PROYECTO ESTADÍSTICA

POR: LESTHER Y JEFFERSON

6/3/2021

CONTEXTO:

-> DATOS DESDE EL 12 DE AGOSTO DEL 2020 HASTA 03 DE JUNIO DEL 2021. REGISTROS RELACIONADOS AL TRÁFICO (VISTAS,LIKES, DISLIKES, COMENTARIOS) DE LOS VÍDEOS DE YOUTUBE; SÓLO DATOS DE MÉXICO

IMPORTACIÓN DE LIBRERÍAS PERTINENTES.

-> message=FALSE para que no aparezcan mensajes en el archivo word.

```
require(stats)
library(dplyr)

## Warning: package 'dplyr' was built under R version 4.0.5

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.0.5

library(corrplot)

## Warning: package 'corrplot' was built under R version 4.0.4
```

LECTURA DEL ARCHIVO.

```
ytmx <- read.csv('./src/MX_youtube_trending_data_fixed.csv')
```

SIGNIFICANCIA A EMPLEAR DURANTE LOS ANALISIS. DECLARACIÓN DE VARIABLE 'BOOLEANA' PARA MANEJAR LAS VALIDACIONES.

```
significancia <- 0.05
rechazarH0 <- FALSE
```

ANÁLISIS EXPLORATORIO

-> A CONTINUACIÓN SE ANALIZAN LAS VARIABLES ÚTILES PARA LOS ANÁLISIS POSTERIORES.

VARIABLES CUANTITATIVAS: view_count, likes, dislikes, comment_count

VARIABLES CUALITATIVAS: comment_disabled, ratings_disabled, category

ANÁLISIS 01

-> ¿EXISTIRÁ ALGUNA RELACIÓN EN LOS VÍDEOS QUE TIENEN LOS COMENTARIOS DESHABILITADOS (O NO) CON RESPECTO A LA CATEGORÍA A LA QUE PERTENECEN?

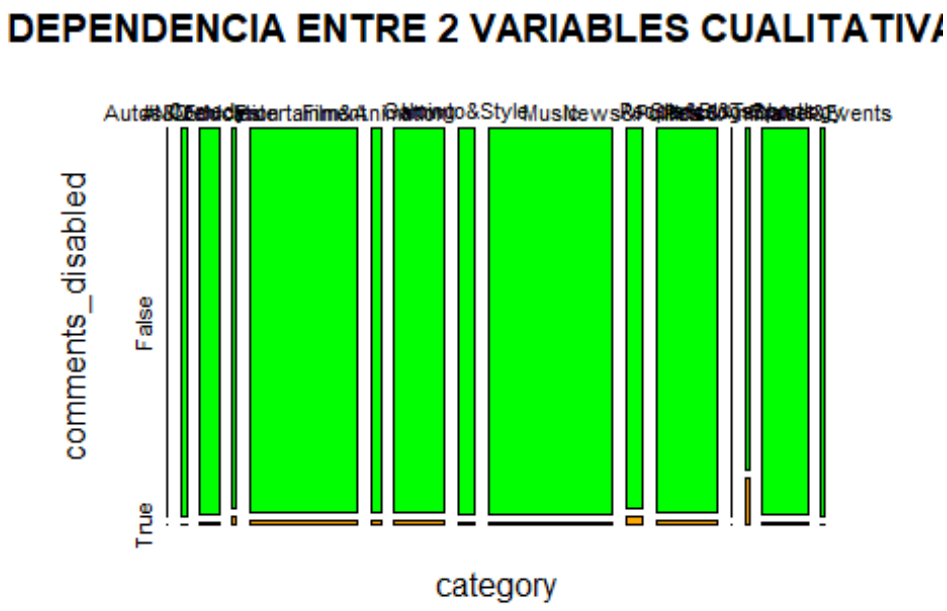
-> VARIABLES: 'comments_disabled', 'category' (cualitativa vs cualitativa)

-> MODELO: TEST CHI CUADRADO

-> GRAFICA: MOSAICO

```
categoryByComentDisabled <-
table(ytmx$category,ytmx$comments_disabled,dnn=c("category","comments_dis
abled"))

mosaicplot(categoryByComentDisabled,main="DEPENDENCIA ENTRE 2 VARIABLES
CUALITATIVAS",col=c('green','orange'))
```



-> VER GRAFICA CONFUSA 01 (CARPETA DE ARCHIVOS UBICADA EN ESTA MISMA CARPETA)

SE PUEDE VALIDAR QUE LA MAYORÍA DE VÍDEOS *NO* TIENEN LOS COMENTARIOS DESHABILITADOS. Y DE LOS QUE SÍ TIENEN LOS COMENTARIOS DESHABILIDADOS PERTENECEN A LAS CATEGORÍAS: ENTRETENIMIENTO, JUEGOS, NOTICIAS Y POLÍTICAS, CIENCIA Y TECNOLOGÍA. ¿EXISTIRÁ ALGUNA RELACIÓN? VEAMOS QUÉ DICE EL MODELO.

Ho: LAS VARIABLES SON INDEPENDIENTES -> NO HAY RELACIÓN: UN VÍDEO PUEDE TENER DESACTIVADOS (O NO) LOS COMENTARIOS INDEPENDIENTEMENTE DE SU CATEOGRÍA.

H1: LAS VARIABLES *NO* SON INDEPENDIENTES -> SÍ HAY RELACIÓN: UN VÍDEO PUEDE TENER DESACTIVADO (O NO) LOS COMENTARIOS PORQUE PERTENECE A DETERMINADA CATEGORÍA.

```
chisq.test(categoryByComentDisabled)
```

```
## Warning in chisq.test(categoryByComentDisabled): Chi-squared
approximation may
## be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  categoryByComentDisabled
## X-squared = 961.84, df = 14, p-value < 2.2e-16
```

```
pvalue <- 2.2e-16
rechazarH0 <- (pvalue<significancia)
rechazarH0
```

```
## [1] TRUE
```

CONCLUSIÓN ANÁLISIS 01

-> SE RECHAZA LA HIPÓTESIS NULA. *NO* EXISTE NINGUNA RELACIÓN ENTRE UN VÍDEO CON LOS COMENTARIOS DESHABILITADOS (O NO) CON RESPECTO A LA CATEGORÍA A LA QUE PERTENECE. SIN EMBARGO, LA GRÁFICA NOS REFLEJÓ QUE ENTRE LOS VÍDEOS QUE TIENEN MÁS COMENTARIOS DESHABILITADOS, PERTENECEN A LA CATEGORÍAS: ENTRETENIMIENTO, JUEGOS, NOTICIAS Y POLÍTICA, CIENCIA Y TECNOLOGÍA.

ANÁLISIS 02

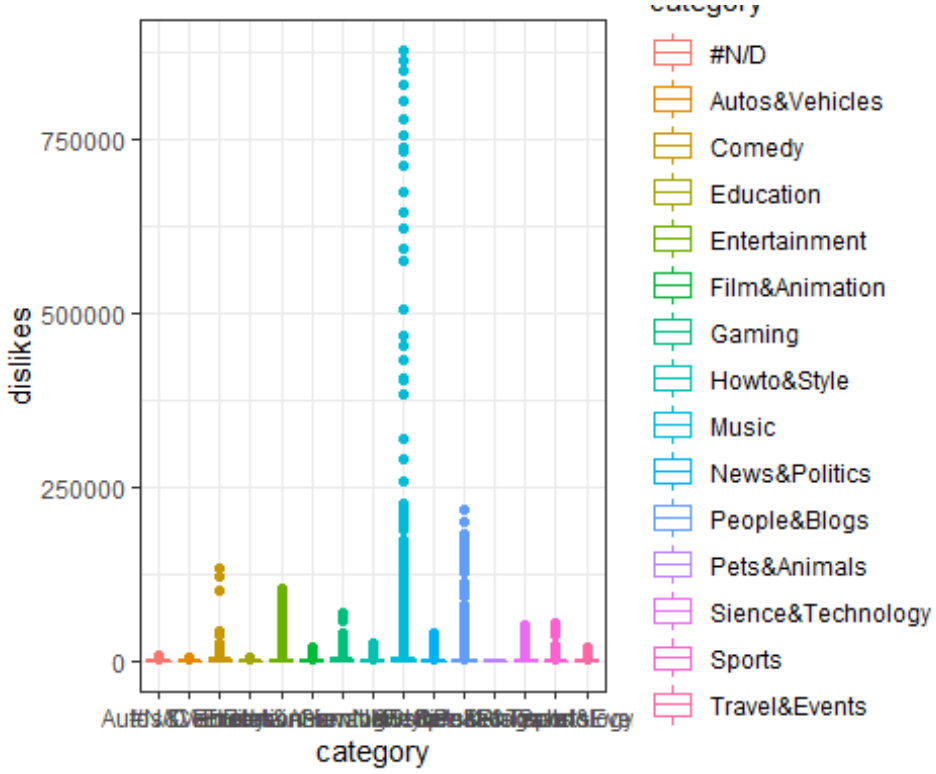
-> ¿EXISTIRÁ ALGUNA DIFERENCIA EN LOS 'DISLIKES' DE LOS VÍDEOS BASADOS EN SU CATEGORÍA?

-> VARIABLES: ‘category’, ‘dislikes’ (cualitativa vs cuantitativa)

-> MODELO: ANOVA

-> GRAFICA: BOXPLOT

```
ggplot(data=ytmx,aes(x=category,y=dislikes,color=category))+geom_boxplot(
)+theme_bw()
```



-> VER GRAFICA CONFUSA 02

SE PUEDE VALIDAR QUE QUE POR LO REGULAR LA CANTIDAD DE DISLIKES ES EL MISMO SIN IMPORTAR LA CATEGORÍA A LA QUE PERTENECEN LOS VÍDEOS. PERO RECORDEMOS QUE ESTAMOS TRABAJANDO EN UNA ESCALA MUY PEQUEÑA, POR LO QUE UNA LIGERA DIFERENCIA EN LAS LÍNEAS DE LA GRÁFICA PODRÍAN REPRESENTAR UNA CANTIDAD NUMÉRICA MAYOR A LA QUE SE PUEDE APRECIAR. TAMBIÉN SE OBSERVA QUE ENTRE LAS CATEGORÍAS QUE PRESENTAN DATOS “ATÍPICOS” MAYORES A LA MEDIA SE ENCUENTRAN: ENTRETENIMIENTO, MÚSICA, NOTICIAS Y POLÍTICA, PERSONAS Y BLOGS. VEAMOS QUÉ NOS DICE EL MODELO.

Ho: NO HAY DIFERENCIA ENTRE MEDIAS DE GRUPOS: LA CANTIDAD DE DISLIKES ES EL MISMO SIN IMPORTAR LA CATEGORÍA

H1: AL MENOS UNO DE LOS GRUPOS ES DIFERENTE: AL MENOS UNA CATEGORÍA PRESENTA UNA CANTIDAD DE DISLIKES DIFERENTE EN LOS VÍDEOS.

```
anova <- aov(ytmx$dislikes~ytmx$category)
summary(anova)
```

```
##                   Df    Sum Sq   Mean Sq F value Pr(>F)
## ytmx$category     14 1.855e+11 1.325e+10   57.65 <2e-16 ***
## Residuals      57784 1.328e+13 2.299e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
pvalue <- 2e-16
significanciaAnova <- 0.001
rechazarH0 <- (pvalue<significanciaAnova)
rechazarH0
```

```
## [1] TRUE
```

CONCLUSIÓN ANÁLISIS 02

-> SE RECHAZA LA HIPÓTESIS NULA. SIGNIFICA QUE AL MENOS UNA CATEGORÍA CUENTA CON UNA CANTIDAD DE DISLIKES EN LOS VÍDEOS DIFERENTE AL RESTO DE LAS CATEGORÍAS. EN ESTE CASO LA GRÁFICA MUESTRA QUE LA CATEOGORÍA QUE DIFIERE BASTANTE CON RESPECTO A LAS DEMÁS CATEGORÍAS Y DONDE EVENTUALMENTE SE ENCUENTRAN VÍDEOS CON MÁS DISLIKES ES 'MÚSICA'. MÁS ADELANTE VEREMOS POR QUÉ ESTA CATEGORÍA PRESENTA MÁS DISLIKES.

ANÁLISIS 03

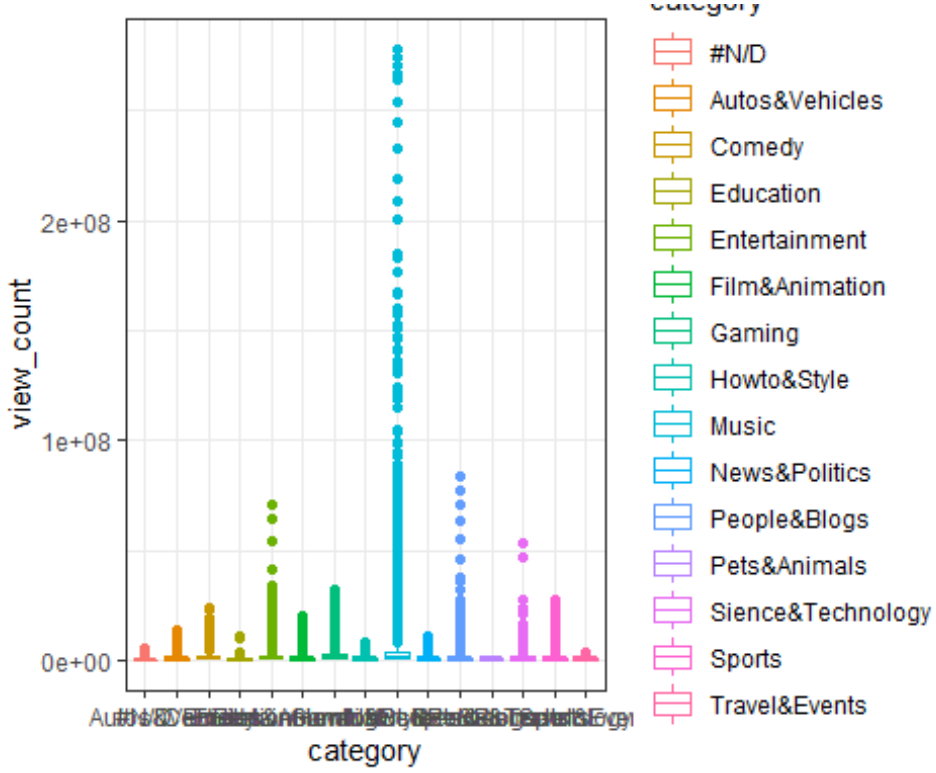
-> ¿EXISTIRÁ ALGUNA DIFERENCIA ENTRE LAS VISITAS DE LOS VÍDEOS BASADOS EN LA CATEGORÍA A LA QUE PERTENECEN? ES DECIR, ¿TODAS LAS CATEGORÍAS POR LO GENERAL TIENEN LAS MISMA CANTIDAD DE VISITAS EN LOS VÍDEOS?

