Laboratorio 1, Estadística Computacional

I. Integrantes

- Juan Pablo León (201473047-0)
- Daniel Águila ()

II. Preguntas

Pregunta 1

Definimos los siguientes tres requerimientos de tipo descriptivo:

1. ¿Qué categoría posee la mayor razón de likes vs dislikes?

2. ¿Qué categoria tiene la mayor cantidad de comentarios?

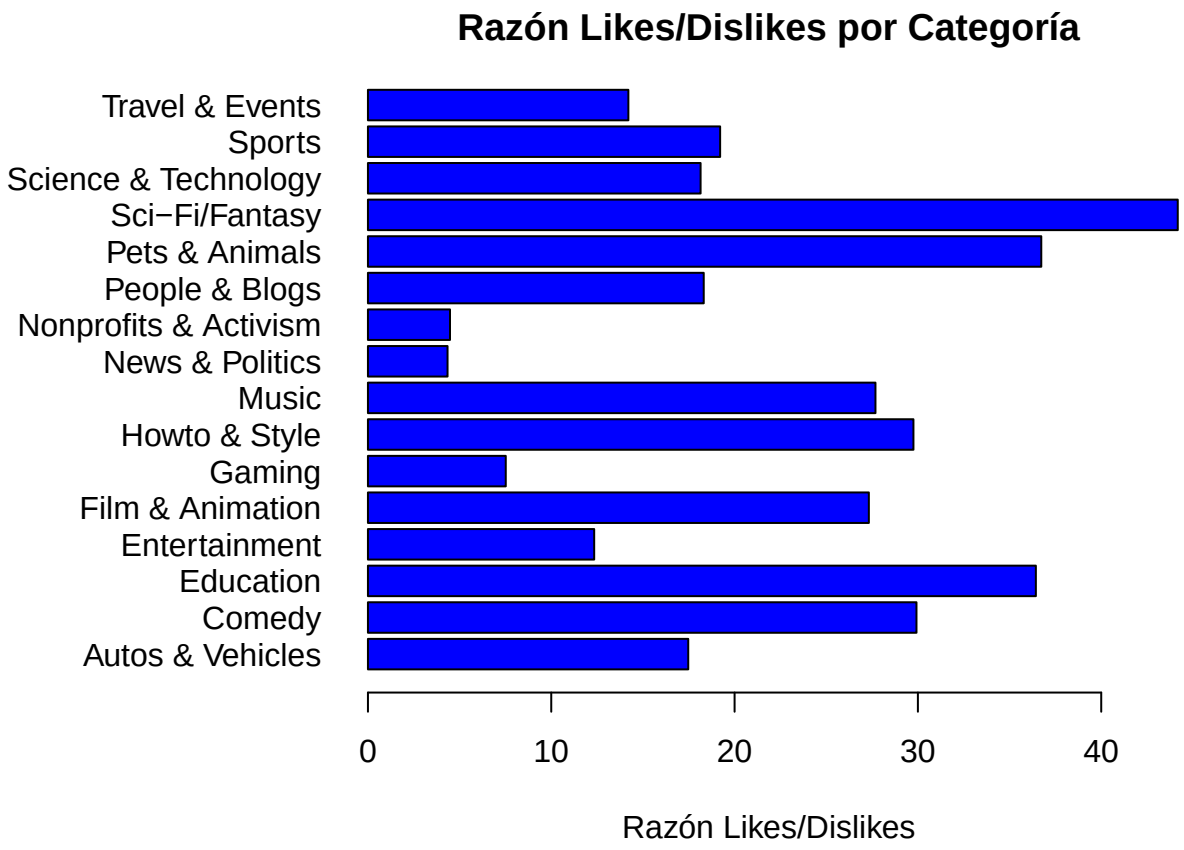
3. ¿Qué categoria tiene la mayor cantidad de videos con los comentarios desactivados?

