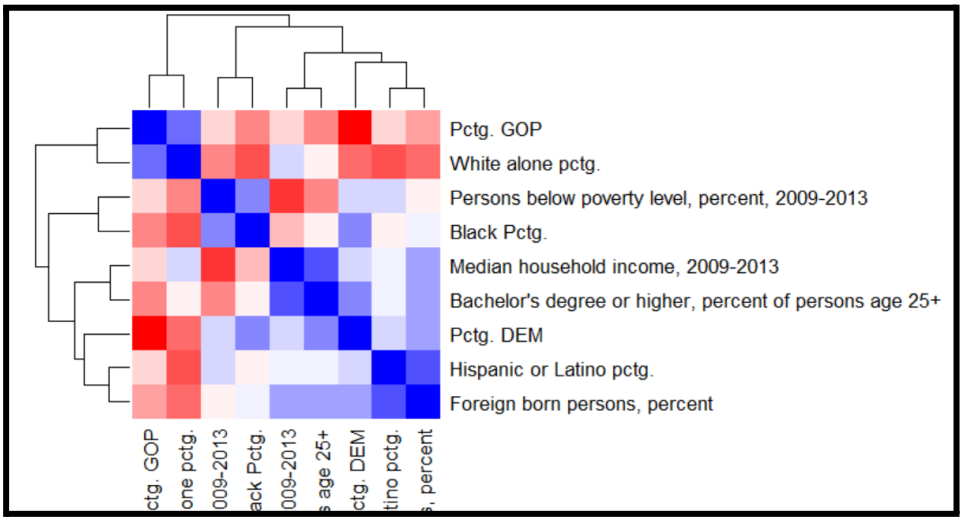
Comenzamos realizando un análisis de PCA sobre nuestro dataset. Para realizar esto, es necesario escalar nuestras variables ya que algunas son valores absolutos, como el income, mientras que otras están en porcentaje. Una vez hecho, podemos plotear el siguiente gráfico que nos muestra el análisis con las dos primeras componentes principales, que acumulan un total del 67% de la varianza acumulada:



Este Biplot de PCA nos deja obtener algunas conclusiones. Por ejemplo, los condados que se encuentran cerca de "White alone pctg." van a tener una población blanca mayor que aquellos que estén en la dirección opuesta, y lo mismo con el resto de variables.

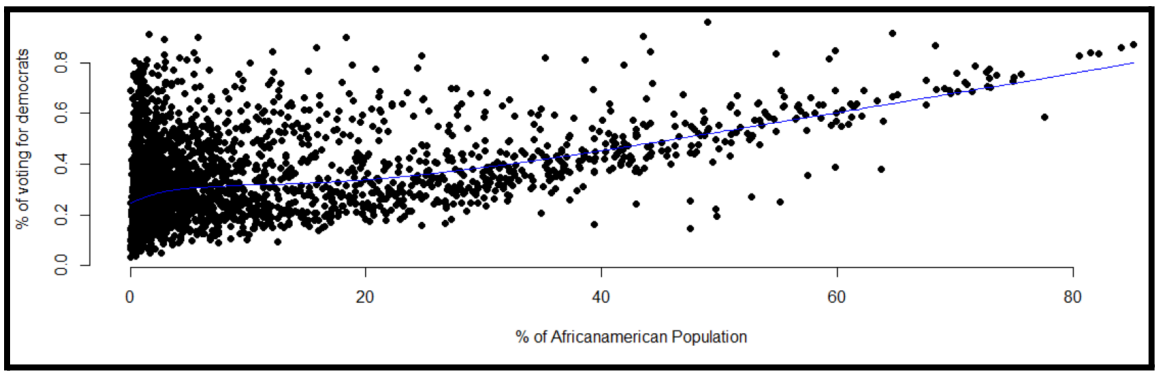
A simple vista podemos ver que las variables "White alone pctg." y "Pctg, GOP" se encuentran muy cerca, demostrando que hay una correlación entre los condados con población blanca y que votan al Partido Republicano (GOP). Por otra parte, las variables "Persons below poverty rate" y "Black Pctg." también se encuentran cerca, mostrando una correlación entre los condados más pobres y de mayoría afroamericana.

Estas correlaciones se nos confirman si hacemos un heatmap:



Como se puede observar, la correlación más negativa es entre los votos demócratas y los republicanos, lo cual resulta obvio al ser excluyentes entre ellas. Tras esto, se pueden observar muchísimas más relaciones. Por ejemplo, los votos republicanos solo tienen una correlación positiva con el porcentaje de habitantes blancos. Es decir, cuanto más blanco sea el condado, más probable es que voten republicano, tal y como indicaba el biplot anterior.

En cuanto a los demócratas, la correlación es positiva con el resto de variables, siendo la más fuerte la que relaciona el voto de izquierdas con un nivel de estudios mayor (Bachelor's degree or higher) y con el porcentaje de población afroamericana. En otros aspectos como puede ser la mediana de ingresos o la población latina, la correlación es positiva pero no tanto, significando que los republicanos tienen una buena performance aquí aunque no lleguen a ganar. A continuación tenemos un scatterplot que muestra claramente la relación entre población afroamericana y voto demócrata:

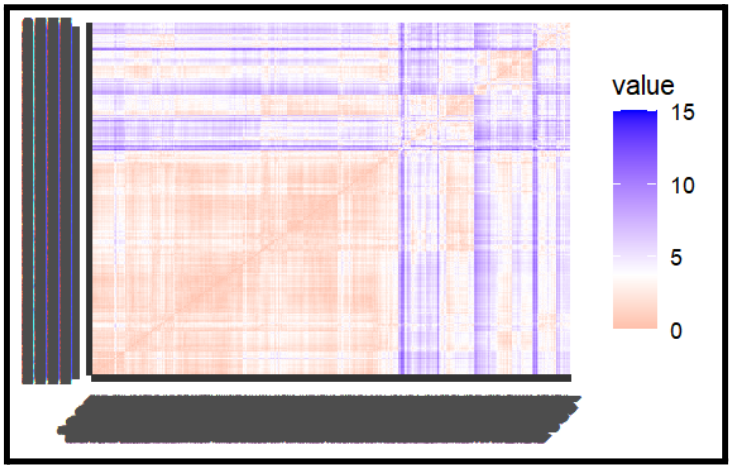


Alejándonos de la óptica política, hay otras correlaciones que nos vislumbra este mapa. Hay una correlación positiva entre el índice de pobreza y el porcentaje de latinos, en menor medida, y de afroamericanos, en mayor medida. Dicha correlación es negativa si comparamos income con nivel de estudios, como resulta lógico.