-> VARIABLES: 'category', 'view_count' (cualitativa vs cuantitativa)

-> MODELO: ANOVA

-> GRAFICA: BOXPLOT

```
ggplot(data=ytmx,aes(x=category,y=view_count,color=category))+geom_boxplo
t()+theme_bw()
```



-> VER GRAFICA CONFUSA 03

SE PUEDE VALIDAR QUE DE NUEVO QUE POR LO REGULAR TODAS PRESENTAN LA MISMA CANTIDAD DE VISITAS. Y ENTRE AQUELLAS CATEGORÍAS QUE CUENTAN CON VÍDEOS CON VISITAS MAYORES A LA MEDIA SE ENCUENTRAN: MÚSICA, ENTRETENIMIENTO, JUEGOS, PERSONAS Y BLOGS. VEAMOS QUÉ DICE EL MODELO.

Ho: NO HAY DIFERENCIA ENTRE MEDIAS DE GRUPOS: TODAS LAS CATEGORÍAS TIENEN LA MISMA CANTIDAD DE VISTAS EN LOS VÍDEOS.

H1: AL MENOS UNO DE LOS GRUPOS ES DIFERENTE: AL MENOS UNA CATEGORÍA TIENE UNA CANTIDAD DE VISTAS DIFERENTE EN LOS VÍDEOS.

```
anova <- aov(ytmx$view_count~ytmx$category)
summary(anova)
```

```
##                   Df    Sum Sq   Mean Sq F value Pr(>F)
## ytmx$category     14 9.244e+16 6.603e+15   164.5 <2e-16 ***
## Residuals      57784 2.320e+18 4.015e+13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
pvalue <- 2e-16
significanciaAnova <- 0.001
rechazarH0 <- (pvalue<significanciaAnova)
rechazarH0
```

```
## [1] TRUE
```

CONCLUSIÓN ANÁLISIS 03

-> SE RECHAZA LA HIPÓTESIS NULA. SIGNIFICA QUE AL MENOS UNA CATEOGRÍA PRESENTA UNA CANTIDAD DE VISTAS DIFERENTE EN LOS VÍDEOS. EN ESTE CASO LA GRÁFICA MUESTRA QUE LA CATEGORÍA "MÚSICA" ES LA QUE PRESENTA MÁS VISTAS CON RESPECTO AL RESTO DE CATEGORÍAS. ESTO QUIERE DECIR QUE… (VER ANÁLISIS SIGUIENTE)

ANÁLISIS 04

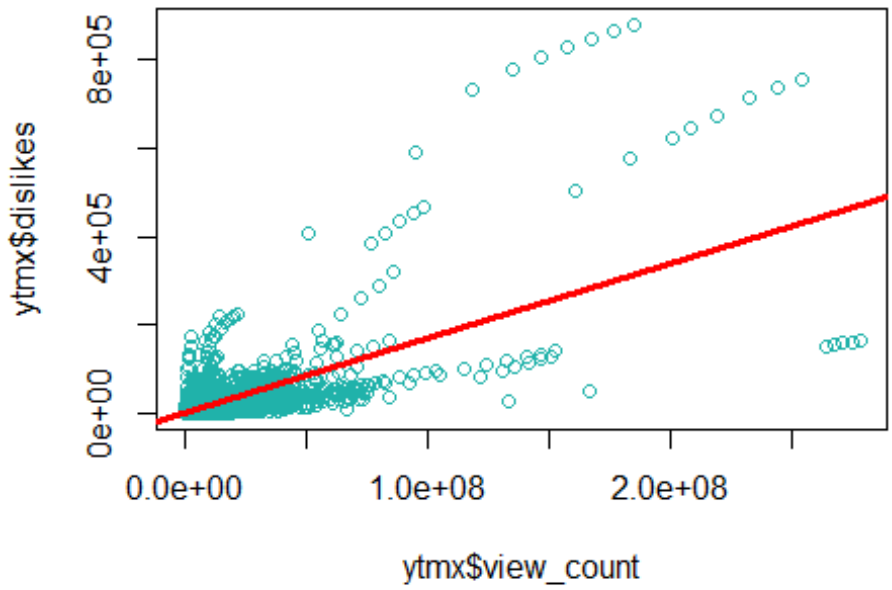
-> ¿EXISTIRÁ ALGUNA RELACIÓN ENTRE LA CANTIDAD DE VISTAS Y LA CANTIDAD DE DISLIKES EN LOS VÍDEOS?

-> VARIABLES: 'view_count', 'dislikes' (cuantitativa vs cuantitativa)

-> MODELO: ANÁLISIS DE CORRELACIÓN

-> GRAFICA: GRÁFICA DE DISPERSIÓN (PLOT)

```
plot(ytmx$view_count,ytmx$dislikes,col='lightseagreen')
abline(lm(ytmx$dislikes~ytmx$view_count), col = 'red', lwd=3)
```



SE PUEDE VALIDAR QUE EL MODELO LINEAL TIENE UN COMPORTAMIENTO POSITIVO ASCENDENTE, LO QUE EVENTUALMENTE QUIERE DECIR QUE A MEDIDA QUE AUMENTA LA VARIABLE INDEPENDIENTE (CANTIDAD DE VISTAS) EVENTUALMENTE AMUENTA TAMBIÉN LA VARIABLE DEPENDIENTE, QUE EN ESTE CASO ES LA CANTIDAD DE DISLIKES EN LOS VÍDEOS. VEAMOS QUÉ NOS DICE EL MODELO.

Ho: LAS 2 VARIABLES SON INDEPENDIENTES: NO HAY RELACIÓN ENTRE LA CANTIDAD DE VISTAS DE UN VÍDEO Y LA CANTIDAD DE DISLIKES QUE ÉSTE TENGA.

H1: LAS 2 VARIABLES *NO* SON INDEPENDIENTES: SÍ HAY RELACIÓN, A MEDIDA QUE AUMENTAN LAS VISTAS EN EL VÍDEO, TAMBIÉN AUMENTARÁ LA CANTIDAD DE DISLIKES.

```
cor(ytmx$view_count,ytmx$dislikes)
```

```
## [1] 0.7204956
```

CONCLUSIÓN ANÁLISIS 04

-> SE RECHAZA LA HIPÓTESIS NULA. COMO SE PUDO VALIDAR GRÁFICAMENTE, EL MODELO LINEAL TIENE UN COMPORTAMIENTO POSITIVO Y ASCENDENTE, EVENTUALMENTE EL MODELO NOS ARROJA UN COEFICIENTE DE CORRELACIÓN (NO PERFECTO) BASTANTE BUENO. POR LO QUE ES VÁLIDO DECIR QUE A MEDIDA QUE UN VÍDEO TENGA MÁS VISTAS, EVENTUALMENTE TENDRÁ UNA CANTIDAD PROPORCIONAL DE DISLIKES. ADICIONAL A ESTO SE LOGRA CORROBORAR QUE LA RAZÓN POR LA CUAL LA CATEGORÍA MÚSICA PRESENTA MÁS DISLIKES ES PORQUE EVENTUALMENTE ES LA CATEGORÍA QUE PRESENTA MÁS VISTAS EN YOUTUBE MÉXICO.

ANÁLISIS 05

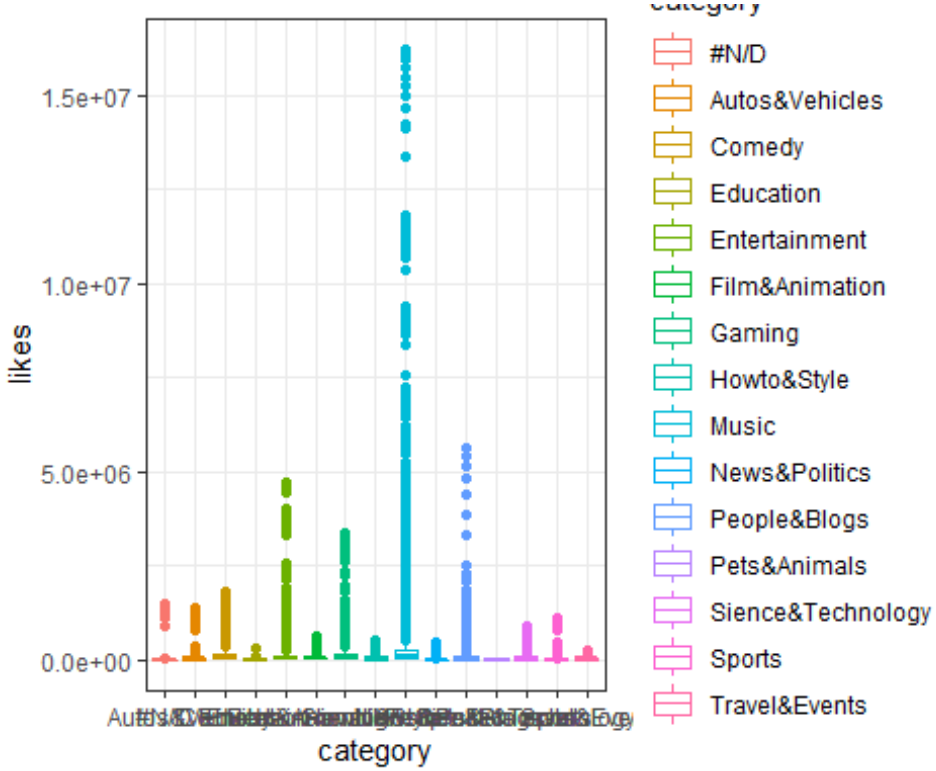
-> AHORA EVALUEMOS LOS LIKES DE LOS VÍDEOS CON RESPECTO A SU CATEOGRÍA. ¿TODOS LAS CATEGORÍAS TENDRÁN LA MISMA CANTIDAD DE LIKES?

-> VARIABLES: ‘category’, ‘likes’ (cualitativa vs cuantitativa)

-> MODELO: ANOVA

-> GRAFICA: BOXPLOT

```
ggplot(data=ytmx,aes(x=category,y=likes,color=category))+geom_boxplot()+t
heme_bw()
```



-> VER GRAFICA CONFUSA 05

SE PUEDE VALIDAR QUE GRÁFICAMENTE CASI TODAS LAS CATEGORÍAS PRESENTAN UNA MISMA CANTIDAD DE LIKES EN LOS VÍDEOS. Y QUE AQUELLAS CATEGORÍAS QUE VARÍAN SON: MÚSCIA, GAMING, ENTRETENIMIENTO, PERSONAS Y BLOGS Y COMEDIA. GRÁFICAMENTE NO TODAS LAS CATEGORÍAS TIENEN LA MISMA CANTIDAD DE LIKES, VEAMOS QUÉ DICE EL MODELO:

Ho: NO HAY DIFERENCIA ENTRE MEDIA DE GRUPOS: TODAS LAS CATEGORÍAS TIENEN LA MISMA CANTIDAD DE LIKES EN LOS VÍDEOS.