En primer lugar, debemos obtener nuestros set de datos y las diferentes variables:

```
datos = read.csv("USvideosWithCategories.csv", header=T)
attach(datos)
```

Luego, para el primer requerimiento formamos un gráfico de barras con el siguiente código:

```
likesPorCategoria = tapply(X=likes, INDEX=category, FUN=sum, na.rm=TRUE)
dislikesPorCategoria = tapply(X=dislikes, INDEX=category, FUN=sum, na.rm=TRUE)
razonLikesDislikes = likesPorCategoria/dislikesPorCategoria

op = par(mar=c(4,10,2,1) + 0.1 )
grafico1 = barplot(razonLikesDislikes, xpd=NA, las=1, horiz=TRUE, col="blue",
        xlab="Razón Likes/Dislikes",
        main="Razón Likes/Dislikes por Categoría")
```

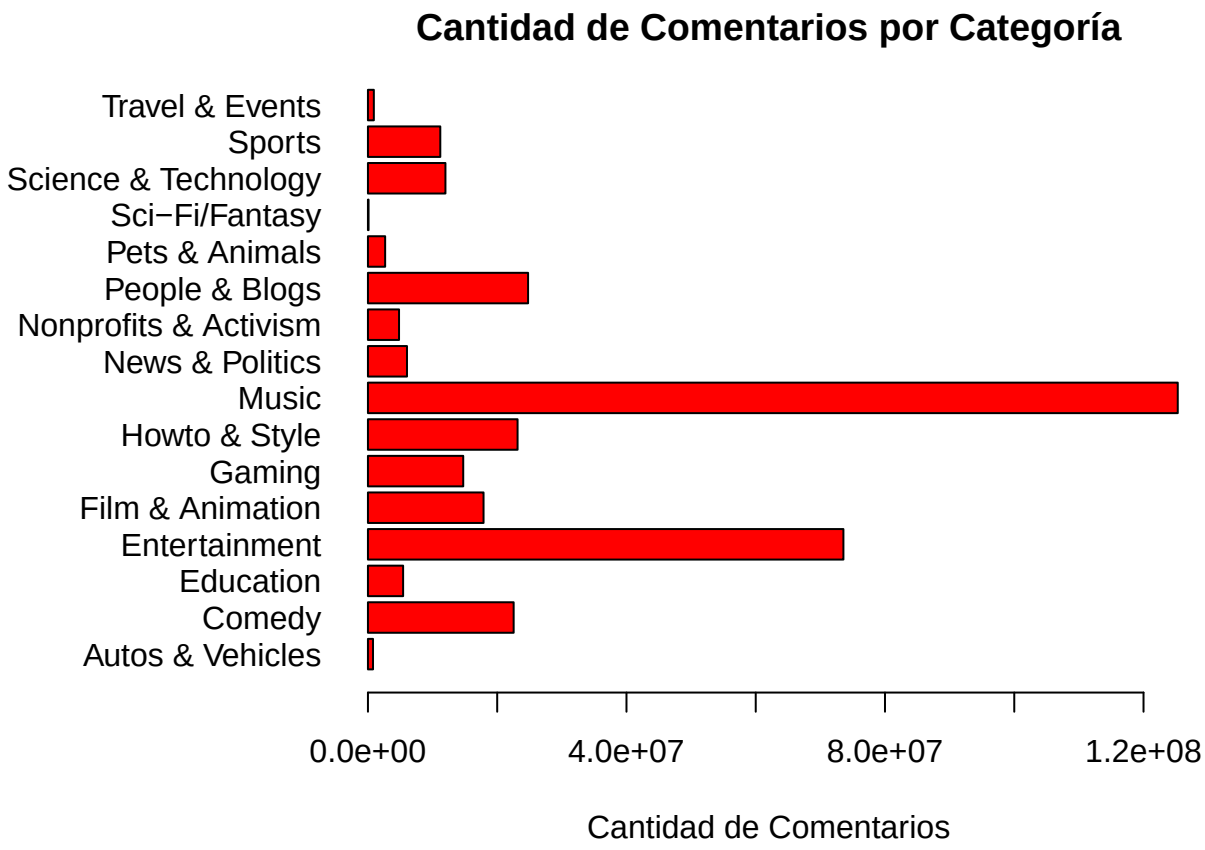



```
par(op)
```

Donde podemos observar cláramente que la categoría de "Sciencia Ficción y Fantasía" tiene la mejor razón likes/dislikes, seguido de cerca por "Mascotas y Animales" y "Educación."