Si echamos un ojo al median income, vemos que tiene una correlación positiva, a parte de con el nivel de estudios, con la población blanca y latina, aunque en menor medida, mientras que es negativa con la población afroamericana, mostrando que hay un problema de educación en la comunidad negra.

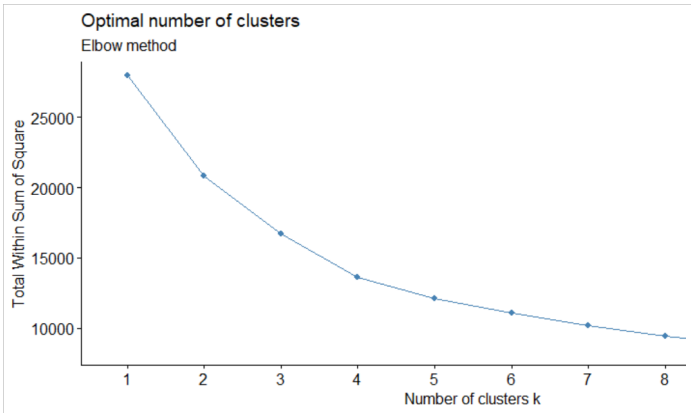
En cuanto al porcentaje de los nacidos fuera de EEUU, vemos que la correlación es más positiva con los latinos que con cualquier otro grupo, mostrando que son el grupo de inmigrantes mayoritario.

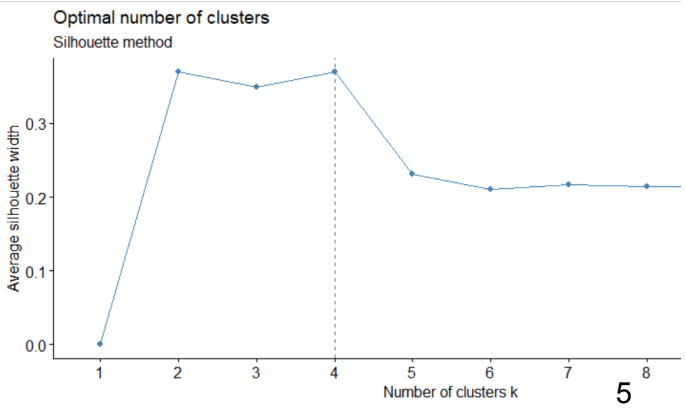
Pasando ya al análisis de clusterización como tal, comenzaremos calculando la matriz de distancias euclídeas:



Como podemos ver, el dataset contiene demasiados datos como para que se pueda interpretar de manera clara las distancias entre todas las observaciones. En total, hay más de 4 millones de elementos en nuestra matriz de distancias.

Tras esto, comenzaremos utilizando la técnica de K–Means en nuestro análisis. Para ello, primero de todo calcularemos cuantos clusters son los óptimos a la hora de realizar el mencionado análisis. Por ello, utilizaremos las dos técnicas que hemos aprendido en clase. La primera es la del “Elbow method”, o “método del codo”, y la segunda la de “Silhouette” o silueta. Obtenemos los siguientes gráficos:

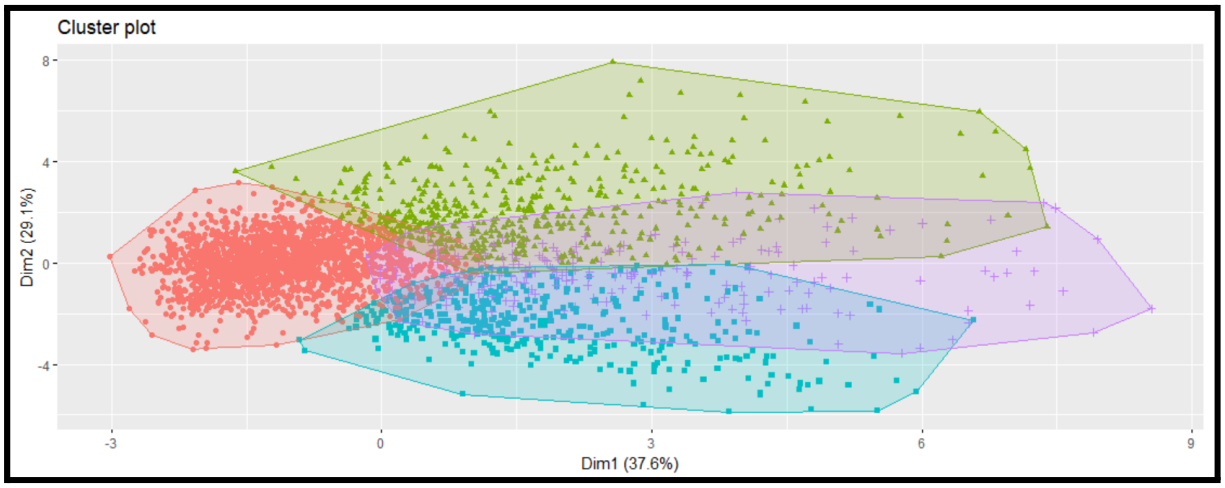


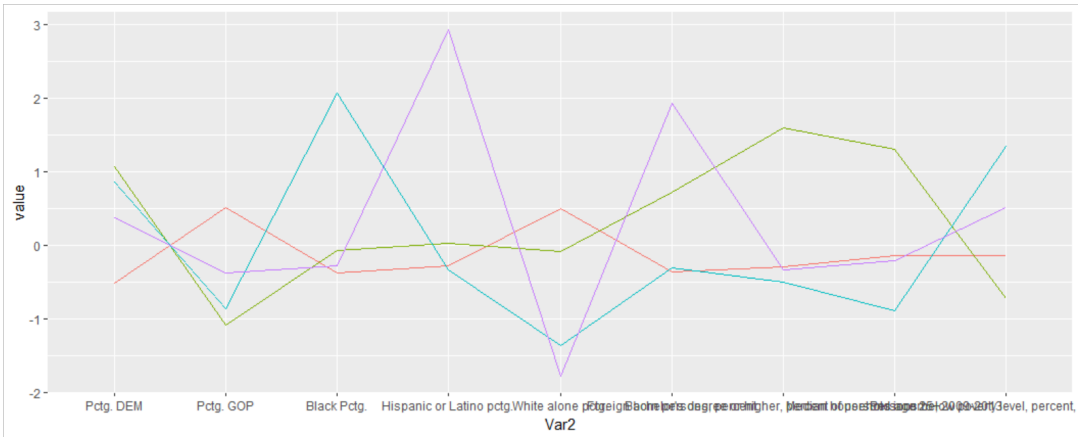


Como podemos ver, el método del codo, que muestra el valor de la Total Within Sum of Squares para cada número de clusters usado, no muestra un punto concreto que nos permita escoger cuantos clusters son los óptimos. Si bien es cierto que lo interesante es coger la cantidad de clusters que minimice la variación intra-clusters, no hay un número muy claro representado en este caso, por lo que nos guiaremos por el método de la silueta.