H1: AL MENOS UNO DE LOS GRUPOS ES DIFERENTE: AL MENOS UNA CATEGORÍA PRESENTA UNA CANTIDAD DE LIKES EN LOS VÍDOES DIFERENTE A LAS DEMÁS CATEGORÍAS.

```
anova <- aov(ytmx$likes~ytmx$category)
summary(anova)
```

```
##                  Df    Sum Sq   Mean Sq F value Pr(>F)
## ytmx$category    14 5.309e+14 3.792e+13   175.4 <2e-16 ***
## Residuals     57784 1.249e+16 2.162e+11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
pvalue <- 2e-16
significanciaAnova <- 0.001
```

```
rechazarH0 <- (pvalue<significanciaAnova)
rechazarH0

## [1] TRUE
```

CONCLUSIÓN ANÁLISIS 05

-> SE RECHAZA LA HIPÓTESIS NULA. LO QUE SIGNIFICA QUE AL MENOS UNA CATEGORÍA PRESENTA UNA CANTIDAD DE LIKES DIFERENTE A LAS DEMÁS CATEGORÍAS. EN ESTE CASO GRÁFICAMENTE SE PUEDE OBSERVAR QUE ES LA CATEGORÍA 'MÚSICA' QUIEN PRESSENTA UNA CANTIDAD DIFERENTE DE LIKES EN LOS VÍDEOS. ACÁ NOS DAMOS CUENTA QUE GRÁFICAMENTE 'MÚSICA' ES LA CATEGORÍA QUE PRESENTA MÁS CANTIDAD TANTO DE LIKES COMO DE DISLIKES, ASÍ QUE… (VER ANÁLISIS SIGUIENTE)

ANÁLISIS 06

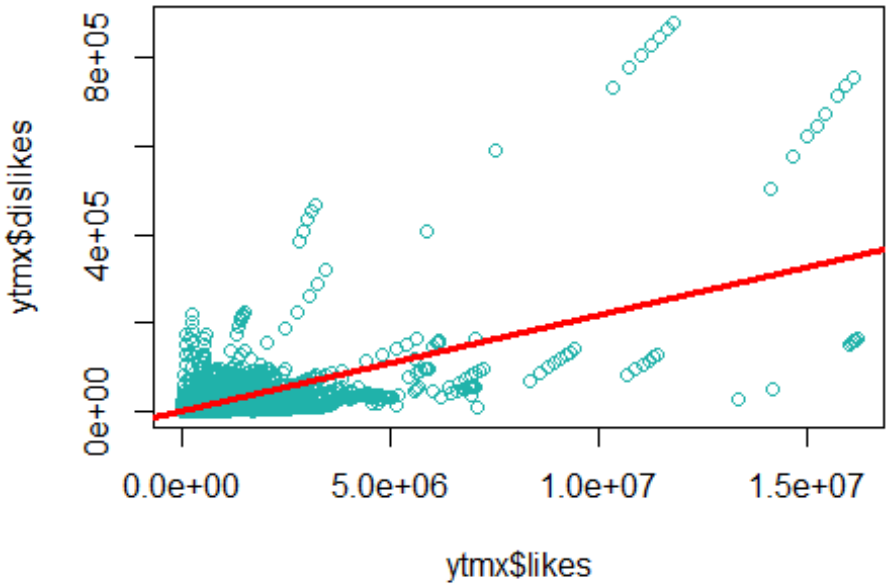
-> ¿HABRÁ ALGUNA RELACIÓN ENTRE LA CANTIDAD DE LIKES CON RESPECTO A LA CANTIDAD DE DISLIKES EN LOS VÍDEOS DE YOUTUBE?

-> VARIABLES: 'LIKES', 'DISLIKES' (cuantitativa vs cuantitativa)

-> MODELO: ANÁLISIS DE CORRELACIÓN

-> GRAFICA: GRÁFICO DE DISPERSIÓN (PLOT)

```
plot(ytmx$likes,ytmx$dislikes,col='lightseagreen')
abline(lm(ytmx$dislikes~ytmx$likes), col = 'red', lwd=3)
```



SE PUEDE VALIDAR QUE EFECTIVAMENTE A MEDIDA QUE AUMENTA LA CANTIDAD DE LIKES EVENTUALMENTE AUMENTA LA CANTIDAD DE DISLIKES EN LOS VÍDEOS, PUES SE OBSERVA TAMBIÉN QUE EL MODELO LINEAL TIENE UN COMPORTAMIENTO POSITIVO ASCENDENTE. GRÁFICAMENTE SÍ HAY RELACIÓN, VEAMOS QUÉ NOS DICE EL MODELO:

Ho: LAS 2 VARIABLES SON INDEPENDIENTES: LA CANTIDAD DE DISLIKES NO DEPENDE DE LA CANTIDAD DE LIKES EN LOS VÍDEOS.

H1: LAS 2 VARIABLES *NO* SON INDEPENDIENTES: LA CANTIDAD DE DISLIKES VARÍA PROPORCIONALMENTE CON RESPECTO A LA CANTIDAD DE LIKES EN LOS VÍDEOS.

```
cor(ytmx$likes,ytmx$dislikes)
```

```
## [1] 0.6796263
```

CONCLUSIÓN ANÁLISIS 06

-> SE RECHAZA LA HIPÓTESIS NULA. EL COEFICIENTE DE CORRELACIÓN NOS INDICA QUE ES VÁLIDO DECIR (POR EJEMPLO) QUE A MEDIDA QUE UN VÍDEO TIENE MÁS LIKES, EVENTUALMENTE TENDRÁ CIERTA CANTIDAD PROPORCIONAL DE DISLIKES EN EL VÍDEO. LO CUAL TIENE SENTIDO YA QUE EN ANÁLISIS POSTERIORES VALIDAMOS GRÁFICAMENTE QUE LA CATEOGORÍA CON MÁS LIKES (MÚSICA) EVENTUALMENTE ES LA CATEGORÍA QUE TAMBIÉN TIENE MÁS CANTIDAD DE DISLIKES CON RESPECTO A LAS OTRAS CATEGORÍAS.

ANÁLISIS 07

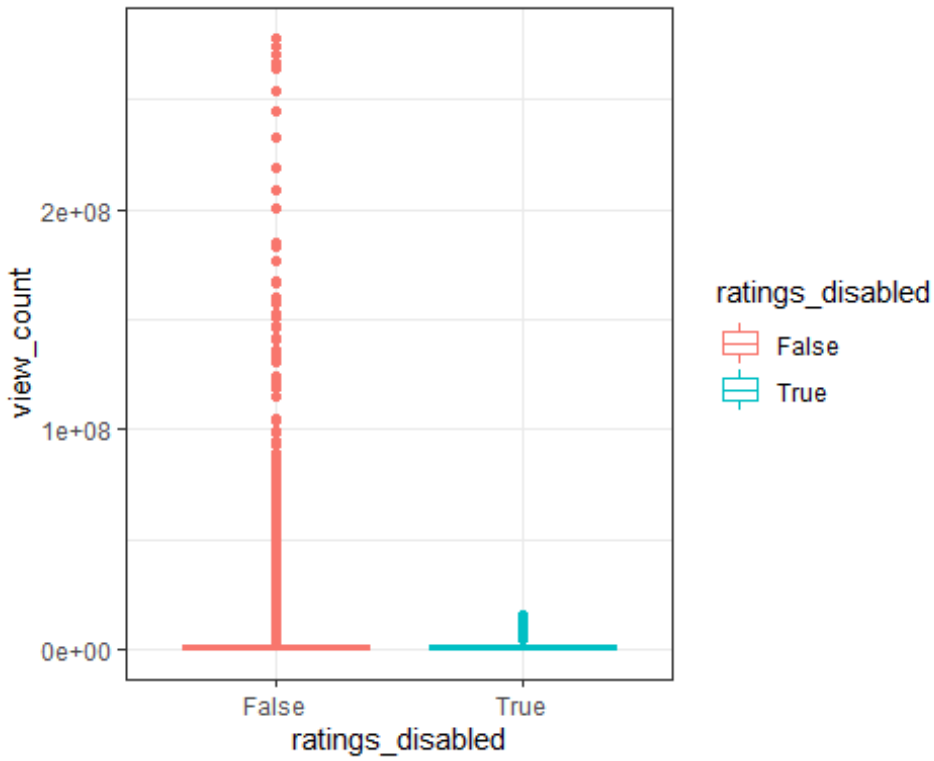
-> ¿EXISTIRÁ DIFERENCIA ENTRE LA CANTIDAD DE VISTAS DE LOS VÍDEOS QUE TIENEN LAS 'CALIFICACIONES DESABILITADAS' CON RESPECTO A LOS QUE NO TIENEN DICHA CARACTERÍSTICA DESHABILITADA?

-> VARIABLES: 'ratings_disabled', 'view_count' (cualitativa vs cuantitativa)

-> MODELO: PRUEBA T

-> GRAFICA: BOXPLOT

```
ggplot(data=ytmx,aes(x=ratings_disabled,y=view_count,color=ratings_disabl
ed))+geom_boxplot()+theme_bw()
```



SE PUEDE VALIDAR GRÁFICAMENTE QUE POR LO GENERAL LAS VISITAS NO VARÍAN, INDEPENDIENTEMENTE SI SE TIENE O NO DESHABILITADA DICHA CARACTERÍSTICA (ratings_disabled). SIN EMBAROG, AQUELLOS QUE *NO* TIENEN DESHABILITADA LAS CALIFICACIONES TIENDEN A TENER MÁS VISTAS, YA QUE HAY GRÁFICAMENTE PODEMOS OBSERVAR QUE AQUELLOS QUE ESTÁN POR ENSIMA DE LA MEDIA POR LO REGULAR PERTENECEN A ESTE GRUPO. VEAMOS QUÉ NOS DICE EL MODELO.

Ho: LA MEDIA DE ratings_disabled (true) == LA MEDIA DE ratings_disabled (false)

H1: LA MEDIA DE ratings_disabled (true) != LA MEDIA DE ratings_disabled (false)

```
t.test(ytmx$view_count~ytmx$ratings_disabled)
```

```
##
##  Welch Two Sample t-test
##
## data:  ytmx$view_count by ytmx$ratings_disabled
## t = 0.37663, df = 157.78, p-value = 0.707
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -392205.5  577024.5
## sample estimates:
## mean in group False  mean in group True
##             2059199             1966789
```

```
pvalue <- 0.707
rechazarH0 <- (pvalue < significancia)
rechazarH0
```

```
## [1] FALSE
```

CONCLUSIÓN ANÁLISIS 07

-> NO SE RECHAZA LA HIPÓTESIS NULA. LO QUE SIGNIFICA QUE LA MEDIA DE VISTAS ES IGUAL, NO IMPORTANDO SI EL CREADOR DE CONTENIDO TIENE O NO LAS CALIFICACIONES DE SUS VÍDEOS DESHABILITADOS. SIN EMBARGO, COMO LO MUESTRA LA GRÁFICA, AQUELLOS DATOS QUE SOBREPASAN LA MEDIA POR LO REGULAR FAVORECEN A AQUELLOS QUE *NO* TIENE DESHABILITADA ESTA CARACTERÍSTICA, ASÍ QUE EN CASO DE PRETENDER LLEGAR A MÁS GENTE, CONVENDRÍA MANTENER DICHA CARACTERÍSTICA HABILITADA PARA TENER LA POSIBILIDAD DE SOBREPASAR LA MEDIA DE VISTAS EN YOUTUBE.

ANÁLISIS 08

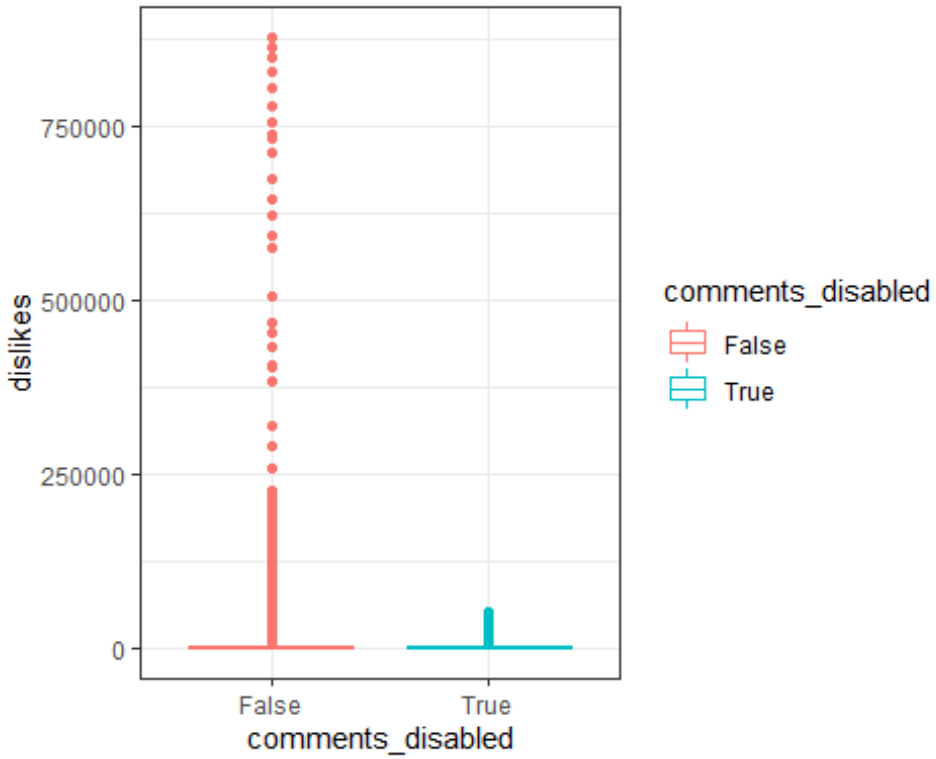
-> ¿HABRÁ ALGUNA DIFERENCIA EN LA CANTIDAD DE DISLIKES DE LOS VÍDEOS QUE TIENEN LOS COMENTARIOS DESHABILITADOS CON RESPECTO A LOS QUE NO TIENEN DICHA CARACTERÍSTICA DESHABILITADA?

-> VARIABLES: ‘comment_disabled’, ‘dislikes’ (cualitativo vs cuantitativo)

-> MODELO: PRUEBA T

-> GRAFICA: BOXPLOT

```
ggplot(data=ytmx,aes(x=comments_disabled,y=dislikes,color=comments_disabl
ed))+geom_boxplot()+theme_bw()
```



SE PUEDE VALIDAR GRÁFICAMENTE QUE POR LO REGULAR SE TIENE LA MISMA CANTIDAD DE DISLIKES, INDEPENDIENTEMENTE SI SE TIENE O NO LOS COMENTARIOS DESHABILITADOS. Y QUE AQUELLOS QUE *NO* TIENEN LOS COMENTARIOS DESHABILITADOS MAYORMENTE PRESENTAN DATOS ATÍPICOS MAYORES A LA MEDIA DE AMBAS CATEGORÍAS. VEAMOS QUÉ DICE EL MODELO.