Similarmente, para nuestro segundo requerimiento tenemos:

```
comentariosPorCategoria = tapply(X=comment_count, INDEX=category, FUN=sum, na.rm=TRUE)

op = par(mar=c(4,10,2,1) + 0.1)
grafico2 = barplot(comentariosPorCategoria, las=1, horiz=TRUE, col="red",
                   xlab="Cantidad de Comentarios", main="Cantidad de Comentarios por Categoría")
```

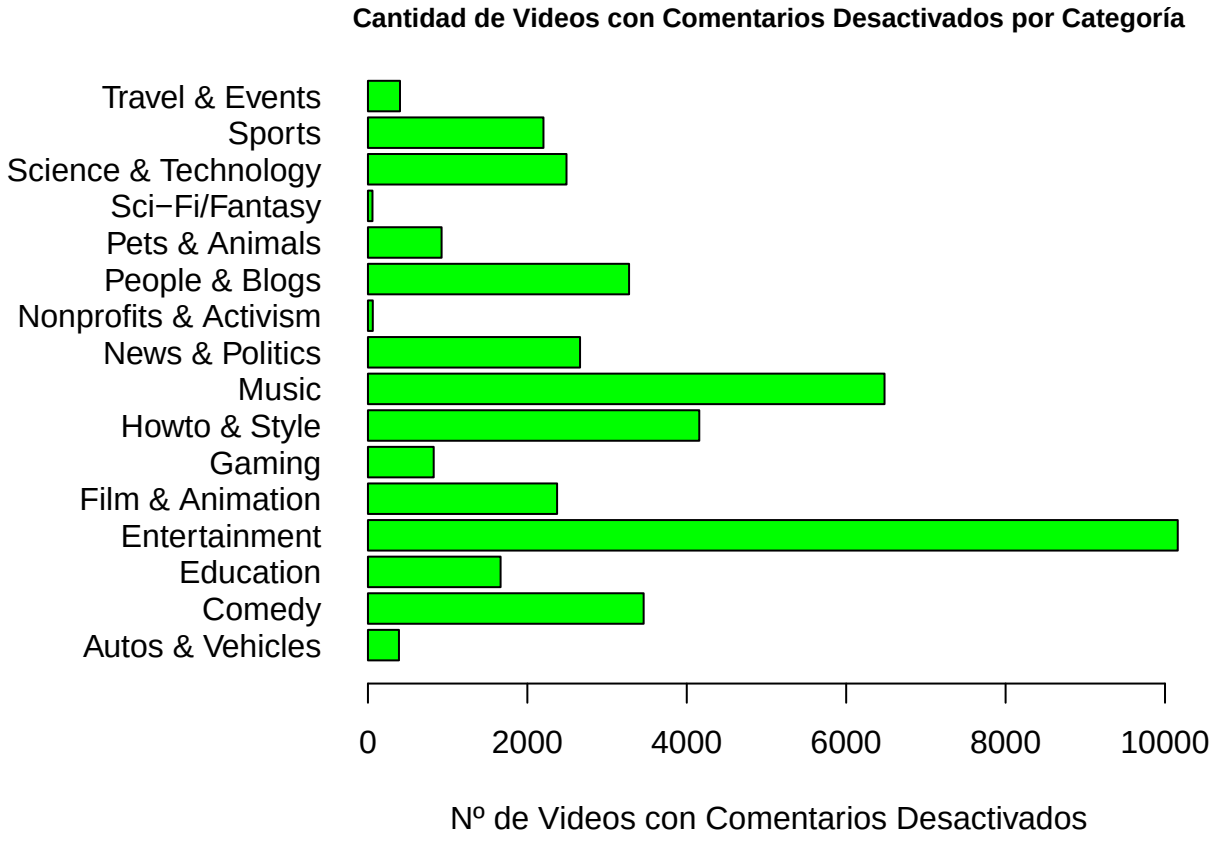



```
par(op)
```

Que nos dice que la categoría con más comentarios es “Música” seguido por “Entretenimiento.”

Finalmente, para nuestro tercer requerimiento se obtuvo el siguiente gráfico:

```
comments_disabled = as.numeric(comments_disabled)
comentariosDesactivadosPorCategoria = tapply(X=comments_disabled, INDEX=category,
                                             FUN=sum, na.rm=TRUE)

op = par(mar=c(4,10,2,1) + 0.1)
grafico3 = barplot(comentariosDesactivadosPorCategoria, las=1, horiz=TRUE,
                   col="green", cex.main=0.85,
                   xlab="Nº de Videos con Comentarios Desactivados",
                   main="Cantidad de Videos con Comentarios Desactivados por Categoría")
```




```
par(op)
```

Donde se aprecia que la categoría con más video con los comentarios desactivados es "Entretenimiento," seguido de "Música."