En el caso del método de la silueta, en esta ocasión lo interesante es maximizar la silhouette width, que en el gráfico de la derecha se puede ver que es máximo cuando el número de clusters escogido es igual a 4, por lo que tomaremos este número.

Tras ejecutar el comando para llevar a cabo un K–Means con un NSTART igual a 40 y K=4, obtenemos el siguiente tras usar la función **`fviz_cluster`** y calcular los centroides:





Para una mejor comprensión, hemos llegado a la conclusión de que la leyenda debería de ser así:

Rojo: tiende republicano. Blancos
Azul: tiende demócrata. Afroamericanos
Morado: tiende demócrata. Latinos
Verde: tiende demócrata. Blancos

Básicamente, el cluster **rojo** esta formado por aquellos condados de mayoría blanca que vota republicano. Es un grupo muy homogéneo y muy abundante. El cluster **azul** se encuentra formado por los condados de mayoría afroamericana que, además, tienen una renta baja. Su intención de voto también suele ser demócrata. Por otra parte, el cluster **morado** se encuentra formado por los condados con alto porcentaje de población latina, nacida en el extranjero y renta media, que también tiene intención de voto demócrata. Por último, el cluster **verde** está formado por habitantes de población blanca, renta alta y voto demócrata.

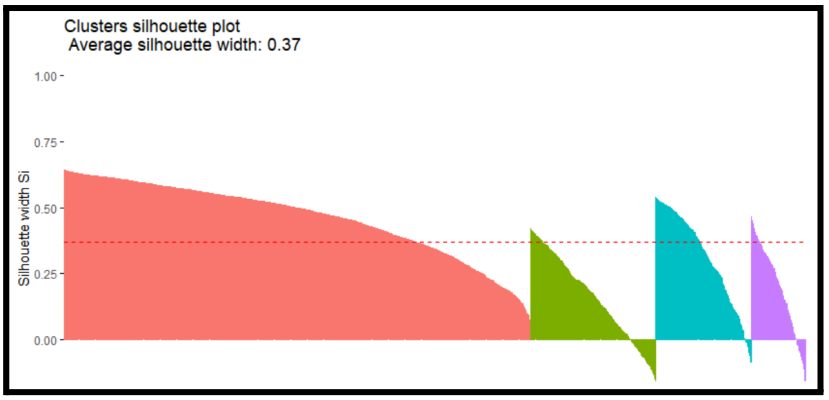
Si echamos un vistazo al tamaño de los clusters, vemos que el rojo (republicano) está formado por 1957 condados, el verde (demócratas blancos) por 525, el azul (demócratas afroamericanos) por 400 y, por último, el morado (demócratas latinos) por 227.

En lo referido a la within cluster variation, el total es 13618.19, coincidiendo con lo que se puede ver en el gráfico del elbow method. Por otra parte, el cluster rojo tendría una variación intra-cluster de 5395, la mayor de los 4 clusters.

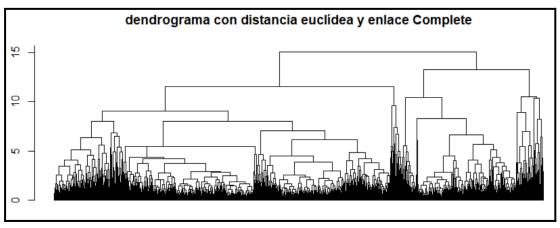
Hemos considerado interesante correr el código de nuevo pero esta vez estableciendo dos clusters en vez de cuatro:

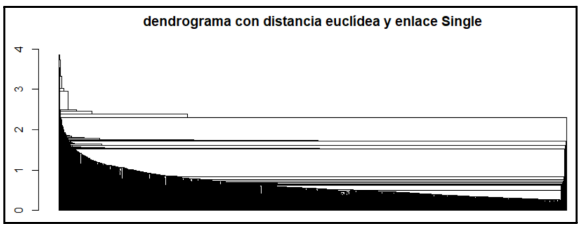
Podemos ver que, en efecto, el cluster **rojo** de la izquierda, está formado por condados de mayoría republicana mientras que el de la derecha **azul** lo estará por demócratas. Si nos fijamos, hay mucha más dispersión dentro del cluster demócrata que en el republicano, demostrando la necesidad de tres clusters para un grupo mucho más heterogéneo que el republicano, el cual coincide casi al completo con el cluster rojo del gráfico anterior, siendo los que no coinciden muy cercanos al centro, lo que significa que el porcentaje de voto entre republicanos y demócratas fue muy cercano al empate.

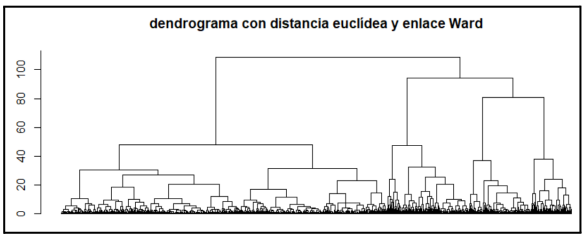
Si recordamos, con el gráfico del método de la silueta, al seleccionar 4 clusters el Si sería superior a 0.3, con el gráfico de abajo podemos confirmar que la media del Si para los clusters sera de 0.37. Además, podemos ver que las 1957 observaciones del cluster rojo se encuentran correctamente clasificadas, puesto que son superiores a 0 (recordemos que 0 significa estar entre dos clusters y negativo que está mal asignado) mientras que en el resto de clusters hay un número pequeño de observaciones mal asignadas, siendo mayor en el cluster verde, que representa a los condados demócratas y blancos, aunque en ningún caso nada preocupante.