Ho: MEDIA DE DISLIKES 'comment_disabled' (TRUE) == MEDIA DISLIKES 'comment_disabled' (FALSE)

H1: MEDIA DE DISLIKES 'comment_disabled' (TRUE) != MEDIA DISLIKES 'comment_disabled' (FALSE)

```
t.test(ytmx$dislikes~ytmx$comments_disabled)

##
##  Welch Two Sample t-test
##
## data:  ytmx$dislikes by ytmx$comments_disabled
## t = -3.1954, df = 488.94, p-value = 0.001487
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2140.816  -510.528
## sample estimates:
## mean in group False  mean in group True
##            2936.236            4261.908
```

```
pvalue <- 0.001487
rechazarH0 <- (pvalue < significancia)
rechazarH0

## [1] TRUE
```

CONCLUSIÓN ANÁLISIS 08

-> SE RECHAZA LA HIPÓTESIS NULA. LO QUE EVENTUALMENTE SIGNIFICA QUE LA MEDIA DE DISLIKES DE LOS VÍDEOS QUE TIENEN LOS COMENTARIOS DESHABILITADOS ES DIFERENTE A LA MEDIA DE AQUELLOS QUE NO TIENEN DICHA CARACTERÍSTICA DESHABILITADA. LO CUAL TIENE SENTIDO, YA QUE MAYORMENTE ALGUNOS CREADORES DE CONTENIDO (POLÉMICO) TIENDEN A DESHABILITAR LOS COMENTARIOS Y EVENTUALMENTE OBTIENE MÁS DISLIKES. TAL Y COMO LO ARROJA EL MODELO, YA QUE TAMBIÉN NOS INDICA QUE LA MEDIA DE DISLIKE DE UN VÍDEO CON COMENTARIO DESHABILITADO = 4261.908, MIENTRAS QUE UN LA MEDIA DE UN VÍDEO CON COMENTARIOS QUE *NO* ESTÁN DESHABILITADOS ES DE 2936.236.

ANÁLISIS 09

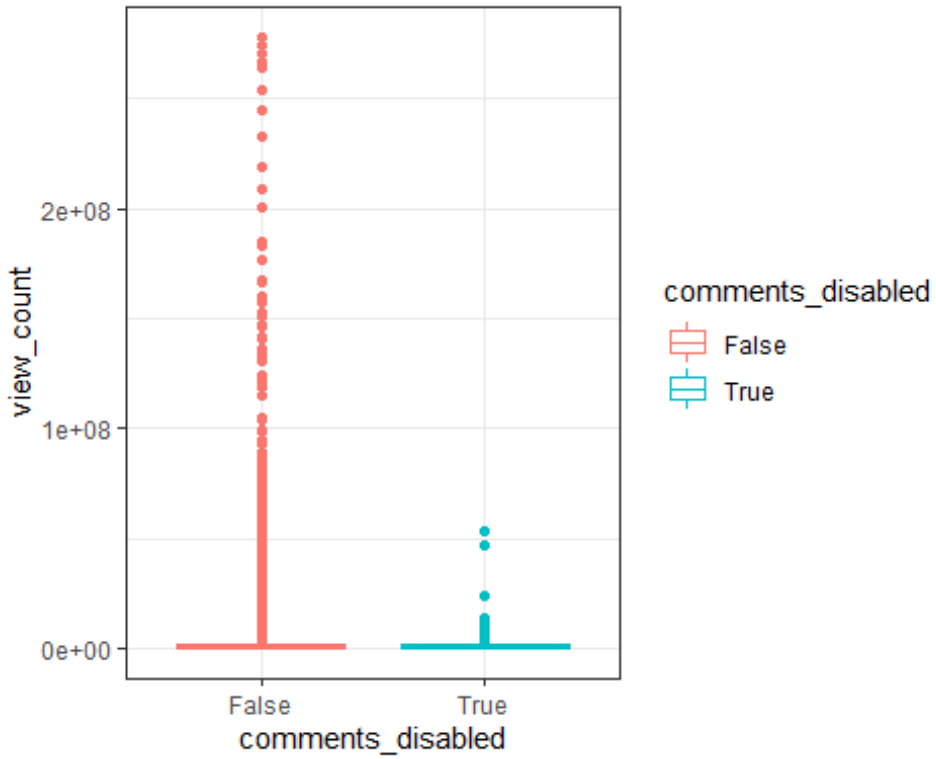
-> SE SABE QUE ALGUNOS CREADORES DE CONTENIDO (POLÉMICO) DESHABILITAN LOS COMENTARIOS PARA OBTENER MÁS VISTAS YA QUE OTROS CREADORES EVENTUALMENTE REACCIONAN A DICHOS VÍDEOS POR *NO* TENER LA POSIBILIDAD DE COMENTAR EN DICHOS VÍDEOS, Y DICHA AUDIENCIA VA A VER EL VÍDEO ORIGINAL. ¿HABRÁ DIFERENCIA EN LAS VISTAS DE LOS VÍDEOS CON COMENTARIOS DESHABILITADOS CON RESPECTO A LOS QUE NO TIENEN DICHA CARACTERÍSTICA DESHABILITADA?

-> VARIABLES: ‘comments_disabled’, ‘view_count’ (cualitativa vs cuantitativa)

-> MODELO: PRUEBA T

-> GRAFICA: BOXPLOT

```
ggplot(data=ytmx,aes(x=comments_disabled,y=view_count,color=comments_disa
bled))+geom_boxplot()+theme_bw()
```



SE PUEDE VALIDAR GRÁFICAMENTE QUE POR LO REGULAR SE TIENE LA MISMA CANTIDAD DE VISTAS EN VIDEOS CON COMENTARIOS TANTO HABILITADOS COMO DESHABILITADOS. Y QUE, COMO HAY MÁS VÍDEOS CON COMENTARIOS HABILITADOS EVENTUALMENTE PODREMOS ENCONTRAR UNA CANTIDAD DE VISTAS MAYOR A LA MEDIA EN DICHOS VÍDEOS. LA GRÁFICA MUESTRA QUE HAY CIERTOS VÍDEOS QUE SOBREPASAN LA MEDIA DE VISTAS PESE A TENER LOS COMENTARIOS DESHABILITADOS (comments_disabled = true). VEAMOS QUE ARROJA EL MODELO…

Ho: MEDIA VISTAS 'comments_disabled' (TRUE) == MEDIA VISTAS 'comments_disabled' (FALSE)

H1: MEDIA VISTAS 'comments_disabled' (TRUE) != MEDIA VISTAS 'comments_disabled' (FALSE)

```
t.test(ytmx$view_count~ytmx$comments_disabled)
```

```
##
##  Welch Two Sample t-test
##
## data:  ytmx$view_count by ytmx$comments_disabled
## t = -0.15583, df = 475.17, p-value = 0.8762
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -581449.7  496004.6
## sample estimates:
## mean in group False  mean in group True
##             2058606             2101328
```

```r
pvalue <- 0.8762
rechazarH0 <- (pvalue<significancia)
rechazarH0
```

```
## [1] FALSE
```

CONCLUSIÓN ANÁLISIS 09

-> NO SE RECHAZA LA HIPÓTESIS NULA. LO QUE SIGNIFICA QUE LA CANTIDAD DE VISTAS EN LOS VÍDEOS ES EL MISMO, INDEPENDIENTEMENTE SI EL CREADOR DE CONTENIDO TIENE O NO DESHABILITADO LOS COMENTARIOS. LA GRÁFICA EVENTUALMENTE MUESTRA QUE EXISTEN MÁS VIDEOS CON COMENTARIOS HABILITADOS QUE SOBREPASAN LA MEDIA DE VISTAS.

-> LAS MEDIAS NO VARÍAN SIGNIFICATIVAMENTE, HE AHÍ DEL PORQUÉ DE LA CONCLUSIÓN DEL MODELO, SIN EMBARGO, SI NOSOTROS HACEMOS UNA DIFERENCIA EN LAS MEDIAS DE VISTAS, TENEMOS QUE…

```r
mediaVistasComentariosDeshabilitados <- 2101328
mediaVistasComentariosHabilitados <- 2058606
diferenciaMedias <- (mediaVistasComentariosDeshabilitados -
mediaVistasComentariosHabilitados)
diferenciaMedias
```

```
## [1] 42722
```

-> UN VÍDEO CON COMENTARIOS DESHABILITADOS TIENE 42722 VISTAS MÁS, CON RESPECTO A LA MEDIA DE AQUELLOS VÍDEOS QUE TIENEN LOS COMENTARIOS HABILITADOS. LO CUAL TIENE SENTIDO YA QUE LOS CREADORES DE CONTENIDO POLÉMICO DESACTIVAN LOS COMENTARIOS PARA OBTENER MÁS VISTAS, Y NO ESTÁN NADA EQUIVOCADOS YA QUE EXISTE EVIDENCIA ESTADÍSTICA DE QUE ESO PUEDE SUCEDER.

ANÁLISIS 10

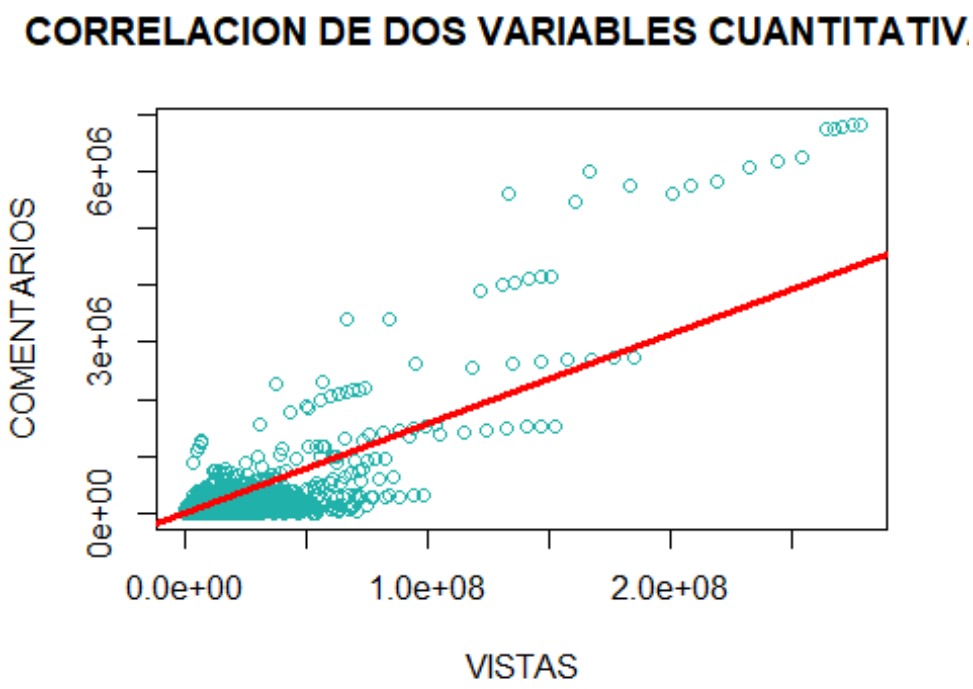
-> ¿EXISTE ALGUNA RELACIÓN ENTRE LA LAS VISTAS DE LOS VÍDEOS Y LA CANTIDAD DE COMENTARIOS QUE LOS USUARIOS DEJAN EN LOS VÍDEOS?

-> VARIABLES: ‘view_count’, ‘comment_count’ (cuantitativa vs cuantitativa)

-> MODELO: ANÁLISIS DE CORRELACIÓN

-> GRAFICA: GRÁFICO DE DISPERSIÓN (PLOT)

```r
plot(ytmx$view_count,ytmx$comment_count, col='lightseagreen',
 main='CORRELACION DE DOS VARIABLES CUANTITATIVAS',
 xlab='VISTAS', ylab = 'COMENTARIOS')
abline(lm(ytmx$comment_count~ytmx$view_count), col = 'red', lwd=3)
```



SE PUEDE VALIDAR GRÁFICAMENTE QUE EXISTE UNA RELACIÓN ENTRE AMBAS VARIABLES, YA QUE EL MODELO LINEAL TIENE UN COMPORTAMIENTO POSITIVO ASCENDENTE. ES DECIR, A MEDID QUE UN VIDEO TENGA MÁS VISTAS EVENTUALMENTE TENDRÁ MÁS COMENTARIOS. VEAMOS QUE NOS DICE EL MODELO…

Ho: LAS 2 VARIABLES SON INDEPENDIENTES: NO HAY RELACIÓN, EL NÚMERO DE COMENTARIOS NO DEPENDE DEL NÚMERO DE VISTAS

H1: LAS 2 VARIABLES *NO* SON INDEPENDIENTES: SÍ HAY RELACIÓN, UN VÍDEO CON MÁS VISTAS EVENTUALMENTE TENDRÁ MÁS COMENTARIOS.

```
cor(ytmx$view_count,ytmx$comment_count)
```

```
## [1] 0.8050224
```

CONCLUSIÓN ANÁLISIS 10

-> SE REHCAZA LA HIPÓTESIS NULA. EL MODELO DE CORRELACIÓN INDICA QUE SÍ EXISTE UNA RELACIÓN ENTRE AMBAS VARIABLES. ESO QUIERE DECIR QUE SI NOSOTROS COMPARAMOS EL NÚMERO DE COMENTARIOS DE DOS VÍDEOS DE YOUTUBE, UNO CON MÁS VISTAS QUE OTRO, EVENTUALMENTE EL VÍDEO CON MÁS VISTAS TENDRÁ MÁS COMENTARIOS. AUNQUE CLARO, NO SE DESCARGA NUNCA LA APARICIÓN DE DATOS ATÍPICOS, YA QUE LA GRÁFICA TAMBIÉN NOS MOSTRÓ QUE EXISTEN CIERTOS VÍDEOS QUE TIENE MÁS COMENTARIOS QUE VISTAS.