Pregunta 2

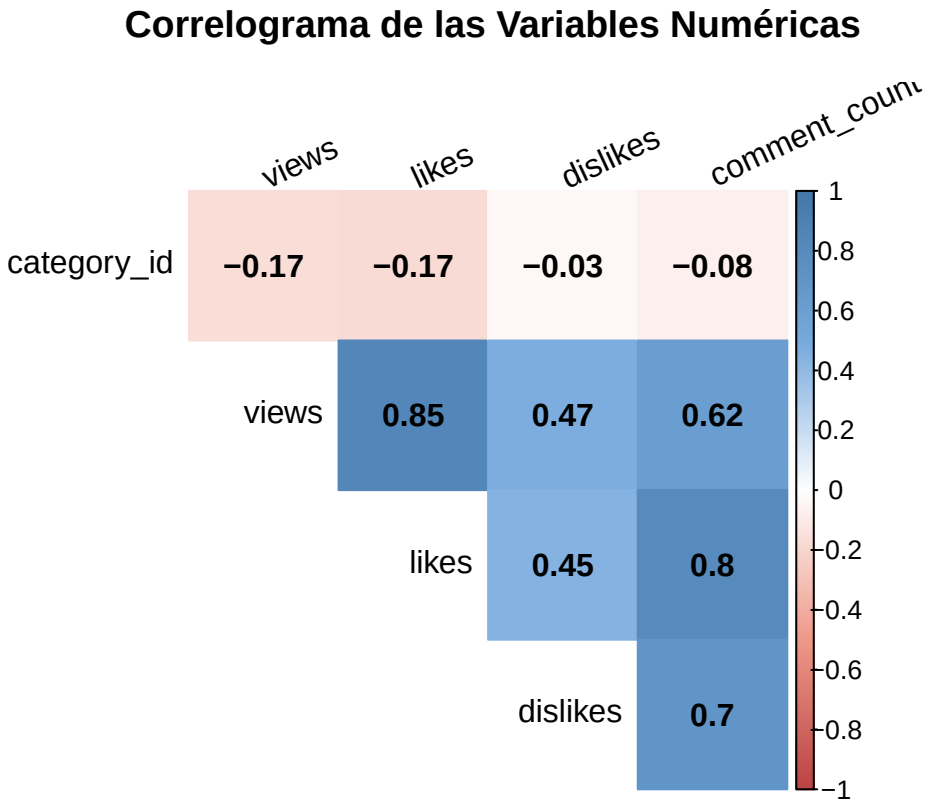
Antes de crear el correlograma debemos asegurarnos de que la columna "X" (la primera columna de la tabla) no sea considerada, pues esta sirve solamente como contador y no entrega información relevante, además aprovechamos de cargar la librería necesaria:

```
datos$X = NULL
library(corrplot)
```

Luego, generamos nuestro correlograma:

```
labels = c("category_id", "views", "likes", "dislikes", "comment_count")
matriz = as.matrix(datos[labels])
matrizCorrelacion = cor(matriz)

col <- colorRampPalette(c("#BB4444", "#EE9988", "#FFFFFF", "#77AADD", "#4477AA"))
corrplot(matrizCorrelacion, method="color", col=col(200),
         type="upper", order="hclust",
         addCoef.col = "black", # Add coefficient of correlation
         tl.col="black", tl.srt=25, #Text label color and rotation
         # Combine with significance
          sig.level = 0.01, insig = "blank",
         # hide correlation coefficient on the principal diagonal
```

```
        diag=FALSE, main="Correlograma de las Variables Numéricas",
        mar=c(0,0,3,0) + 0.1)
```



Aquello de interés:

- No existe una correlación lineal entre la categoría del video y el resto de las variables numéricas.
- Existe una correlación positiva que tiende a ser del tipo directa entre la cantidad de visitas que tiene un video y su cantidad de likes.
- Existe una correlación positiva que tiende a ser del tipo directa entre la cantidad visitas que tiene un video y su cantidad de comentarios.
- Existe una correlación positiva más alta entre la cantidad likes y comentarios que tiene un video que la cantidad de dislikes y comentarios.