Ahora pasaremos a realizar le mismo análisis de clustering pero esta vez con el método jerárquico en lugar de K–Means. Para ello, primero hay que decidir que tipo de distancia y de enlace queremos usar. En su momento usamos la distancia euclídea para calcular la matriz de distancias, por lo que seguiremos usando esta. En cuanto al tipo de enlace, hemos hecho rápidamente un dendrograma para los enlaces Complete, Single y Ward:

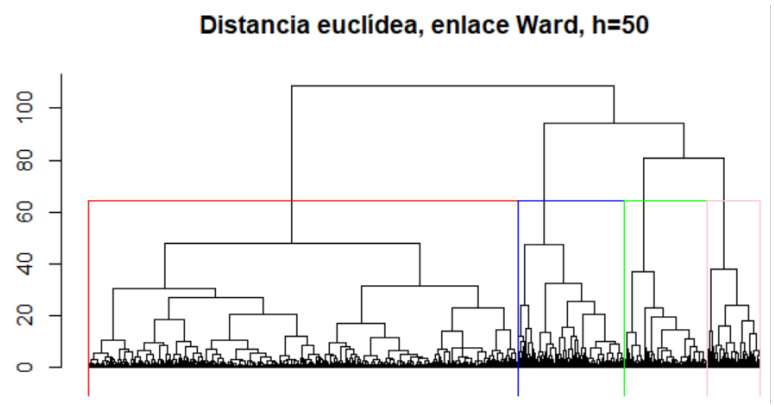


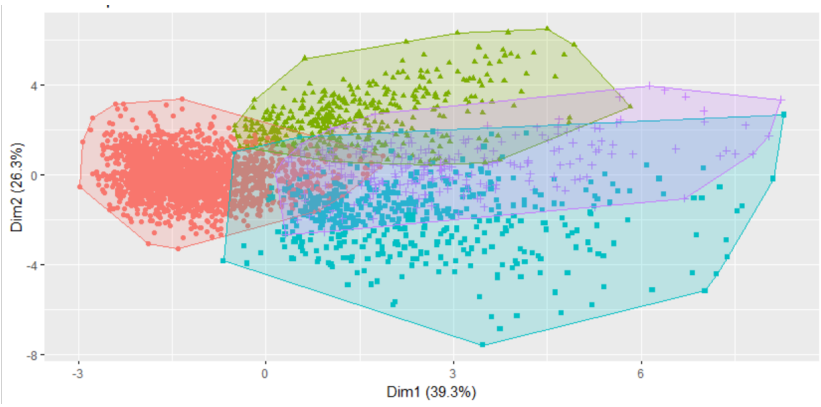




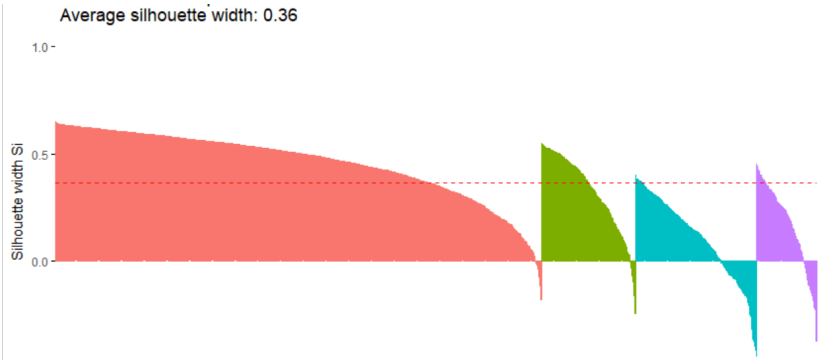
En vista de los resultados obtenidos, hemos decidido quedarnos con el enlace Ward por ser más sencillo. El Single quedó siempre descartado desde un punto de vista teórico, pero el plot nos lo ha reconfirmado.

En función de los análisis de within-cluster variation y de silhouette hechos previamente, volvemos a escoger hacer 4 clusters, que sería como cortar el dendrograma a una altura próxima a 50. Aquí tenemos como se dividen los cluster visualmente:





De primeras ya se puede vislumbrar que en el método jerárquico el clúster afroamericano será mucho más grande que por K–Means. A simple vista parece que los grupos están mucho menos separados que por el método anterior. Decidimos llevar a cabo, una vez más, el análisis de silueta para compararlo con el anterior. Obtenemos lo siguiente:



De por sí, el average silhouette width es 0.36, por lo que es menor que el que obtuvimos al hacerlo por K–Means. Esto significa que las observaciones están peor clasificadas, pues recordemos que nuestro objetivo siempre es maximizar este índice. De manera visual, se ve como ahora sí que hay observaciones mal clasificadas dentro del cluster rojo, algo que no sucedía en el análisis anterior. De la misma forma, el cluster azul toma valores muy negativos en casi la mitad de condados, mostrando que las observaciones no están en el cluster correcto, algo que no sucedía con K–Means. Por todo esto, hemos llegado a la conclusión de que nos quedaremos con el primer análisis en lugar de con el jerárquico.

Partimos de la base de que desconocemos la base de votantes de los partidos políticos en EEUU. De esta forma, este análisis nos ha permitido descubrir como son los electores de tanto el Partido Demócrata como del Partido Republicano. Gracias al clustering, hemos podido tomar datos de los condados y saber como tratar a cada cluster:

1. El cluster **rojo**, formado por condados de mayoría blanca y de clase media, tiende a votar republicano. Estamos hablando de un grupo muy homogéneo al que para conseguir sus votos probablemente haya que tomar un discurso a favor de una fiscalidad baja y evitar discursos que tengan que ver con temas sociales o raciales. Su nivel de educación es medio por lo que no es un tema importante.
2. El cluster **azul**, formado por condados afroamericanos de renta baja, tiende a votar demócrata. Para poder lograr sus votos, el discurso a favor de los impuestos bajos no funcionará debido a sus bajos ingresos, por lo que funcionará mejor hablar de planes sociales y de educación, donde también tienen un déficit. El tema racial también jugará un papel importante.
3. El cluster **morado** o latino, también tiende a un voto demócrata, pero no tanto como los afroamericanos. Tienen un nivel de renta mayor que otras minorías, incluso muy cercano a la renta de los republicanos blancos, por lo que un discurso de fiscalidad baja puede ayudar a rascar votos. Su nivel de estudios también es medio, no es un tema fundamental, a diferencia de en los afroamericanos. Por último, la mayoría de los que forman parte de este cluster han sido inmigrantes, por lo que el discurso pro-inmigracióon ayudará a ganar sus votos.
4. En el cluster **verde**, de blancos de altos ingresos que votan demócrata, triunfarán discursos relacionados con la conciencia social y la desigualdad debido a su alto nivel de estudios que les ha llevado a tener una ideología que tienda más hacia el progresismo que hacia el conservadurismo.

CONCLUSIÓN

Tras haber realizado este extenso análisis, tenemos que decir que estamos más que satisfechos con los resultados obtenidos. No esperábamos que se pudiesen obtener unos grupos tan claros con los distintos tipos de votantes que hay en Estados Unidos, de forma en la que la estrategia a seguir pudiese ser tan clara.