ANÁLISIS 11

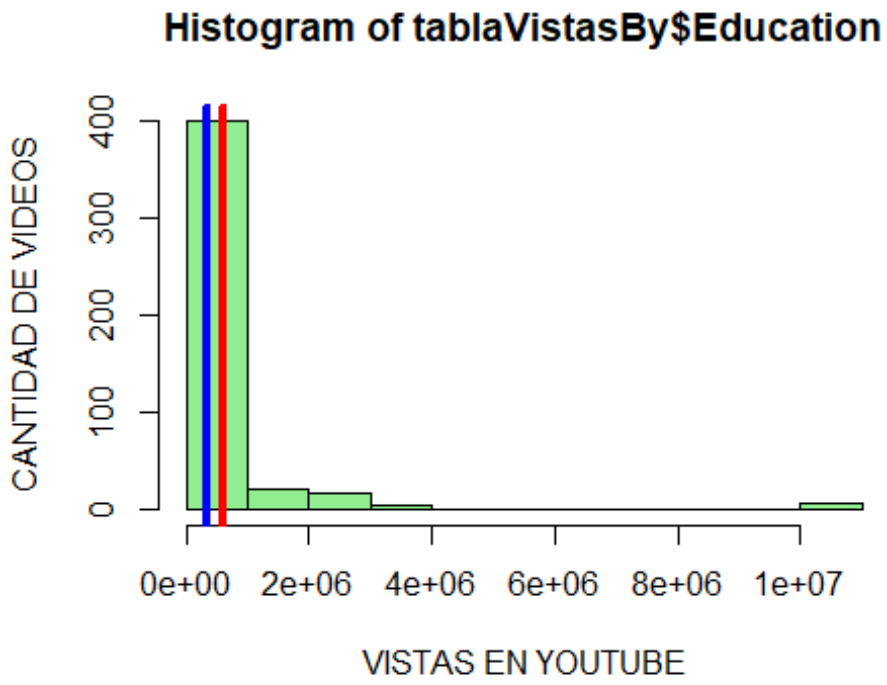
-> ¿CUÁL SERÁ LA DISTRIBUCIÓN DE LAS VISTAS DE LOS VÍDEOS CUYO CONTENIDO ESTÉ CATEGORIZADO COMO 'EDUCATIVO'? ¿HABRÁ NORMALIDAD EN LAS VISTAS DE LOS VÍDEOS RELACIONADOS A LA EDUCACIÓN?

-> VARIABLES: 'view_count', 'category'

-> MODELO: shapiro.test (para validar la normalidad de los datos)

-> GRAFICA: histograma

```
tablaVistasBy <- split(ytmx$view_count,ytmx$category)
hist(tablaVistasBy$Education,col='lightgreen',ylab = 'CANTIDAD DE
VIDEOS',xlab = 'VISTAS EN YOUTUBE')
abline(v = mean(tablaVistasBy$Education), col='red', lwd=4)
abline(v = median(tablaVistasBy$Education), col='blue', lwd=4)
```



SE PUEDE VALIDAR GRÁFICAMENTE QUE *NO* SE TIENE UNA DISTRIBUCIÓN SIMÉTRICA (FORMA ACAMPANADA) SINO MÁS BIEN SE TIENE UNA *DISTRIBUCIÓN SESGADA A LA DERECHA*, YA QUE LA MEDIANA (LÍNEA VERTICAL AZUL) ES MENOR QUE LA MEDIA (LÍNEA VERTICAL ROJA). EN LA GRÁFICA SE OBSERVA QUE EXISTEN ALREDEDOR DE 400 VÍDEOS RELACIONADOS A LA EDUCACIÓN QUE OSCILAN ENTRE 0 Y 800,000 VISTAS. A PARTIR DE DICHO PUNTO, LA CANTIDAD DE VÍDEOS QUE LLEGA A MÁS VISTAS TIENDE A DESCENDER. POR EJEMPLO, SE OBSERVA QUE APROXIMADAMENTE 25 VÍDEOS LLEGAN A 2,000,000 (2e+06) DE VISTAS.

CONVERSION:
https://www.google.com/search?q=2e%2B06+%3D+%3F&oq=2e%2B06+%3D+%3F&aqs=chrome..69i57.5094j0j9&sourceid=chrome&ie=UTF-8

AHORA EVALUEMOS LA NORMALIDAD DE LOS DATOS.

Ho: HAY NORMALIDAD EN LOS DATOS

H1: *NO* HAY NORMALIDAD EN LOS DATOS

```
shapiro.test(tablaVistasBy$Education)

##
##  Shapiro-Wilk normality test
##
## data:  tablaVistasBy$Education
## W = 0.35374, p-value < 2.2e-16

pvalue <- 2.2e-16
rechazarH0 <- (pvalue < significancia)
rechazarH0

## [1] TRUE
```

CONCLUSIÓN ANÁLISIS 11

-> SE RECHAZA LA HIPÓTESIS NULA. EXISTE EVIDENCIA SUFICIENTE QUE INDICA QUE NO HAY NORMALIDAD EN LOS DATOS. LO CUAL PUDIMOS VALIDAR TANTO EN LA GRÁFICA COMO EN EL MODELO. PUES SE CONSIDERA QUE UN CONJUNTO DE DATOS TIENE NORMALIDAD SI LA MEDIA, MEDIANA Y LA MODA EQUIVALEN A LO MISMO. EN ESTE CASO SE CORROBORÓ QUE LA MEDIANA ES MENOR QUE LA MEDIA. LA GRÁFICA TAMBIÉN MUESTRA QUE APROXIMADAMENTE 400 VÍDEOS (COMO MÁXIMO) CUBREN ESTE ANÁLISIS GRÁFICO Y ESTADÍSTICO DE LOS 57,799 REGISTROS DE VÍDEOS CONTENIDOS EN ESTE DATASET. ES DECIR, APROXIMADAMENTE EL 0.69% DE LOS VÍDEOS EN YOUTUBE MÉXICO TRATAN SOBRE EDUCACIÓN. (400*100)/57799

********** DATASET 02

CONTEXTO:

-> DATOS SOBRE TWEETS CON EL HASHTAG #BITCOIN Y #BTC. TWETTER ES LA RED SOCIAL DONDE MÁS SE HABLA DE BITCOIN. DATO INTERESANTE: El 19 de enero de 2021, Elon Musk colocó #Bitcoin en su perfil de Twitter tuiteando “En retrospectiva, era inevitable”, lo que provocó el precio. para subir brevemente alrededor de $ 5000 en una hora a $ 37,299.

LECTURA DEL ARCHIVO

```
btc_tweets <- read.csv('./src/Bitcoin_tweets.csv')
```

ANÁLISIS 12

-> ¿EXISTIRÁ DIFERENCIA ENTRE LA CANTIDAD DE SEGUIDORES DE LOS USUARIOS VERIFICADOS QUE PUBLICAN SOBRE BITCOIN, CON RESPECTO A AQUELLOS QUE PUBLICAN SOBRE BITCOIN PERO QUE NO ESTÁN VERIFICADOS?

-> VARIABLES: ‘user_verified’, ‘user_followers’ (cualitativa vs cuantitativa)

-> MODELO: PRUEBA T

-> GRAFICA: BOXPLOT

```
ggplot(data=btc_tweets,aes(x=user_verified,y=user_followers,color=user_ve
rified))+geom_boxplot()+theme_bw()

## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```

SE PUEDE VALIDAR GRÁFICAMENTE QUE *EXISTE DIFERENCIA EN LA CANTIDAD DE SEGUIRODES* QUE TIENE UN *USUARIO VERIFICADO* QUE PUBLICA SOBRE BITCOIN, RESPECTO A AQUELLOS QUE *NO ESTÁN VERIFICADOS*. VEAMOS QUÉ NOS DICE EL MODELO…

Ho: MEDIA SEGUIDORES USUARIOS VERIFICADOS == MEDIA SEGUIDORES USUARIOS *NO* VERIFICADOS

H1: MEDIA SEGUIDORES USUARIOS VERIFICADOS != MEDIA SEGUIDORES USUARIOS *NO* VERIFICADOS

```
t.test(btc_tweets$user_followers~btc_tweets$user_verified)

##
##  Welch Two Sample t-test
##
## data:  btc_tweets$user_followers by btc_tweets$user_verified
## t = -14.431, df = 626.01, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -331022.8 -251723.2
## sample estimates:
## mean in group False  mean in group True
##            3418.997          294792.024
```

```
pvalue <- 2.2e-16
rechazarH0 <- (pvalue < significancia)
rechazarH0

## [1] TRUE
```

CONCLUSIÓN ANÁLISIS 12

-> SE RECHAZA LA HIPÓTESIS NULA. LO QUE SIGNIFICA DE QUE EXISTE UNA DIFERENCIA ENTRE LA CANTIDAD DE SEGUIDORES DE LOS USUARIOS VERIFICADOS QUE PUBLICAN SOBRE BITCOIN, CON RESPECTO A LOS *NO* VERIFICADOS. SI NOSOTROS LEEMOS LOS REQUISITOS DE VERIFICACIÓN DE TWITTER VEMOS QUE UNO DE LOS REQUISITOS ES "SER UNA PERSONA INFLUYENTE", POR ENDE AQUELLAS PERSONAS INFLUYENTES EVENTUALMENTE TIENDEN A TENER MÁS SEGUIDORES Y FINALMENTE MÁS ATENCIÓN SOCIAL, SIENDO CAPACES INCLUSO DE GENERAR CAMBIOS TAN RADICALES COMO LO ES EL CASO DE ELON MUSK AL CAMBIAR DRÁSTICAMENTE EL VALOR DE UNA CRIPTOMONEDA.

REF -> https://help.twitter.com/es/managing-your-account/about-twitter-verified-accounts

-> https://www.xataka.com/criptomonedas/elon-musk-cambia-su-bio-twitter-para-poner-bitcoin-valor-se-dispara-5-000-dolares-hora

ANÁLISIS 13

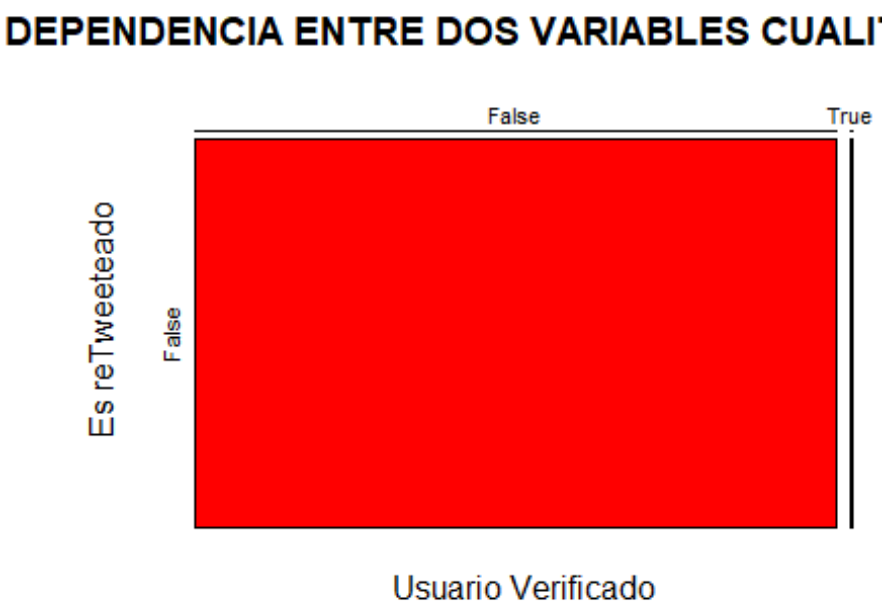
-> ¿EXISTIRÁ ALGUNA RELACIÓN EN SI UN TWEET ES "RETWEETEADO" (POR EL MISMO USUARIO QUE SE AUTENTICA) CON RESPECTO A SI SE TRATA O NO DE UNA CUENTA VERIFICADA?

-> VARIABLES: 'is_retweet', 'user_verified' (cualitativa vs cualitativa)

-> MODELO: CHI CUADRADO

-> GRAFICA: MOSAICO

```
retweetByVerified <-
table(btc_tweets$user_verified,btc_tweets$is_retweet, dnn = c("Usuario Verificado","Es reTweeteado"))
mosaicplot(retweetByVerified, main="DEPENDENCIA ENTRE DOS VARIABLES CUALITATIVAS", col=c("green","red"))
```



SE PUEDE VALIDAR QUE PRÁCTICAMENTE NO SE TIENEN RETWEETS POR PARTE DE LOS USUARIOS QUE PUBLICAN DICHOS TWEETS EN RELACIÓN A BITCOIN, TANTO DE USUARIOS VERIFICADOS COMO DE NO VERIFICADOS. VEAMOS QUE NOS DICE EL MODELO…

Ho: LAS 2 VARIABLES SON INDEPENDIENTES: EL RETWEET NO DEPENDE DE LA VERIFICACIÓN DEL USUARIO.

H1: LAS 2 VARIABLES *NO* SON INDEPENDIENTES: QUE UN TWEET SEA RETWEETEADO POR EL USUARIO QUE SE AUTENTICA PROBABLEMENTE ES PORQUE EL USUARIO ESTÁ VERIFICADO.