Pregunta 3

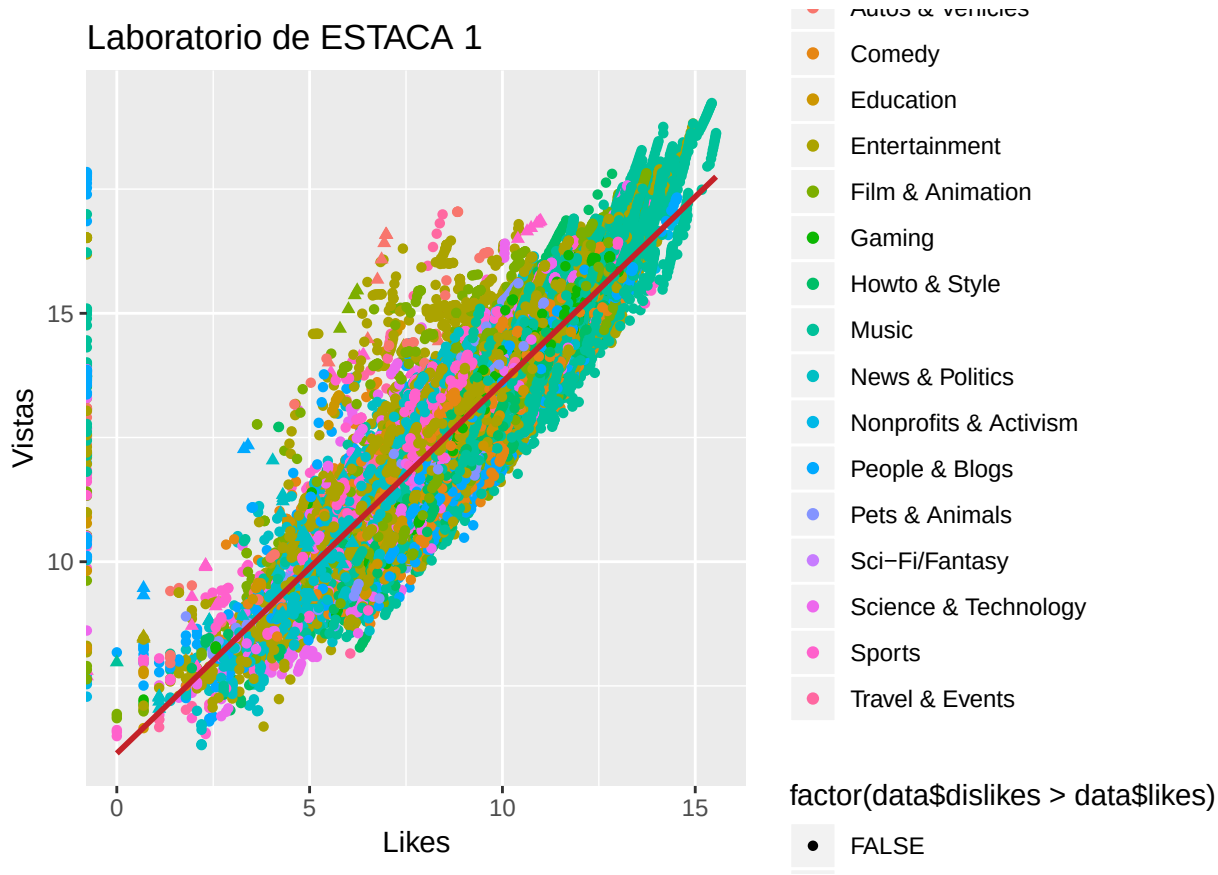
```
library(ggplot2)
library(labdsv)
library(rmarkdown)
```

```
data<-read.csv("USvideosWithCategories.csv", stringsAsFactors=FALSE)
cuenta = 0
mayor = 0
for(i in 2:18){
  for(j in 2:18){
    if(typeof(data[[i]][1]) == 'integer' && typeof(data[[j]][1]) == 'integer'){
```

