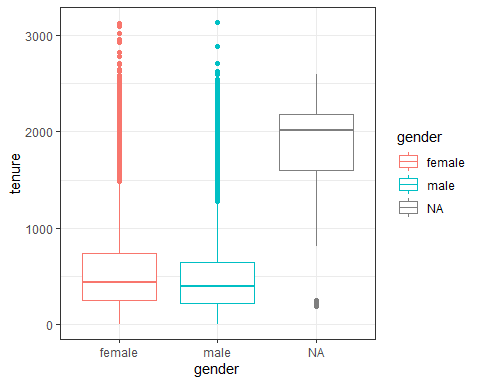
Proyecto\_Final

Emerson Christian Andres Perez Chen 2490-19-13220 Y Marco Ederson Yol Raxcaco 2490-19-19523

31/5/2021

# No.1

#El genero puede ser un factor que altere de manera drastica las oportundades de los   
#usuarios a ser tenderncia en Facebook  
  
#H0: El genero no afecta (Genero = Tendencia)  
#H1: El genero si afecta (Genero =/= Tendencia)  
  
#Grafico (ggplot)  
ggplot(data=Facebook, aes(x=gender,y=tenure, color=gender ))+ geom\_boxplot()+theme\_bw()



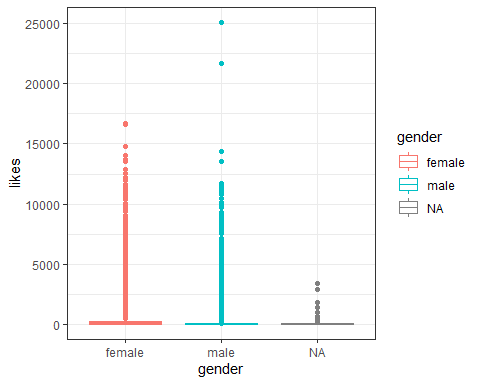
#Metodo (Anova)  
anova<- aov(Facebook$tenure~Facebook$gender)  
summary(anova)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Facebook$gender 1 1.807e+08 180678965 883.4 <2e-16 \*\*\*  
## Residuals 98824 2.021e+10 204525   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 177 observations deleted due to missingness

#P-value = 2e-16  
  
#el P-value es mucho menor a 0.05 esto quiere decir que se rechaza la hipotesis nula esto quiere decir  
#que segun el genero afecta las posibilidades de ser tendencia  
  
#Analisis  
#Segun el analisis anova el genero si afecta en ser tendencia en Facebook pero como se puede ver  
#en la grafica, los usuarios que si dicen su genero (masculino y gemenino) tiene la misma media de tendencia  
#en donde el genero femenino tiene mas datos atipicos que el masculino que significa que hay mas mujeres   
#siendo tendencia por distintas cuestriones que los hombres y tambien podemos notar que los usuaio que no definieron  
#su genero tiene una mayor media de ser tendencia esto significa que muchos usuasios sin definir genero en su perfil   
#tienen a ser mas tendencia e igual tienen datos atipicos bajos en donde se ve que no siempre se es tendencia con ese  
#factor  
  
#segun nuestros datos de investigacion  
#https://www.euskadi.eus/contenidos/noticia/liburua\_sexismoa\_gazteak\_7/es\_def/adjuntos/sexismo\_gizarte\_sareetan\_c.pdf  
#Este comportamiento es generado gracias al miedo o desigualdad que existe en las redes sociales ya que las personas  
#pueden presentar en redes sociales ya que al tener mas informacion en su perfil pueden ser un mayor blanco de burlas o   
#molestias de los demas usuarios y a diferencia de los que no comparten la informacion de su genero se sienten mas libres   
#con respecto a dar sus opiniones que pueden llegar a ser mas tendencia a diferencia de los que si definiron su genero  
#pero como en todo no siempre es asi ya que tambien se dice que las mujeres tienen a dar mas su opinion o publicar mas a   
#diferencia de los hombres asi que hay mas datos atipicos en mujeres que en hombres a pesar de tener una media que rosa lo  
#similar

# No.2

#En Facebok el genero es un factor que puede alterar los likes que da un usuario  
  
#H0: El genero no afecta (Genero = likes)  
#H1: El genero si afecta (Genero =/= likes)  
  