Hemos encontrado correlaciones muy interesantes que enseñan de una manera muy acertada la realidad estadounidense. Por ejemplo, aquellos condados con poca educación resultan ser los que menos ingresos reciben y que terminan votando por partidos a favor de programas sociales.

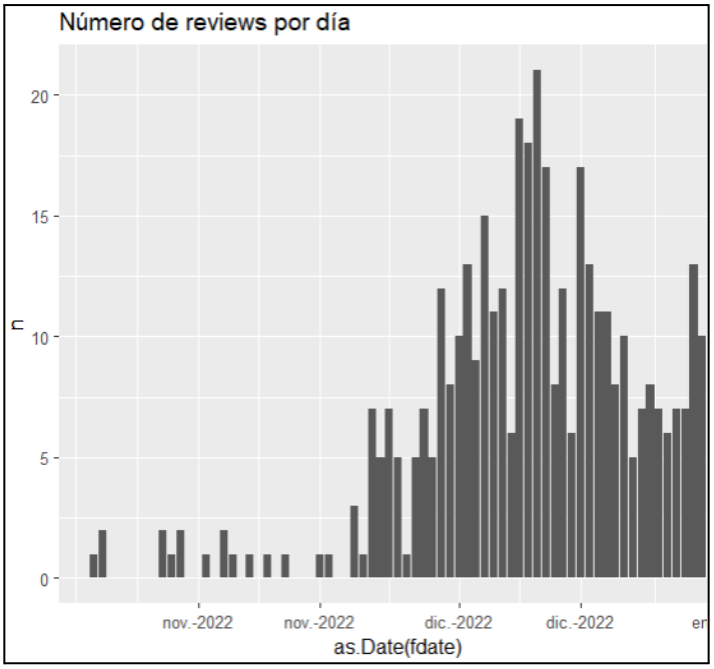
Nos hemos quedado con ganas de poder ampliar este análisis incorporando nuevas variables, como podría ser el número de habitantes por condado, de manera que se viese la división rural–urbana norteamericana.

(*NOTA: Adjuntamos aquí las estadísticas de cada variable ya que arriba no nos dejaba añadirla)

Vemos que la media de votos al GOP es del 66.7%, lo cual no simboliza que ese sea el porcentaje de voto a nivel nacional, ya que se está calculando la media dividida por número de condados y no por número de habitantes. Cada condado tiene una población distinta.

```
County Name        State Abreviation       Fips         Pctg. DEM         Pctg. GOP
Length:3109        Length:3109        Min.   : 1001   Min.   :0.03247   Min.   :0.04251
Class :character   Class :character   1st Qu.:19037   1st Qu.:0.21336   1st Qu.:0.57861
Mode  :character   Mode  :character   Median :29207   Median :0.29922   Median :0.70078
                                      Mean   :30647   Mean   :0.33265   Mean   :0.66735
                                      3rd Qu.:46005   3rd Qu.:0.42139   3rd Qu.:0.78664
                                      Max.   :56045   Max.   :0.95749   Max.   :0.96753
 Black Pctg.     Hispanic or Latino pctg. White alone pctg. Foreign born persons, percent
Min.   : 0.000   Min.   : 0.000           Min.   : 0.00     Min.   : 0.000
1st Qu.: 0.800   1st Qu.: 2.100           1st Qu.:65.70     1st Qu.: 1.200
Median : 2.400   Median : 3.800           Median :84.70     Median : 2.500
Mean   : 9.283   Mean   : 9.045           Mean   :77.36     Mean   : 4.488
3rd Qu.:10.900   3rd Qu.: 9.300           3rd Qu.:93.20     3rd Qu.: 5.500
Max.   :85.100   Max.   :95.800           Max.   :98.60     Max.   :51.300
Bachelor's degree or higher, percent of persons age 25+
Min.   : 3.20
1st Qu.:13.70
Median :17.50
Mean   :19.74
3rd Qu.:23.40
Max.   :74.40
Median household income, 2009-2013 Persons below poverty level, percent, 2009-2013
Min.   : 19986                     Min.   : 0.9
1st Qu.: 38105                     1st Qu.:12.1
Median : 44094                     Median :16.0
Mean   : 45780                     Mean   :16.7
3rd Qu.: 50881                     3rd Qu.:20.3
Max.   :122238                     Max.   :53.2
```

TEXT MINING



Para el ejercicio de text mining propuesto hemos elegido un producto de la tienda de Amazon España, en concreto el Echo Dot, que es un altavoz inteligente con el que se puede interactuar con Alexa. El código específico de este producto en la tienda es 'B09B8X9RGM'.
Como muestra cogemos las primeras 40 páginas de comentarios, que teniendo en cuenta que hay 10 comentarios por página, hacen un total de 400 reviews. Podemos ver que estas se concentran principalmente en el mes de Diciembre, lo que es lógico debido al mayor número de compras que se realizan de este objeto para regalos tanto de Navidad como de Reyes, y por consiguiente aumentan el número de reviews. Haciendo summary sobre los datos obtenemos también que las valoraciones son, en general, positivas.

```
          titulo           fecha          texto              rating     
 Genial       :  8   2022-12-10: 21   Length:400         Min.   :1.00  
 Perfecto     :  7   2022-12-08: 19   Class :character   1st Qu.:4.00  
 Excelente    :  5   2022-12-09: 18   Mode  :character   Median :5.00  
 Muy bueno    :  5   2022-12-11: 17                      Mean   :4.52  
 Buen producto:  4   2022-12-15: 17                      3rd Qu.:5.00  
 Me encanta   :  4   2022-12-04: 15                      Max.   :5.00  
 (Other)      :367   (Other)   :293                                    
```

El primer paso para realizar el análisis, después de haber cargado y leído los textos gracias a la función *scrape_amazon* consiste en la creación del corpus. Sin embargo, antes realizamos ciertos pasos previos:

- Limpieza del patrón en el texto palabra.palabra:
  grep("[A-z]\\.[A-z]",mis_reviews$text)
  mis_reviews$texto<-gsub("\\.(?=[A-z])", " ",mis_reviews$texto, perl = TRUE)
- Limpieza de caracteres especiales y emoticonos:
  mis_reviews$texto<-gsub('[^[:alnum:][:blank:]¿?&/!¡\\.\\-]',"",mis_reviews$texto)

Después de esto construimos el corpus y lo almacenamos en el objeto micorpus:
micorpus<-corpus(mis_reviews$texto)