```
retweetByVerified

##                   Es reTweeteado
## Usuario Verificado         False
##              False      2 127798
##              True       0    627
```

LA TABLA ACUTALMENTE NOS INDICA QUE TAN SOLO DOS USUARIOS *NO VERIFICADOS* HAY RETWEETEADO SU TWEET. APLICANDO EL MODELO, TENEMOS QUE:

```
chisq.test(retweetByVerified)

## Warning in chisq.test(retweetByVerified): Chi-squared approximation
may be
## incorrect

##
##  Pearson's Chi-squared test with Yates' continuity correction
```

```
##
## data:  retweetByVerified
## X-squared = 8.7852e-22, df = 1, p-value = 1
```

CONCLUSIÓN ANÁLISIS 13

-> NO SE RECHAZA LA HIPÓTESIS NULA. QUE UN TWEET SEA O NO RETWEETEADO, NO TIENE NINGUNA RELACIÓN CON RESPECTO A SI EL USUARIO ESTÁ O NO VERIFICADO. ADICIONAL A ESTO, LA GRÁFICA DE MOSAICO NOS INDICA QUE CASI TODOS LOS USUARIOS QUE PUBLICAN SOBRE BITCOIN *NO* HACEN UN RETWEET DE SU PROPIO *TWEET*.

ANÁLISIS 14

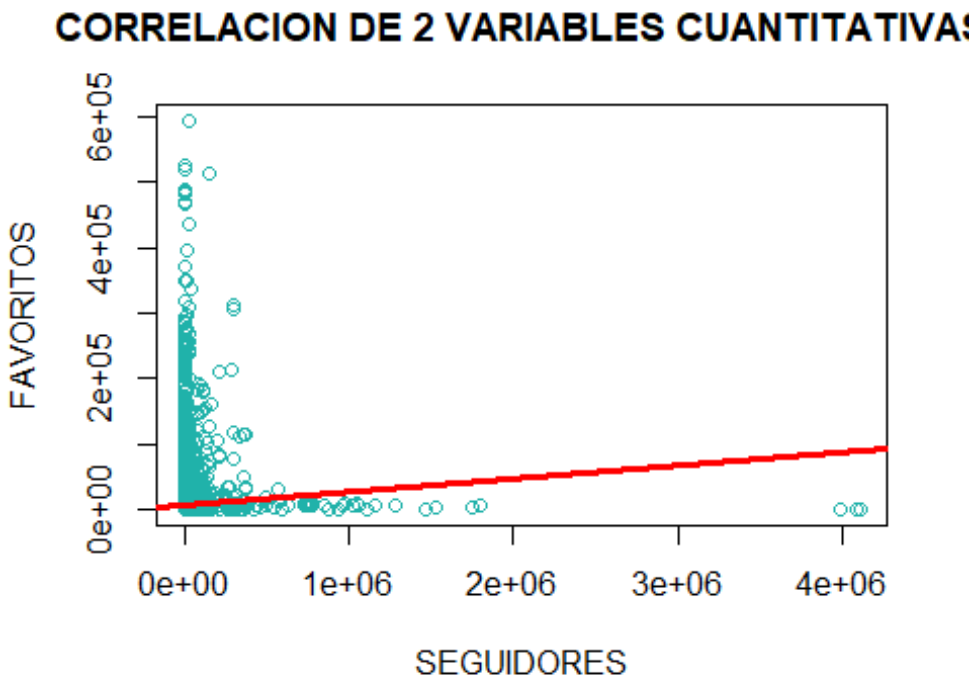
-> ¿EXISTIRÁ ALGUNA RELACIÓN ENTRE LA CANTIDAD DE FAVORITOS DE LOS USUARIOS QUE PUBLICAN SOBRE TWEETER CON RESPECTO A LA CANTIDAD DE SEGUIDORES QUE ELLOS TIENEN?

-> VARIABLES: 'user_followers', 'user_favourites' (cuantitativa vs cuantitativa)

-> MODELO: CORRELACIÓN DE PEARSON

-> GRAFICA: DIAGRAMA DE DISPERSIÓN (PLOT)

```
plot(btc_tweets$user_followers,btc_tweets$user_favourites
,col='lightseagreen',
 main='CORRELACION DE 2 VARIABLES CUANTITATIVAS',
 xlab='SEGUIDORES', ylab = 'FAVORITOS')
abline(lm(btc_tweets$user_favourites~btc_tweets$user_followers), col =
'red', lwd=3)
```



SE PUEDE VALIDAR GRÁFICAMENTE QUE EL MODELO LINEAL TIENE UN COMPORTAMIENTO POSITIVO ASCENDENTE, LO QUE FINALMENTE SIGNIFICA QUE SÍ EXISTE UNA RELACIÓN ENTRE AMBAS VARIABLES, VEAMOS QUE NOS DICE EL MODELO.

Ho: LAS 2 VARIABLES SON INDEPENDIENTES. NO HAY RELACIÓN: LA CANTIDAD DE FAVORITOS DEL USUARIO NO DEPENE DE LA CANTIDAD DE SEGUIDORES QUE TENGA.

H1: LAS 2 VARIABLES *NO* SON INDEPENDIENTES. SÍ HAY RELACIÓN: UN CIERTO NÚMERO DE SEGUIDORES POR LO REGULAR TAMBIÉN SIGNIFICA QUE EL USUARIO TENGA CIERTO NÚMERO DE FAVORITOS.

```
cor.test(btc_tweets$user_followers,btc_tweets$user_favourites,method=c("p
earson"))
```

```
##
##  Pearson's product-moment correlation
##
## data:  btc_tweets$user_followers and btc_tweets$user_favourites
## t = 17.34, df = 128424, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.04287144 0.05378426
## sample estimates:
##       cor
## 0.04832929
```

```
pvalue <- 2.2e-16
rechazarH0 <- (pvalue < significancia)
rechazarH0
```

```
## [1] TRUE
```

CONCLUSIÓN ANÁLISIS 14

-> SE RECHAZA LA HIPÓTESIS NULA. LAS VARIABLES *NO* SON INDEPENDIENTES, LO QUE EVENTUALMENTE SIGNIFICA QUE SÍ EXISTE CIERTA RELACIÓN ENTRE LA CANTIDAD DE SEGUIDORES CON RESPECTO A LA CANTIDAD DE FAVORITOS QUE EL USUARIO TIENE EN SU CUENTA DE TWITTER.

ANÁLISIS 15

-> ¿EXISTIRÁ ALGUNA RELACIÓN ENTRE LA CANTIDA DE AMIGOS Y LA CANTIDAD DE FAVORITOS DE UN USUARIO QUE PUBLICA EN TWITTER SOBRE BITCOIN?

-> VARIABLES: 'user_friends', 'user_favourites' (2 variables cuantitativas)

-> MODELO: CORRELACIÓN DE PEARSON

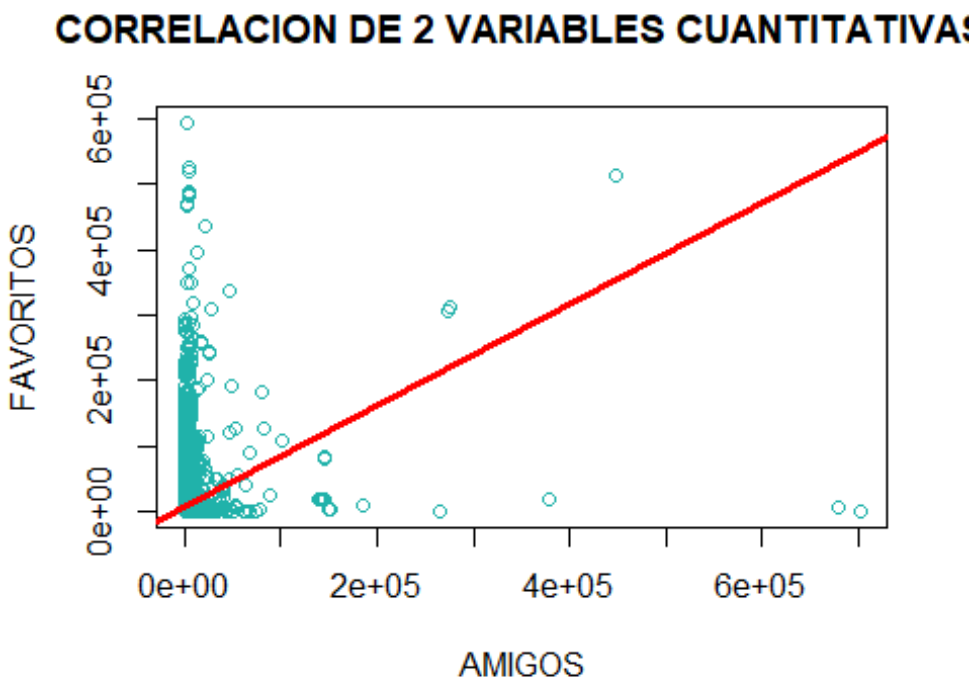
-> GRAFICA: DIAGRAMA DE DISPERSIÓN (PLOT)

ES NECESARIO CONVERTIR DICHA VARIABLE A NUMERIC, YA QUE EL DATAFRAME LO TRAE COMO STRING O CHAR.

```
NumUser_friends <- as.numeric(btc_tweets$user_friends)
```

```
## Warning: NAs introduced by coercion
```

```
plot(NumUser_friends,btc_tweets$user_favourites,col='lightseagreen',
 main='CORRELACION DE 2 VARIABLES CUANTITATIVAS',
 xlab='AMIGOS', ylab = 'FAVORITOS')
abline(lm(btc_tweets$user_favourites~NumUser_friends), col = 'red',
lwd=3)
```



SE PUEDE VALIDAR GRÁFICAMENTE QUE EXISTE UN RELACIÓN ENTRE AMBAS VARIABLES YA QUE EL MODELO LINEAL TIENE UN COMPORTAMIENTO POSITIVO ASCENDENTE, LO QUE EVENTUALMENTE SIGNIFICA QUE A MEDIDA QUE LOS USUARIOS DE TWITTER QUE PUBLICAN SOBRE BITCOIN TIENE MÁS AMIGOS, EVENTUALMENTE TIENEN MÁS USUARIOS COMO FAVORITOS. VEAMOS QUÉ DICE EL MODELO:

Ho: LAS 2 VARIABLES SON INDEPENDIENTES -> NO HAY RELACIÓN.

H1: LAS 2 VARIABLES *NO* SON INDEPENDIENTES -> SÍ HAY RELACIÓN.

```
cor.test(NumUser_friends,btc_tweets$user_favourites,method=c("pearson"))
```

```
##
##  Pearson's product-moment correlation
##
## data:  NumUser_friends and btc_tweets$user_favourites
## t = 83.729, df = 128423, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2223233 0.2326955
## sample estimates:
##       cor
## 0.2275158
```

```
pvalue <- 2.2e-16
rechazarH0 <- (pvalue < significancia)
rechazarH0
```

```
## [1] TRUE
```

CONCLUSIÓN ANÁLISIS 15

-> SE RECHAZA LA HIPÓTESIS NULA, LO QUE EVENTUALMENTE SIGNIFICA QUE HAY RELACIÓN ENTRE AMBAS VARIABLES. A MEDIDA QUE DICHOS USUARIOS TIENEN MÁS AMIGOS EVENTUALMETNE TIENEN MÁS REGISTROS MARCADOS COMO "FAVORITOS".

********** DATASET 03

CONTEXTO: CONTEO DE TODAL DE VOTOS DE LOS CANDIDATOS PRESIDENCIALES DE ESTADOS UNIDOS 2020, TOMANDO COMO REFERENCIA EL ESTADO Y EL PARTIDO POLÍTICO AL QUE REPRESENTAN.

LECTURA DE ARCHIVO

```
county <- read.csv('./src/president_county_candidate.csv')
summary(county)
```

```
##     state              county           candidate            party
##  Length:32177       Length:32177       Length:32177       Length:32177
##  Class :character   Class :character   Class :character   Class
:character
##  Mode  :character   Mode  :character   Mode  :character   Mode
:character
##
##
##
##   total_votes          won
##  Min.   :      0   Length:32177
##  1st Qu.:      3   Class :character
##  Median :     34   Mode  :character
##  Mean   :   4960
##  3rd Qu.:    745
##  Max.   :3028885
```

ANÁLISIS 16

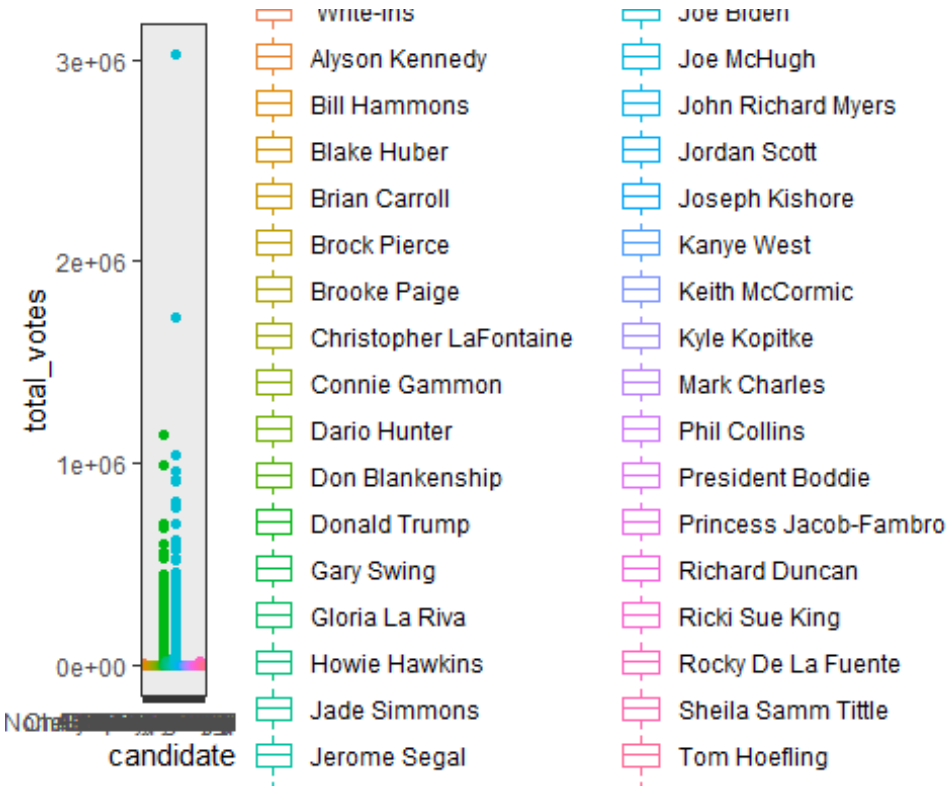
-> ¿EXISTE DIFERENCIA EN EL TOTAL DE VOTOS DE LOS CANDIDATOS?

-> VARIABLES: ‘candidate’, ‘total_votes’ (cualitativa vs cuantitativa)

-> MODELO: ANOVA

-> GRAFICA: BOXPLOT

```
ggplot(data=county,aes(x=candidate,y=total_votes,color=candidate))+geom_b
oxplot()+theme_bw()
```



-> VER GRAFICA CONFUSA 16

SE PUEDE VALIDAR GRÁFICAMENTE QUE PARA “NUESTRA SORPRESA” POR LO GENERA LOS VOTOS SON IGUALES A EXCEPCIÓN DE DONALD TRUMP Y JOE BIDEN. EVIDENTEMENTE LA CANTIDAD DE VOTOS ES DIFERENTE, SIN EMBARGO VEAMOS LO QUE NOS DICE EL MODELO:

Ho: NO HAY DIFERENCIAS ENTRE MEDIAS DE GRUPOS; TODOS LOS CANTIDATOS TIENEN LA MISMA CANTIDAD DE VOTOS.

H1: AL MENOS UNO DE LOS GRUPOS ES DIFERENTE; AL MENOS UN CANTIDATO TIENE UNA CATIDAD DE VOTOS DIFERENTE.