  
#Grafico (ggplot)  
ggplot(data=Facebook, aes(x=gender,y=likes, color=gender ))+ geom\_boxplot()+theme\_bw()



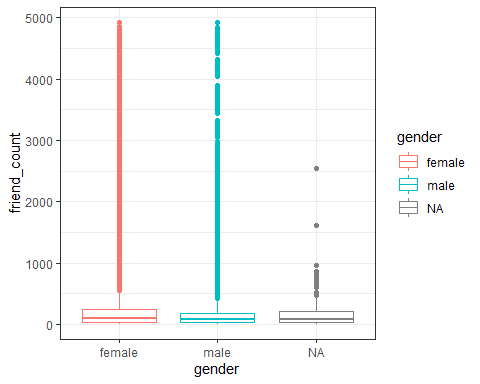
#Metodo (Anova)  
anova<- aov(Facebook$likes~Facebook$gender)  
summary(anova)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Facebook$gender 1 7.338e+08 733772630 2290 <2e-16 \*\*\*  
## Residuals 98826 3.166e+10 320390   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 175 observations deleted due to missingness

#P-value = 2e-16  
  
#Al se el P-value mucho menor a 0.05 quiere decir que se rechaza la hipotesis nula que significa que el   
#genero afecta en los likes que realiza un usuario   
  
#como se ve en la grafica casi todos los generos tiene una meedia simirar pero con gran diferencia en datos atipicos  
#se nota que el genero masculino tiene likes mas altos pero menos constantes a diferencia del genero fenenino  
#en donde los likes no son tan altos mero mas constantes   
  
#https://www.vice.com/es/article/zndpjy/por-que-los-hombres-dan-like-literalmente-a-todos-los-perfiles-de-tinder  
#Esto puede deberse a que los gustos y posteos de los usuarios, los hombres ven mas publicaciones de humor en donde   
#se vuelven mas tendencia y como va al publico de humor los hombres destacan en esto a diferencia de las mujeres que   
#son mas reservadas con los gustos que comparte que genere menos likes que puede dar, a diferencia de el analisis anterior  
#los que no definin su genero no realizan muchos likes que es a causa de mantener su privasidad con respecto a quien es

# No. 3

#El genero afecta en la en la cantidad de amigos que conoce esto quiere decir que no son amistades que facebook le recomendo  
#si no que a amistades que el conoce o viven cerca a el usuario que es gracias a facebook  
  
#H0: El genero no afecta (Genero = conteo de amistades)  
#H1: El genero si afecta (Genero =/= conteo de amistades)  
  
#Grafico (ggplot)  
ggplot(data=Facebook, aes(x=gender,y=friend\_count, color=gender ))+ geom\_boxplot()+theme\_bw()



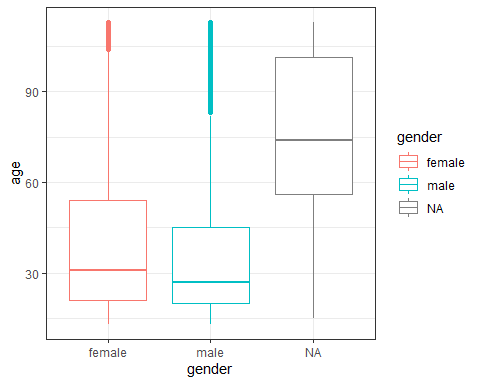
#Metodo (Anova)  
anova<- aov(Facebook$friend\_count~Facebook$gender)  
summary(anova)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Facebook$gender 1 1.412e+08 141213423 949.7 <2e-16 \*\*\*  
## Residuals 98826 1.470e+10 148698   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 175 observations deleted due to missingness

#P-value = 2e-16  
#Al ser el P-value mucho menor a 0.05 se rechaza la hipotesis nunla lo que quiere decir que el genero si  
#afecta en en el conteo de amistades de los usuarios  
  
#La grafica indica que tanto masculino como femenino tiene nos datos similares tanto en media como en los datos atipicos  
#pero hay una peque?a diferencia que entre las mujeres los datos atipicos son mas constantes que los hombres que quiere decir que  
#las mujeres en un minimo caso puede tener mas amistade que los hombre, a gran diferencia de los usuarios sin su genero de definido  
#en donde tiene menos amistades en donde siempre es una constante que quiere tener su privacidad apartada de las   
#redes sociales compartiendo menos informacion con personas o amistades que puede tener