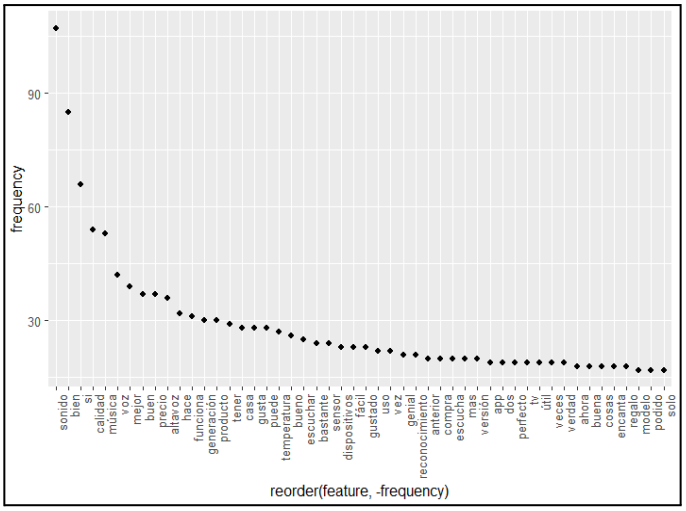
Utilizamos la función *docvar* para añadir fecha, rating y el tipo de valoración a cada documento del corpus:
docvars(micorpus, c("Date", "Rating"))<-mis_reviews[,c(2,4)]
docvars(micorpus, "posneg")<-ifelse(micorpus$Rating<=3, "NEGATIVO", "POSITIVO")

Tokenizamos y luego realizamos la DTM, quitando números, puntuación, y tras realizar la matriz quitamos stopwords y términos que no aporten mucho valor:
mistokens<-tokens(micorpus,remove_numbers = TRUE, remove_punct = TRUE)

midfm<-dfm(mistokens)
midfm<-dfm_remove(midfm, pattern = stopwords("spanish"))
midfm<-dfm_remove(midfm, pattern = c("echo", "dot", "alexa", "amazon"))

No parece necesario hacer stemming.

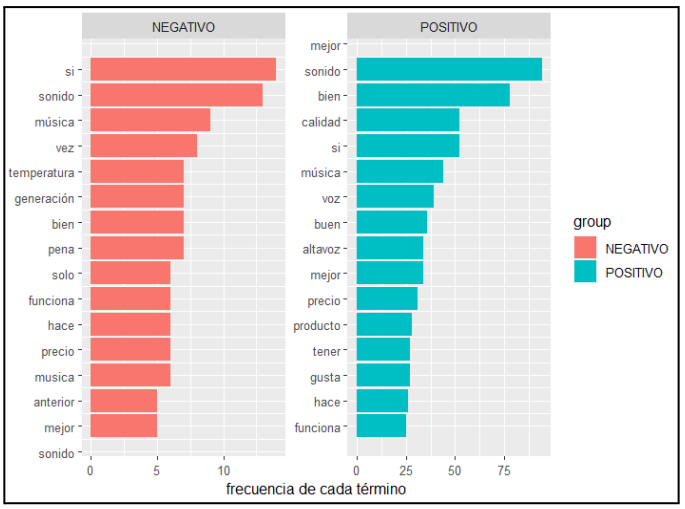
Obtenemos el siguiente gráfico de frecuencias de términos en el conjunto del corpus en general. No sorprende observar que el término más presente sea 'sonido'. A priori parece un término neutral, pero luego analizaremos en qué contextos se ha utilizado. El segundo término es bien, que sí que nos atreveremos a decir que es positivo. El tercer término es si, dependiendo si es condicional o no tendrá polaridad positiva.



features_freq<-textstat_frequency(midfm, n = 50)

ggplot(features_freq, aes(x = reorder(feature,-frequency), y = frequency)) + geom_point() + theme(axis.text.x = element_text(angle = 90, hjust = 1))

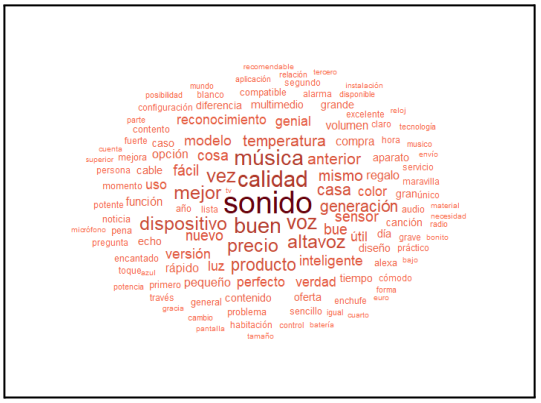
Para evaluar la frecuencia de los términos en reseñas positivas (<=4 estrellas) como negativas(>4 estrellas) diseñamos dos barplots para cada categoría. Los tres términos más repetidos en las reseñas positivas son sonido, bien y calidad; en las reseñas negativas son si, sonido y música.



freq_posneg <- textstat_frequency(midfm, n = 15, groups = posneg)

ggplot(data = freq_posneg, aes(x = nrow(freq_posneg):1, y = frequency, fill=group)) + geom_bar(stat="identity") + facet_wrap(~group, scales = "free") + coord_flip() + scale_x_continuous(breaks = nrow(freq_posneg):1, labels = freq_posneg$feature) + labs(x = NULL, y = "frecuencia de cada término")

Para apoyar este análisis de términos, creamos una nube de palabras lematizadas, para interpretar visualmente cuáles son los tokens (post-lematización) que más aparecen en el DTM.



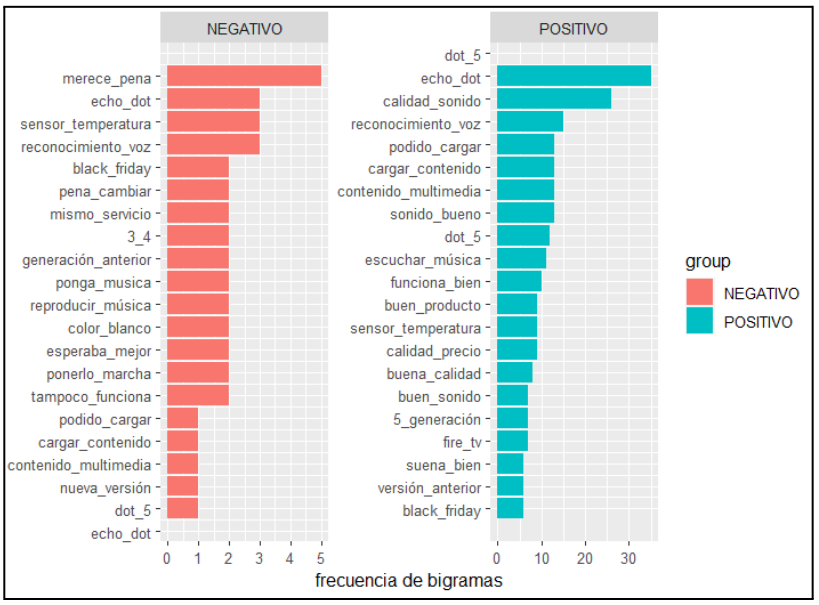
anot_plot <- anot_texto %>% filter(upos %in% c("NOUN","ADJ")) %>% count(lemma, sort=T)