```
anova <- aov(county$total_votes~county$candidate)
summary(anova)

##                      Df    Sum Sq   Mean Sq F value Pr(>F)
## county$candidate     37 1.863e+12 5.035e+10   41.06 <2e-16 ***
## Residuals         32139 3.941e+13 1.226e+09
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

pvalue <- 2e-16
significanciaAnova <- 0.001
rechazarH0 <- (pvalue<significanciaAnova)
rechazarH0

## [1] TRUE
```

CONCLUSIÓN ANÁLISIS 16

-> SE RECHAZA LA HIPÓTESIS NULA, ESTO SIGNIFICA QUE AL MENOS UN CANDIDATO TIENE UNA CANTIDAD DISTINTA DE TOTAL DE VOTOS CON RESPECTO A LOS OTROS CANDIDATOS. LO CUAL ES REALMENTE CIERTO YA QUE COMO BIEN SABEMOS LOS QUE ENCABEZABAN LA LISTA DE CANDIDATOS ERAN: DONALD TRUMP Y JOE BIDEN, TAL Y COMO LO MOSTRÓ LA GRÁFICA TAMBIÉN ANTERIORMENTE.

ANÁLISIS 17

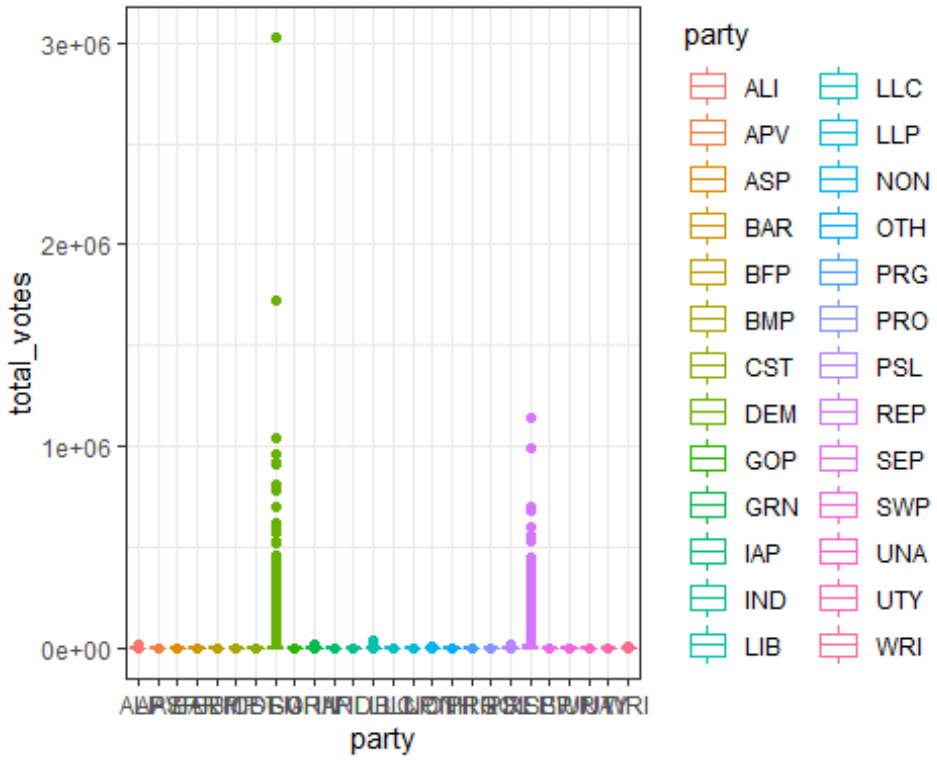
-> ¿EXISTIRÁ DIFERENCIA EN EL TOTAL DE VOTOS BASADO EN EL PARTIDO POLÍTICO DE LOS CANDIDATOS?

-> VARIABLES: 'party', 'total_votes' (cualitativa vs cuantitativa)

-> MODELO: ANOVA

-> GRAFICA: BOXPLOT

```
ggplot(data=county,aes(x=party,y=total_votes,color=party))+geom_boxplot()
+theme_bw()
```



-> VER GRAFICA CONFUSA 17

SE PUEDE VALIDAR GRÁFICAMENTE QUE POR LO GENERAL LOS PARTIDOS POLÍTICOS TIENEN LA MISMA CANTIDAD DE VOTOS, A EXCEPCIÓN IGUALMENTE DEL PARTIDO REPUBLICANDO Y DEMÓCRATA, AL CUAL PERTENECE DONALD TRUMP Y JOE BIDEN, RESPECTIVAMENTE. POR LO CUAL TIENE SENTIDO CON RESPECTO AL ANÁLISIS ANTERIOR.

https://www.google.com/search?q=a+que+partido+politico+pertenece+donald+trump&oq=a+que+partido+politico+pertenece+donald+trump&aqs=chrome..69i57.9064j0j7&sourceid=chrome&ie=UTF-8

https://www.google.com/search?q=a+que+partido+politico+pertenece+joe+biden&oq=a+que+partido+politico+pertenece+joe+biden&aqs=chrome..69i57.3230j0j7&sourceid=chrome&ie=UTF-8

Ho: NO HAY DIFERENCIAS ENTRE MEDIAS DE GRUPOS; TODOS LOS PARTIDOS POLÍTICOS TIENEN LA MISMA CANTIDAD DE VOTOS.

H1: AL MENOS UNO DE LOS GRUPOS ES DIFERENTE; AL MENOS UN PARTIDO POLÍTICO TIENE UNA CATIDAD DE VOTOS DIFERENTE.

```
anova <- aov(county$total_votes~county$party)
summary(anova)
```

```
##                  Df    Sum Sq   Mean Sq F value Pr(>F)
## county$party     25 1.863e+12 7.451e+10   60.79 <2e-16 ***
## Residuals    32151 3.941e+13 1.226e+09
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

pvalue <- 2e-16
significanciaAnova <- 0.001
rechazarH0 <- (pvalue<significanciaAnova)
rechazarH0

## [1] TRUE
```

CONCLUSIÓN ANÁLISIS 17

-> SE RECHAZA LA HIPÓTESIS NULA, LO QUE SIGNIFICA QUE AL MENOS UN PARTIDO POLÍTICO TIENE UNA CANTIDAD DE VOTOS DIFERENTE, EN ESTE CASO GRÁFICAMENTE SE PUDO CORROBORAR QUE LOS PARTIDOS POLÍTICOS CON UNA CANTIDAD DE VOTOS SON EL PARTIDO DEM (DEMÓCRATA) Y EL REP (REPUBLICANO).

ANÁLISIS 18

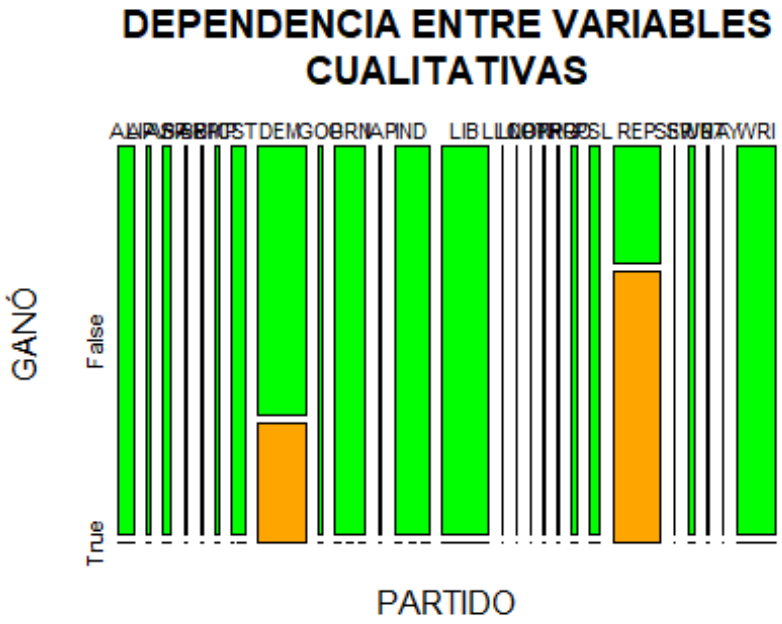
-> ¿EXISTIRÁ ALGUNA RELACIÓN ENTRE EL PARTIDO POLÍTICO CON RESPECO A SÍ GANÓ O NO POR VOTOS LAS ELECCIONES?

-> VARIABLES: ‘won’, ‘party’ (2 variables cualitativas)

-> MODELO: CHI CUADRADO

-> GRAFICA: MOSAICO

```
wonByParty <- table(county$party,county$won, dnn = c("PARTIDO","GANÓ"))
mosaicplot(wonByParty, main="DEPENDENCIA ENTRE VARIABLES
CUALITATIVAS", col=c("green","orange"))
```



-> VER GRAFICA CONFUSA 18

SE PUEDE VALIDAR GRÁFICAMENTE QUE AQUELLOS PARTIDOS QUE GANARON “MÁS VECES” SON EL PARTIDO REPUBLICANO Y EL DEMÓCRATA. “MÁS VECES” PORQUE ES UN REGISTRO DE VOTOS BASADO EN LOS ESTADOS -> CONDADOS DE ESTADOS UNIDOS.

Ho: LAS 2 VARIABLES SON INDEPENDIENTES: NO HAY RELACIÓN, UN PARTIDO GANÓ LAS ELECCIONES EN LOS ESTADOS, NO PORQUE HAYA SIDO {X} PARTIDO.

H1: LAS 2 VARIABLES *NO* SON INDEPENDIENTES: SÍ HAY RELACIÓN, LAS VICTORIAS DE ELECCIONES PRESIDENCIALES EN LOS ESTADOS -> CONDADOS DEPENDE DEL PARTIDO POLÍTICO.

```
chisq.test(wonByParty)
```

```
## Warning in chisq.test(wonByParty): Chi-squared approximation may be
incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  wonByParty
## X-squared = 16215, df = 25, p-value < 2.2e-16
```

```
pvalue <- 2.2e-16
rechazarH0 <- (pvalue<significancia)
rechazarH0
```

```
## [1] TRUE
```

CONCLUSIÓN ANÁLISIS 18

-> SE RECHAZA LA HIPÓTESIS NULA. LO QUE FINALMENTE SIGNIFICA QUE LAS 2 VARIABLES *NO* SON INDEPENDIENTES, ES DECIR, SÍ EXISTE UNA RELACIÓN ENTRE EL PARTIDO POLÍTICO Y EN EL CRITERIO DE QUE SI GANÓ O NO LAS ELECCIONES. ESTO PROBABLEMENTE SE DEBA A FACTORES EXTERNOS, PRINCIPALMENTE LA INVERSIÓN EN CAMPAÑAS POLÍTICAS POR PARTE DE LOS CANDIDATOS PRESIDENCIALES.

ANÁLISIS 19

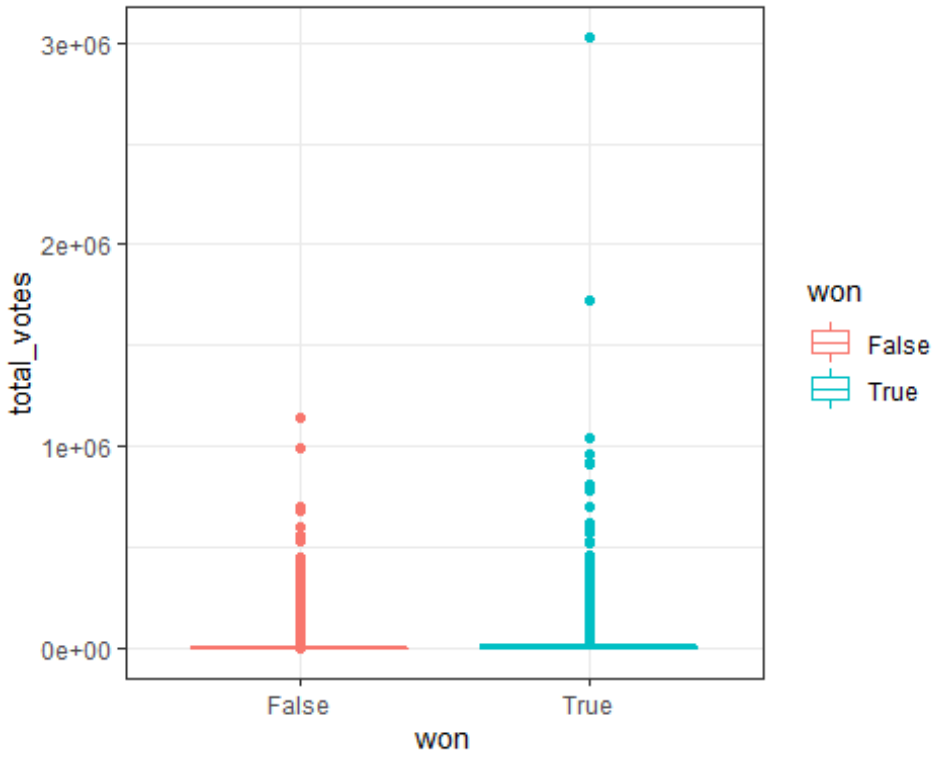
-> LÓGICAMENTE LOS PARTIDOS QUE GANARON LAS ELECCIONES EVENTUALMENTE TENDRÍAN QUE TENER MÁS VOTOS. ENTONCES. VALIDAR SI EXISTE DIFERENCIA EN EL TOTAL DE VOTOS DE LAS ELECCIONES PRESIDENCIALES BASADAS EN SI GANARON O NO LAS ELECCIONES.

-> VARIABLES: 'won', 'total_votes' (cuantitativa vs cualitativa)

-> MODELO: PRUEBA TEST

-> GRAFICA: BOXPLOT