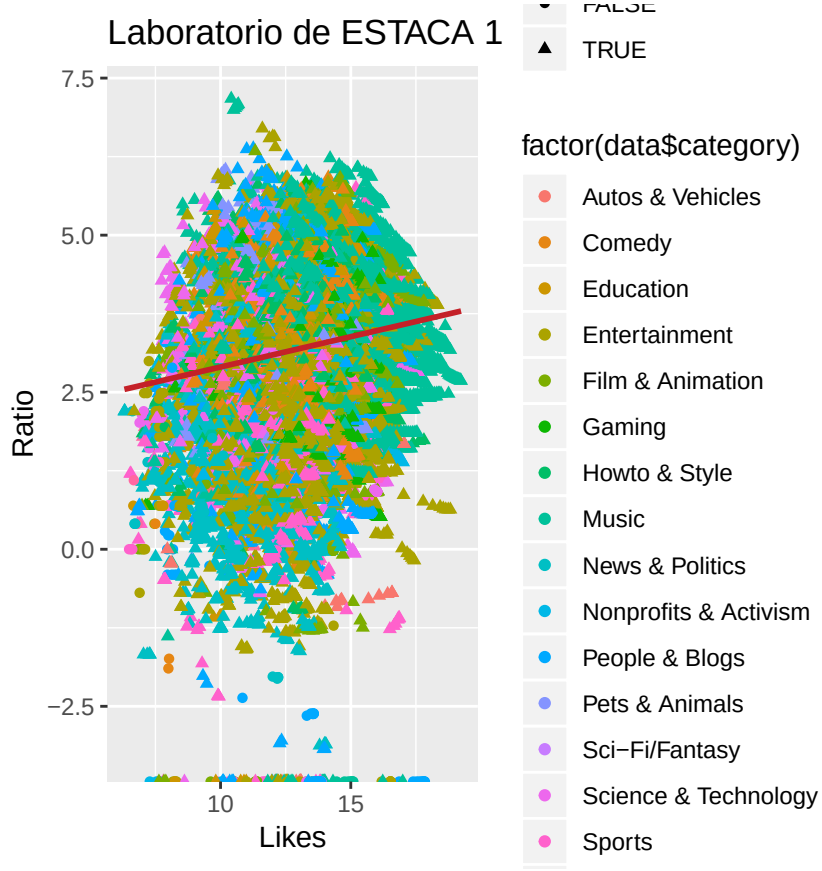
```
      #print("numeric")
      #print(cor(data[i], data[j], method=c("pearson")))
      if(mayor < cor(data[i], data[j]) && data[i] != data[j] ){
        mayor = cor(data[i], data[j], method=c("pearson"))
      }
    }
  }

}
print(mayor)
```

```
##           likes
## views 0.8491765
```

```
ggplot(data, aes(x=log(data$likes), y=log(data$views)))+geom_point(aes(color = factor(data$category),sha
```




```
ggplot(data, aes(x=log(data$views), y=log(data$likes/(data$dislikes + 1))))+geom_point(aes(shape=factor
```