ggplot(anot_plot %>% filter(n>5), aes(label = lemma, size = n, color=n)) + geom_text_wordcloud(eccentricity = 1) + scale_size_area(max_size = 10) + theme_minimal() + scale_color_gradient(low="#fb6a4a", high="#67000d")

Como comentamos anteriormente, hay términos que aparecen tanto en reseñas positivas como en reseñas negativas teniendo alta frecuencia en ambas. Para contextualizar algunos de estos, quedándonos con el término música y el término sonido, nos servimos de la librería *kwic()* para extraer algunos ejemplos de los mismos. Con ello observamos que algunas opiniones comentan que el sonido ha mejorado con respecto a la versión anterior del producto, mientras que otros comentan que al sonido le falta potencia. Con respecto al término música observamos que la utilización del dispositivo para poner música es algo bastante frecuente.

```
[text347, 24]          quizá levemente una pequeña mejora en el | sonido | aunque por el contrario la app Alexa
 [text357, 5]                                Me ha gustado el | sonido | y no me ha gustado que hay
 [text359, 4]                                   La calidad de | sonido | es la misma que el anterior modelo
[text361, 14]               pensando que era mejor en cuanto a | sonido | este Rrhh urbe más volumen pero en
 [text364, 2]                                             El | sonido | es bueno. Y el reconocimiento de
 [text371, 5]                              Me esperaba un mejor | sonido | . Alexa te puede hacer suscribirte a
[text374, 11]                          me esperaba otra cosa... | sonido | normalucho.
[text376, 15]              ecucha fuerte. Le falta potencia de | sonido |
[text377, 64]          igual con el resto solo puede reproducir | música | el titular de la cuenta. El
[text377, 93] que pueden pedirle verbalmente a Alexa reproducir | música | no pueden ver los altavoces en la
[text378, 78]            en internet le preguntas y ya Reproduce | música | aleatoria cuando quieras Yo me lo compré
[text378, 90]              me lo compré para quee despedirte con | música | y cumple su función muy bien.
[text379, 18]              formato de la generación anterior. El | sonido | mejora ligeramente y ahora muestra la temperatura
[text379, 66]        directamente para usarlos en estéreo o con | sonido | envolvente solo es posible con un FireTV
 [text383, 4]                                   Una vez pongo | música | cuando empieza a sonar hace ruidos de
[text385, 24]              radio pone alarmas y escucha algo de | música | . Todos muy contentos. Yo alguna
[text385, 35]                contentos. Yo alguna vez he puesto | música | de grupos que me gustan y sin
[text385, 67]                  antojo para que te suscribas a la | música | unlimited 10 euros más al mes.
 [text389, 5]                            Carece de conexión de | sonido | a través de hacha
[text393, 11]                como 4 generación. Pero no vincula | sonido | multiestancia ni se puede usar con los
[text394, 12]            perfectamente. Ahora le pides temas de | música | y te pone lo que le da
[text395, 20]                          con el fire stick TV y el | sonido | va por libre. La aplicación es
```

Para ampliar esta contextualización realizamos un análisis mediante bigramas. Estas son parejas de palabras que nos ayudarán a entender mejor la contextualización de las mismas. Por ejemplo, en este bigrama vemos que tres de las críticas más habituales son sobre que el producto no merece la pena, sobre el sensor de temperatura y sobre el reconocimiento de voz. Mientras que las alabanzas más habituales son sobre la calidad del sonido, el reconocimiento de voz y sobre el contenido. Como términos más neutrales intuimos que alguna gente lo compró en black friday o que lo utiliza con el Amazon Fire TV.




```
      echo_dot    calidad_sonido reconocimiento_voz    podido_cargar cargar_contenido contenido_multimedia
            38                27                18                14               14                14
  sonido_bueno           dot_5 sensor_temperatura escuchar_música buen_producto     funciona_bien
            14                13                12               11               10                10
 calidad_precio     black_friday     buena_calidad     5_generación         fire_tv       buen_sonido
             9                 8                 8                8                8                 7
     app_alexa     4_generación
             7                 7
```

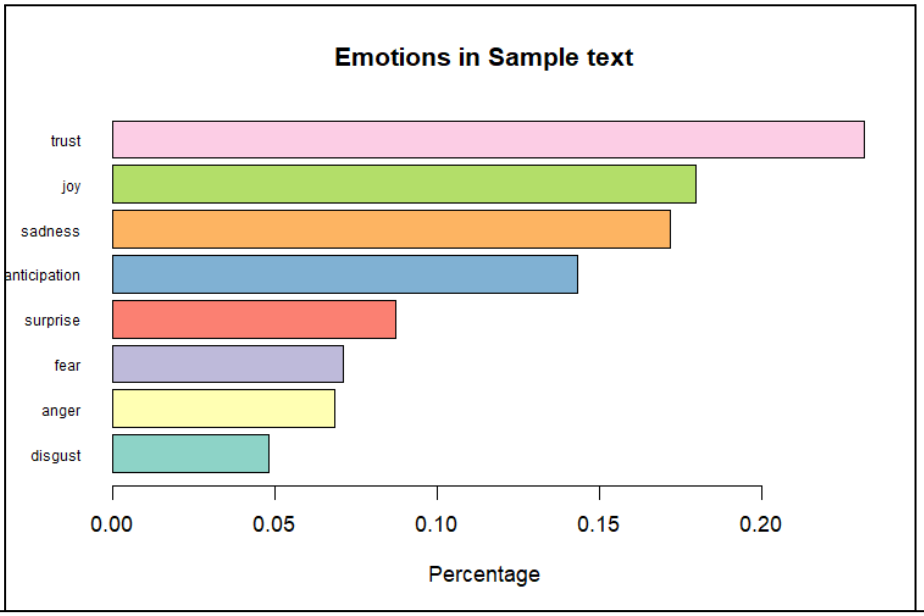
Tras estos pasos comenzamos con el Análisis de Sentimiento de las reseñas para lo que nos serviremos de la librería *syuzhet*. Sirviendonos del siguiente código obtenemos el porcentaje de comentarios con polaridad positiva, neutra y negativa, obteniendo que la mayoría de comentarios son neutro, habiendo pocos abiertamente negativos.