```
ggplot(data=county,aes(x=won,y=total_votes,color=won))+geom_boxplot()+the
me_bw()
```



LA GRÁFICA NO NOS MUESTRA UNA DIFERENCIA SIGNIFICANTE, SIN EMBARGO, RECORDEMOS QUE SON DATOS A GRAN ESCALA, ES DECIR, ESTAMOS HABLANDO DE MILLONES DE VOTOS, POR LO QUE VISUALMENTE NO PODRÍAMOS APRECIAR UNA DIFERENCIA, ASÍ QUE APLICAMOS EL MODELO POARA VER QUÉ NOS INDICA…

Ho: MEDIA DEL 'total_votes' DE LOS PARTIDOS QUE GANARO == MEDIA DEL 'total_votes' DE LOS PARTIDOS QUE *NO* GANAR

H1: MEDIA DEL 'total_votes' DE LOS PARTIDOS QUE GANARO != MEDIA DEL 'total_votes' DE LOS PARTIDOS QUE *NO* GANAR

```
t.test(county$total_votes~county$won)
```

```
##
##  Welch Two Sample t-test
##
## data:  county$total_votes by county$won
## t = -16.405, df = 4714.6, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -21902.60 -17226.53
## sample estimates:
## mean in group False  mean in group True
##             2143.32            21707.89
```

```
pvalue <- 2.2e-16
rechazarH0 <- (pvalue < significancia)
rechazarH0
```

```
## [1] TRUE
```

CONCLUSIÓN ANÁLISIS 19

-> SE RECHAZA LA HIPÓTESIS NULA. LO QUE EVENTUALMENTE SIGNIFICA QUE EL TOTAL DE VOTOS ES DIFERENTE DE AQUELLOS PARTIDOS QUE GANARON CON RESPECTO A LOS QUE NO GANARON. LO CUAL TIENE SENTIDO LÓGICO YA QUE SI UN PARTIDO GANA ES PORQUE EVENTUALMENTE LOGRÓ ACUMULAR MÁS VOTOS. POR OTRA PARTE, AL APLICAR EL MÓDELO NOS ARROJA LA MEDIA DE CADA GRUPO, SIENDO PARA AQUELLOS PARTIDOS QUE PERDIERON UNA MEDIA DE 2,143.32 VOTOS, MIENTRAS QUE AQUELLOS PARTIDOS QUE GANARON TIENEN UNA MEDIA DE 21,707.89

ANÁLISIS 20

-> ¿CUÁL SERÁ LA DISTRIBUCIÓN DE LOS DATOS DE LOS VOTOS DE DONALD TRUMP? ¿HABRÁ NORMALIDAD EN SUS VOTOS?

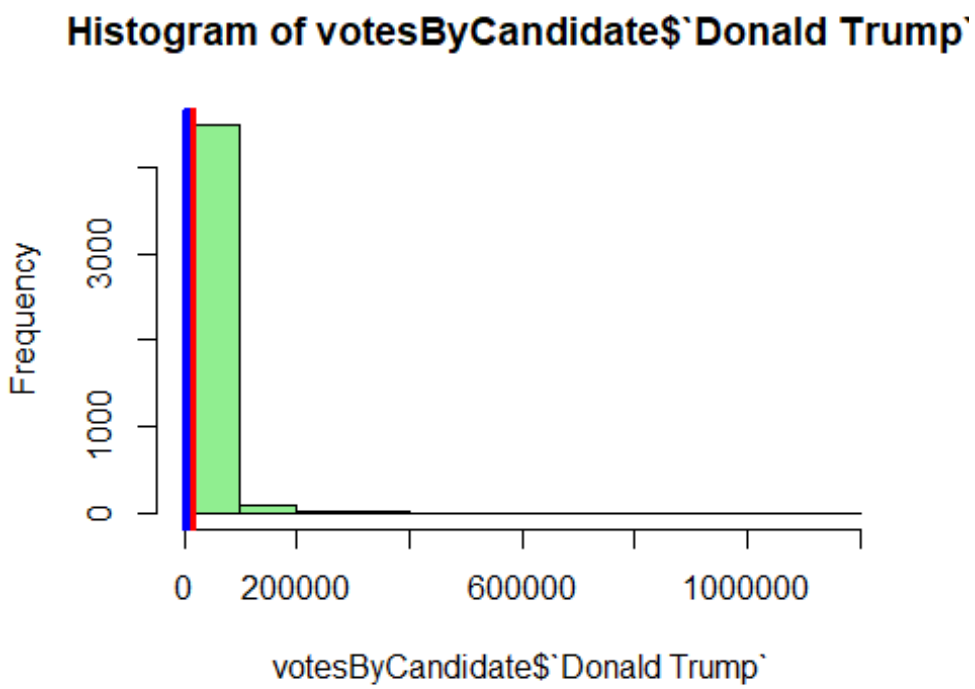
-> MODELO: SHAPIRO (PARA VALIDAR LA NORMALIDAD DE LOS DATOS)

-> GRAFICA: HISTOGRAMA (PARA VALIDAR LA DISTRIBUCIÓN)

```
votesByCandidate <- split(county$total_votes,county$candidate)
```

DONALD TRUMP

```
hist(votesByCandidate$`Donald Trump`,col='lightgreen')
abline(v = mean(votesByCandidate$`Donald Trump`), col='red', lwd=4)
abline(v = median(votesByCandidate$`Donald Trump`), col='blue', lwd=4)
```



SE PUEDE VALIDAR QUE LOS VOTOS DE DONALD TRUMP POR CONDADO OSCILAN ENTRE 0 Y 100,000 VOTOS, A PARTIR DE DICHO PUNTO LOS VOTOS EMPIEZAN A DESCENDER DRÁSTICAMENTE. PUDIENDO INCLUSO LLEGAR A TENER ALREDEDOR DE 200,000 VOTOS EN APROXIMADAMENTE 50 O 100 CONDADOS. TIENE UN COMPORTAMIENTO ASIMÉTRICO SESGADO A LA DERECHA YA QUE LA MEDIANA (LÍNEA VERTICAL AZUL) ES MENOR QUE LA MEDIA (LÍNEA VERTICAL ROJA). GRÁFICAMENTE NO HAY NORMALIDAD EN LOS DATOS, SIN EMBARGO CORROBOREMOSLO A TRAVÉS DEL MODELO.

Ho: HAY NORMALIDAD EN LOS DATOS

H1: *NO* HAY NORMALIDAD EN LOS DATOS

```
shapiro.test(votesByCandidate$`Donald Trump`)

##
##  Shapiro-Wilk normality test
##
## data:  votesByCandidate$`Donald Trump`
## W = 0.3171, p-value < 2.2e-16

pvalue <- 2.2e-16
rechazarH0 <- (pvalue < significancia)
rechazarH0

## [1] TRUE
```

CONCLUSIÓN ANÁLISIS 20

-> SE RECHAZA LA HIPÓTESIS NULA. ES DECIR, SE TIENE EVIDENCIA TANTO GRÁFICA COMO ESTADÍSTICA (A TRAVÉS DEL MODELO) DE QUE LOS VOTOS DE DONALD TRUMPO POR CONDADO, NO TIENEN UN COMPORTAMIENTO NORMAL, SINO UN COMPORTAMIENTO ASIMÉTRICO SESGADO A LA DERECHA.

ANÁLISIS 21

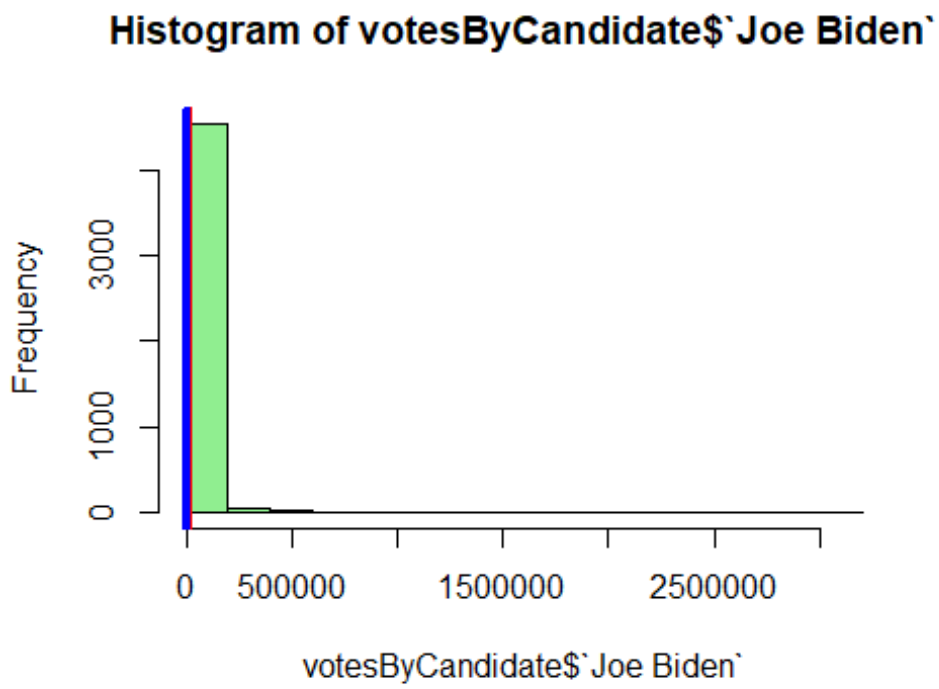
-> ¿CUÁL SERÁ LA DISTRIBUCIÓN DE LOS DATOS DE LOS VOTOS DE JOE VIDEN? ¿HABRÁ NORMALIDAD EN SUS VOTOS?

-> MODELO: SHAPIRO (PARA VALIDAR LA NORMALIDAD DE LOS DATOS)

-> GRAFICA: HISTOGRAMA (PARA VALIDAR LA DISTRIBUCIÓN)

JOE VIDEN

```
hist(votesByCandidate$`Joe Biden`,col='lightgreen')
abline(v = mean(votesByCandidate$`Joe Biden`), col='red', lwd=4)
abline(v = median(votesByCandidate$`Joe Biden`), col='blue', lwd=4)
```



SE PUEDE VALIDAR GRÁFICAMENTE QUE LOS VOTOS NO TIENEN UN COMPORTAMIENTO NORMAL, TAMBIÉN QUE TIENEN UN COMPORTAMIENTO SIMÉTRICO SESGADO A LA DERECHA YA QUE LA MEDIANA ES MENOR QUE LA MEDIA. ADICIONAL A ESTO, ACÁ PODEMOS CORROBORAR POR QUÉ JOE BIDEN GANÓ LA PRESIDENCIA EN ESTADOS UNIDOS, YA QUE SI VOVLEMOS A LA GRÁFICA ANTERIOR DE DONALD TRUMP, LOS VOTOS POR CONDADO OSCILAN ENTRE 0 Y 100,000 VOTOS, MIENTRAS QUE JOE BIDEN ACUMULÓ ENTRE 0 Y APROXIMADAMENTE 250,000 VOTOS POR CONDADO. YA SE SABE QUE NO HAY NORMALIDAD EN LOS DATOS, SIN EMBARGO, HAY QUE CORROBORARLO A TRAVÉS DE UN MODELO.

Ho: HAY NORMALIDAD EN LOS DATOS

H1: *NO* HAY NORMALIDAD EN LOS DATOS

```
shapiro.test(votesByCandidate$`Joe Biden`)

##
##  Shapiro-Wilk normality test
##
## data:  votesByCandidate$`Joe Biden`
## W = 0.1863, p-value < 2.2e-16

pvalue <- 2.2e-16
rechazarH0 <- (pvalue < significancia)
rechazarH0

## [1] TRUE
```

CONCLUSIÓN ANÁLISIS 21

-> SE RECHAZA LA HIPÓTESIS NULA, LO QUE EVENTUALMENTE SIGNIFICA QUE NO EXISTE NORMALIDAD EN LOS VOTOS A FAVOR DE JOE BIDEN, EVENTUALMENTE SE LOGRÓ CORROBORAR QUE CON BASE EN EL ANÁLISIS ANTERIOR, JOE BIDEN ACUMULÓ MÁS VOTOS POR CONDADO QUE DONALD TRUMP. Y DE QUE AMBOS DATOS EVENTUALMENTE TIENEN UN COMPORTMIENTO ASIMÉTRICO SESGADO A LA DERECHA.