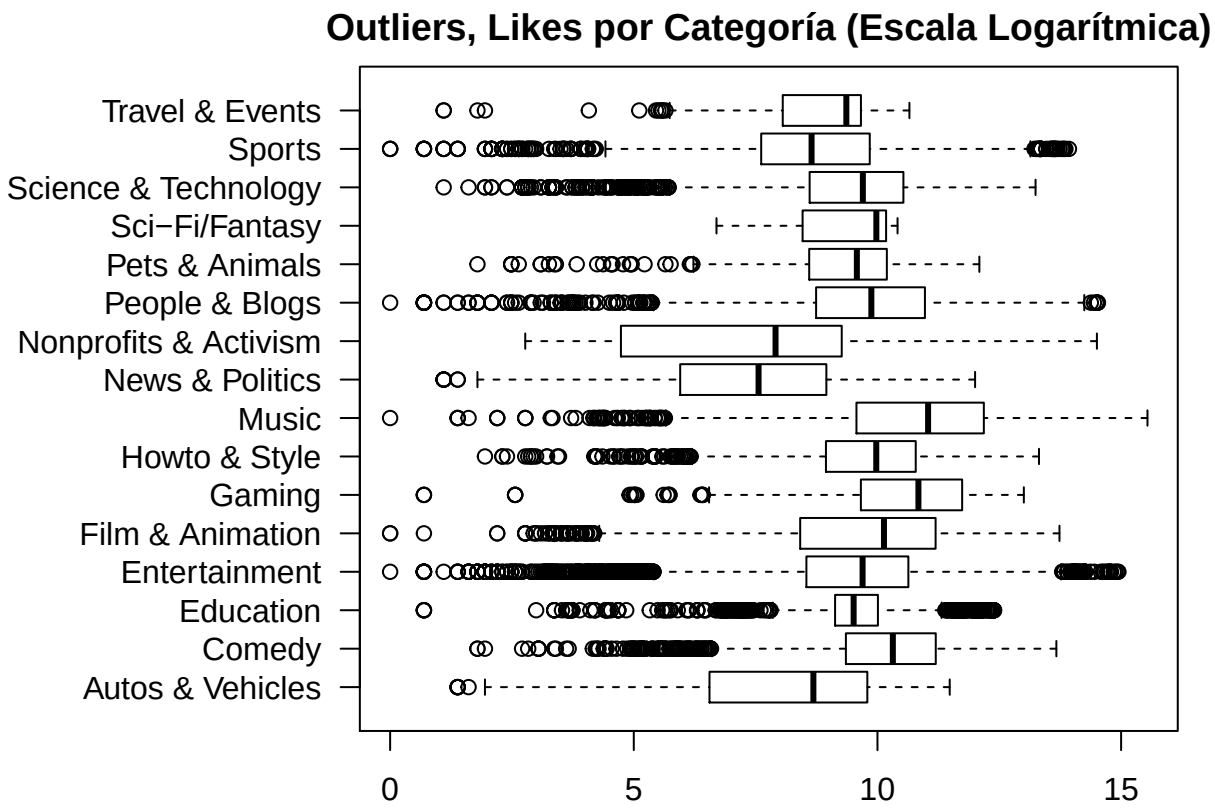
Encontramos que las dos variables con mayor correlacion son el número de vistas y los likes con un valor de 0.85, tiene sentido si pensamos que a mayor numero de vistas existe mayor probabilidad existe que el video tenga una gran cantidad de likes. Ademas se aplico distinto color a cada punto con el fin de distingir a que categoria pertenecen, finalmente si muestran en forma circulas los puntos donde el número de likes es mayor al de dislikes y en forma de triangulos en el caso contrario, notando una clara mayoria en el numero de likes que de dislikes