# No.4

#El gener puede afectar en la edad de los usuarios   
  
#H0: El genero no afecta (Genero = conteo de amistades)  
#H1: El genero si afecta (Genero =/= conteo de amistades)  
  
#Grafico (ggplot)  
ggplot(data=Facebook, aes(x=gender,y=age, color=gender ))+ geom\_boxplot()+theme\_bw()



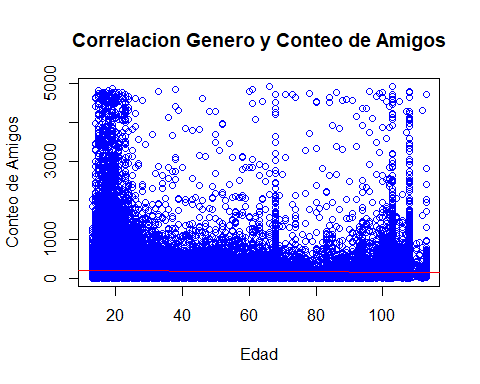
#Metodo (Anova)  
anova<- aov(Facebook$age~Facebook$gender)  
summary(anova)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Facebook$gender 1 342637 342637 679.9 <2e-16 \*\*\*  
## Residuals 98826 49802193 504   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 175 observations deleted due to missingness

#P-value = 2e-16  
  
#al ser e P-value mucho menor a 0.05 quiere decir que se rechaza la hipotesis nula que significa que   
#dependiendo la edad si hay diferencia de edades  
  
#como se ve en la grafica el genero femenino tieneuna media un poco mas alta que los hombres que significa   
#existen mas usuarios feneminos con una edad que ronda los 30 a?os y menos hombres pero tambien se nota que los   
#usuarios sin genero definido hay una mayor cantidad con edades mayores esto puede deberse a que son cuentas en dode  
#queren tener privacidad y no compartin sus datos que causa que coloquen edades mayores o que no corresonden y colocan  
#mas altas para desapersivir a atacantes

# No.5

#La edad puede ser un factor que tiene relacion on el conteo de amigos que se tiene en la red  
#social Facebook  
  
#H0 = Si tiene conrrelacion entre los datos   
#H0 = No tiene correclacion entre los datos  
  
#Grafico(Plot)  
plot(Facebook$age,Facebook$friend\_count, col="Blue", main = "Correlacion Genero y Conteo de Amigos", ylab = "Conteo de Amigos", xlab = "Edad")  
modelolineal<-lm(Facebook$friend\_count~Facebook$age)  
abline(modelolineal,col ="red")



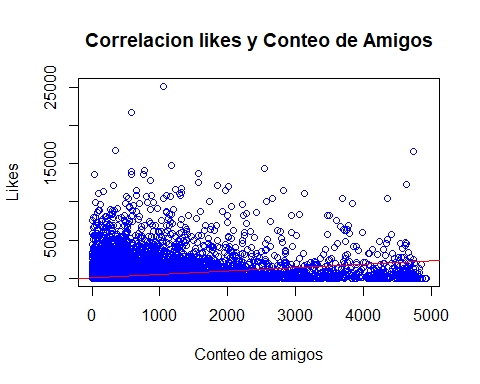
#Metodo(Correlacion)  
cor(Facebook$friend\_count,Facebook$age)

## [1] -0.02740737

#-0.02740737  
  
#en este caso el valor de correlacion es negativo esto quiere decir que se rechaza la hipotesis nula   
  
#La correacion negativa quiere decir que entre mas alto es un dato el otro es bajo como en este ejmplo se nota que   
#la correlacion es negatica en donde se ve un gran conteo de amistades en usuarios entre los 20 a?os y otro pico en   
#los que sobrepasan lo 100 a?os y la linea de correlacion siempre se mantiene bajo por no tener ni un tipo de asociacion  
  
#A a edad de 15 a 30 a?os los jovenes y adultos tienden a socializar mas y tener mas amistades que decae en los 40 por cuestiones  
#de trabajo o problemas que se nota un pico entre los 70 y otro gran pico en los mayores de 100 a?os en donde se son influidos a   
#las redes sociales y familiares queridos o amigos tiende a amistarse en dicha red social

# No.6

#el conteo de amigos total tiene alguna correlacion con los likes que da   
  
#H0 = Si tiene conrrelacion entre los datos   
#H0 = No tiene correclacion entre los datos  
  