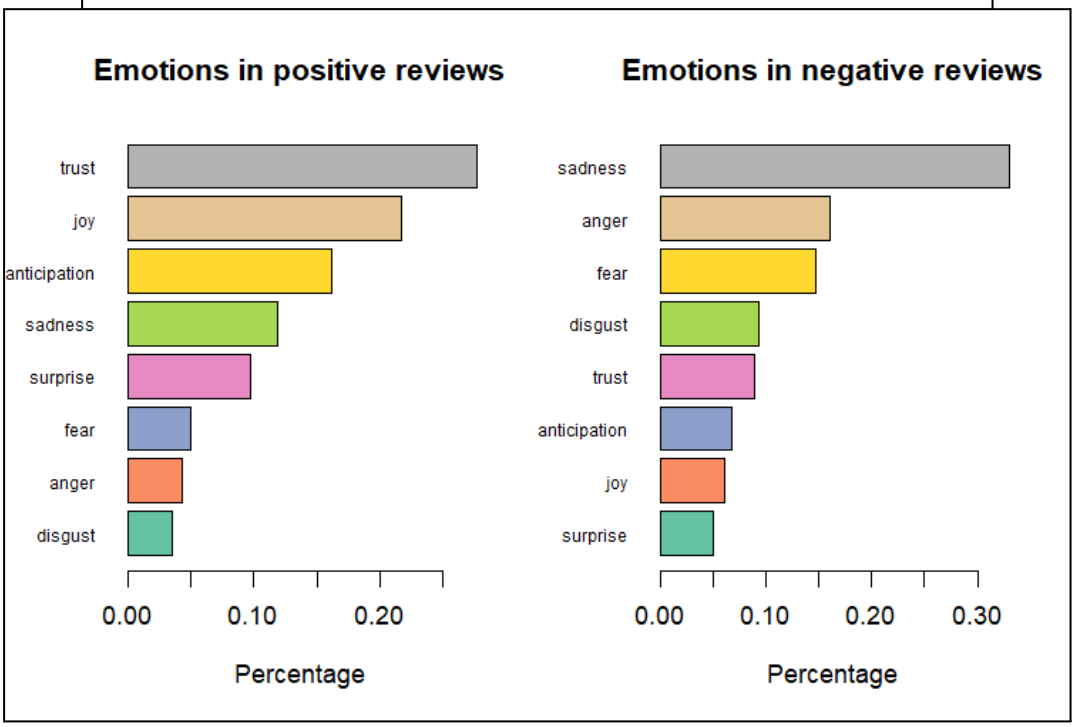
```
positivos neutros negativos
     13.5    80.5         6
```

```
reviews<-mis_reviews %>% mutate(nrc_polarity= get_sentiment(mis_reviews$texto, method="nrc"),
    bing_polarity=get_sentiment(mis_reviews$texto, method="bing"))

reviews %>% summarise(positivos = 100 * sum(nrc_polarity > 0) / n(), neutros = 100 *
    sum(nrc_polarity == 0) / n(), negativos = 100 * sum(nrc_polarity < 0) / n())
```

Además, podemos obtener barplots para representar el peso que tiene cada emoción en el total del corpus o el porcentaje sobre las reviews con polaridad positiva y negativa. Observamos que la emoción predominante en el corpus es la confianza, que se transmite también en las reviews positivas, mientras que en las negativas se transmite tristeza como la emoción predominante.





Conclusiones

Como conclusión del proceso de text mining obtenemos que la mayoría de reviews son positivas y que seguramente haya habido un aumento de ventas en diciembre.

Gracias al proceso de tokenización y a la formación de la matriz de términos vemos que los topics más importantes para los consumidores son la música y el sonido, por lo que independientemente de las valoraciones que se hagan sobre esos aspectos, que parecen repartidas, Amazon deberá centrarse en optimizar esos dos atributos de producto puesto que son los más comentados.

Dados los bigramas, somos capaces de concretar más en la problemática del producto, y vemos que se debería perfeccionar el sensor de temperatura, la utilidad general del producto (puesto que se pone en duda que valga la pena comprarlo) y la calidad de los dispositivos vendidos durante el black friday, ya que se intuye que existe cierta problemática en particular con la gente que compró el producto el black friday. Como aspectos positivos se reconoce principalmente la calidad del sonido que ofrece el aparato, la facilidad de carga y el contenido multimedia del producto. También a tener en cuenta que hay cierto número de usuarios que lo usan con Amazon Fire TV y que aparenta que se compatibilizan bien.

Respecto al análisis de sentimientos inducimos que el producto tiene un efecto neutral sobre el consumidor, hecho representado en las reseñas, de las que no se desprende ni gran pasión ni adversión. De las reviews positivas el valor que se desprende es confianza, por lo que deducimos que nuestro producto es estable y no da fallos inesperados. El segundo factor que se desprende sobre todas las reviews es alegría y el tercero tristeza. Estos sentimientos resultan difíciles de interpretar al nivel de producto, aunque podemos encender ambos como reacciones post compra del artículo y con alegría como un componente positivo de satisfacción con lo comprado y tristeza lo opuesto. Que el producto no genere mayor exaltación ni animadversión es consistente al ver que los sentimientos de sorpresa, disgusto, miedo e ira son residuales.

Tras estas conclusiones, podemos enumerar ciertas recomendaciones para Amazon, el productor:

- Potenciar la producción y publicidad del producto, ya que parece que satisface al consumidor
- Dedicar la mayoría de recursos a la hora de asegurar una excelente experiencia en lo relativo al sonido y al uso que se le da para escuchar música
- Perfeccionar el sensor de temperatura, que parece que falla
- Aumentar el número de utilidades del producto
- Vigilar la venta que se realiza durante el black friday
- El reconocimiento de voz supone problemas en algunos casos y en otros se reconoce como bueno, así que recomendaría asegurarse de que funciona correctamente, pero como última prioridad
- Consideraría importante introducir algún aspecto en el producto que suponga una mayor atracción o sorpresa grata, tratando de desarrollar quizás atributos originales o innovativos que hagan que el producto dé más que hablar, ya que el producto se encuentra bien valorado pero tiene margen para arriesgar y despertar algo más en el consumidor