Pregunta 4

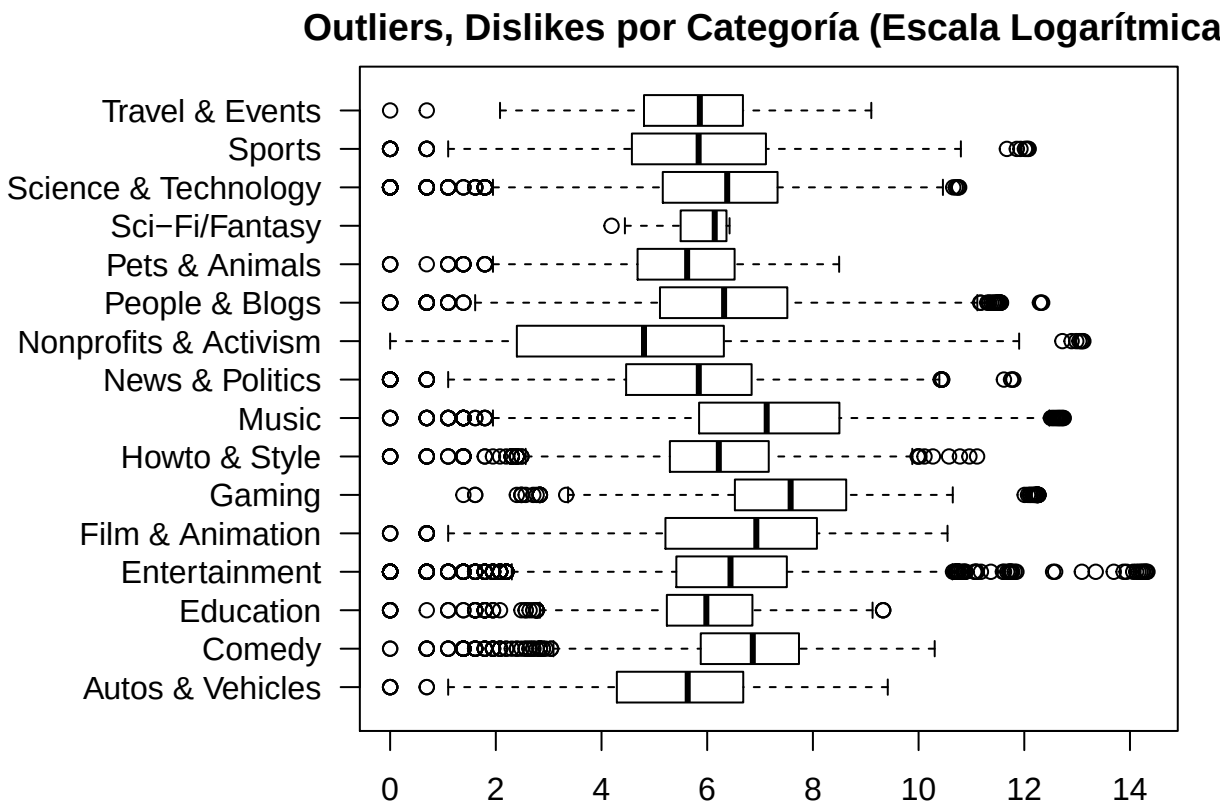
Pregunta 5

Para observar los outliers de los likes y dislikes entre las distintas categorías generamos los siguientes gráficos de cajas:

```
op = par(mar=c(3,10,2,1) + 0.1)
boxplot(log(likes) ~ category, data=datos, las=1, horizontal=TRUE,
        main="Outliers, Likes por Categoría (Escala Logarítmica)",
        xlabel="Cantidad de Likes")
```




```
par(op)

op = par(mar=c(3,10,2,1) + 0.1)
boxplot(log(dislikes) ~ category, data=datos, las=1, horizontal=TRUE,
        main="Outliers, Dislikes por Categoría (Escala Logarítmica)",
        xlabel="Cantidad de Likes")
```




```
par(op)
```

Para el caso de la cantidad de likes:

- Las categorías "Sciencia Ficción/Fantasía," "Sin Fines de Lucro & Activismo," "Noticias y Política" y "Autos & Vehículos" no tienen casi valores extremos fuera de sus gráficos de cajas.
- Se puede observar una gran cantidad de outliers tanto sobre como por debajo del gráfico de caja de las categorías "Entretenimiento" y "Educación."
- La gran mayoría de las categorías poseen outliers que se encuentran por debajo de sus gráficos de cajas.

Para el caso de la cantidad de dislikes:

- Se puede observar que la cantidad de outliers en las categorías de "Deportes," "Sciencia y Tecnología," "Mascotas y Animales," "Gente y Blogs" y "Películas y Animación" se ha reducido de manera apreciable en comparación con el caso de la cantidad de likes.
- La cantidad de outliers en las categorías "Sciencia Ficción/Fantasía," "Sin Fines de Lucro & Activismo," "Noticias y Política" y "Autos & Vehículos" es similar a la cantidad de outliers en el caso de la cantidad de likes.
- La mayoría de las categorías tienden a tener outliers tanto debajo como por sobre su gráfico de caja, de manera que parecen estár mejor distribuídos que en el caso de la cantidad de likes.

III. Conclusiones

Dentro de las conclusiones pudimos observar que:

1. La categoría que posee la mayor razón de cantidad de likes versus cantidad de dislikes es la categoría de "Sciencia Ficción/Fantasía," seguida de "Mascotas y Animales" y "Comedia."

2. La categoría que posee la mayor cantidad de comentarios es la de "Música," seguida de lejos por la

categoría de “Entretenimiento.”

3. La categoría que cuenta con la mayor cantidad de videos cuyos comentarios han sido desactivados es la categoría de “Entretenimiento”, seguida de las categorías de “Música” y “Howto & Style.”
4. No existe una correlación del tipo lineal entre la categoría de un video y su cantidad de visitas, likes, dislikes y comentarios.
5. Tanto la cantidad de visitas y cantidad de likes como la cantidad de likes y la cantidad de comentarios de un video, tienen una correlación positiva que tiende a ser del tipo directa, o sea, mientras mayor sea una a su vez mayor será la otra.
6. Los puntos extremos/outliers tienden a estar por debajo del gráfico de cajas en el caso de la cantidad de likes según categoría que en el caso de la cantidad de dislikes, donde la distribución de las anomalías tiende a ser más equilibrada.
7. La mediana en los gráficos de cajas del caso de la cantidad de likes tiende a estar por sobre la mediana en el caso de la cantidad de dislikes, lo que puede explicar el por qué de la conclusión anterior.