#Grafico(Plot)  
plot(Facebook$friend\_count,Facebook$likes, col="Blue", main = "Correlacion likes y Conteo de Amigos", ylab = "Likes", xlab = "Conteo de amigos")  
modelolineal<-lm(Facebook$likes~Facebook$friend\_count)  
abline(modelolineal,col ="red")



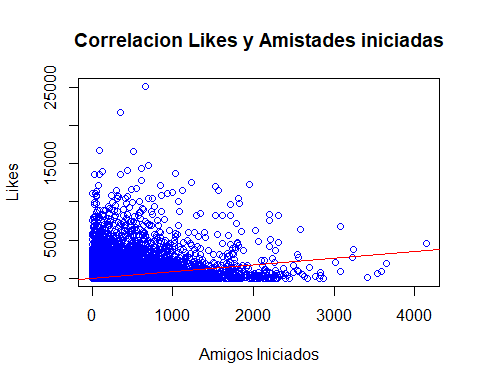
#Metodo(Correlacion)  
cor(Facebook$friend\_count,Facebook$likes)

## [1] 0.2980169

#0.2980169  
  
#La correlacion es minima asi que se puede decir que se hacepta la hipotesis nula pero de manera minima   
#ya que si existe una correlacion que no es negativa pero lo tiene de manera minima   
  
#en la grafica se explica que menos de 1000 amigos se tienden muchos usuarios a dar pocos likes pero tiene un pico elevado  
#en donde se da mucho en unas ocaiones pero a muchos amigos menos usuarios dan likes y si lo dan no sobrepasa de  
#gran manera

# No.7

#Los likes que da tiene alguna correlacion con los amigos que inicio en facebook esto quiere decir que son amigos que conocio  
#gracias a la red social y   
  
#H0 = Si tiene conrrelacion entre los datos   
#H0 = No tiene correclacion entre los datos  
  
#Grafico(Plot)  
plot(Facebook$friendships\_initiated,Facebook$likes, col="Blue", main = "Correlacion Likes y Amistades iniciadas", ylab = "Likes", xlab = "Amigos Iniciados")  
modelolineal<-lm(Facebook$likes~Facebook$friendships\_initiated)  
abline(modelolineal,col ="red")



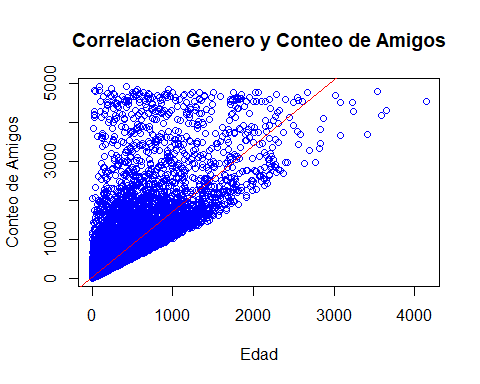
#Metodo(Correlacion)  
cor(Facebook$friendships\_initiated,Facebook$likes)

## [1] 0.2855923

#0.2855923  
  
#en ese caso el valor de correlacion es minima pero no negativa asi que se nota que los usuarios que no tienen  
#muchos amigos iniciados tienden a dar mas likes que los que tienen mas amigos iniciados en donde el promedio y la   
#constante de likes es menos esto es causa de que se inician amigos pero no se tiene mucha confianza para publicar o   
#dar opiniones de los posteos de los amigos

# No.8

#Tiene alguna correlacion el conteo de amigos total con los amigos iniciados  
  
#H0 = Si tiene conrrelacion entre los datos   
#H0 = No tiene correclacion entre los datos  
  
#Grafico(Plot)  
plot(Facebook$friendships\_initiated,Facebook$friend\_count, col="Blue", main = "Correlacion Genero y Conteo de Amigos", ylab = "Conteo de Amigos", xlab = "Edad")  
modelolineal<-lm(Facebook$friend\_count~Facebook$friendships\_initiated)  
abline(modelolineal,col ="red")



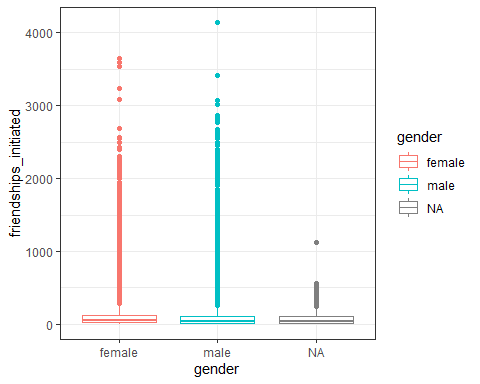
#Metodo(Correlacion)  
cor(Facebook$friend\_count,Facebook$friendships\_initiated)

## [1] 0.82585

# 0.82585  
  
#En este caso se nota que el valor de correlacion es mayor que quiere decir que si tiene correlacion en este caso  
#se nota que el total de amigos que se tiene porque concoce previamente tiene una gran correlacion con los nuevos amigos  
#inicados recomendados por la plataforma  
  
#esto da a entender que el que tiene mas amigos esta mas abierto a inicar mas amigos nuevos en redes sociales a los que son   
#cerrados que tambien tienden a ser menos abiertos a conocer nuevos amigos

# No.9

#El Genero es un factor que afecta con los amigos que inicia en la plataforma de facebook  
  
#H0: El genero no afecta (Genero = Amigos Iniciados)  
#H1: El genero si afecta (Genero =/= Amigos Iniciados)  
  
#Grafico (ggplot)  
ggplot(data=Facebook, aes(x=gender,y=friendships\_initiated, color=gender ))+ geom\_boxplot()+theme\_bw()



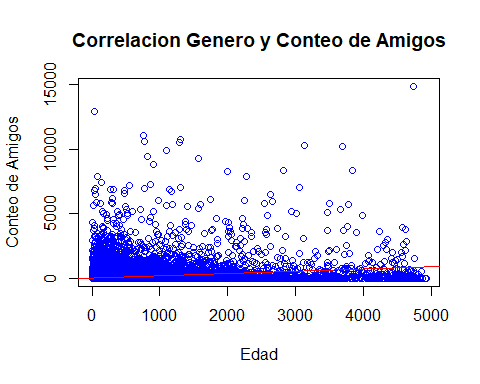
#Metodo (Chi2)  
anova<- aov(Facebook$friendships\_initiated~Facebook$gender)  
summary(anova)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Facebook$gender 1 2.800e+06 2799565 78.55 <2e-16 \*\*\*  
## Residuals 98826 3.522e+09 35640   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 175 observations deleted due to missingness

#Pvalue = 2e-16  
#En este casi se nota que el P-value es el valor minimo esto quiere decir que no se acepta la hipotesis nula  
#esto da a entender que el genero afecta con los amigos que se inician en la plataforma de fecebook en donde se   
#nota en la grafica que en el genero mastulino tiene un mayor pico de amigos iniciados pero menos constantes y las   
#mujeres tiene un menor pico de amistades iniciadas pero mas constantes con respecto a los datos atipicos   
#y se nota que en los usuarios que no definieron su genero tienen menos datos atipicos en donde no se inician amigos nuegos  
#que sigue siendo una constante en todo el dataset, en donde los que no tiene su Genero definido comparten menos informacion  
#obre ellos para mantener su privacidad apartada de las redes sociales

# No.10

#Los likes desde internet tienen una correlacion con el conteo de amigos   
  
#H0 = Si tiene conrrelacion entre los datos   
#H0 = No tiene correclacion entre los datos  
  
#Grafico(Plot)  
plot(Facebook$friend\_count,Facebook$www\_likes, col="Blue", main = "Correlacion Genero y Conteo de Amigos", ylab = "Conteo de Amigos", xlab = "Edad")  
modelolineal<-lm(Facebook$www\_likes~Facebook$friend\_count)  
abline(modelolineal,col ="red")



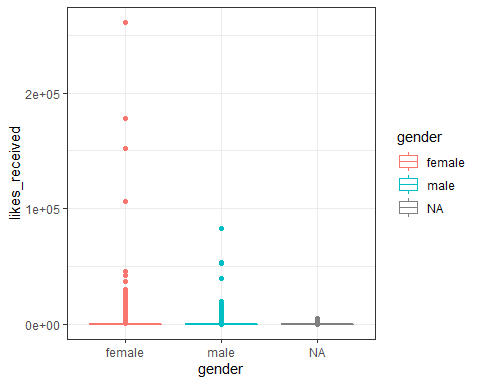
#Metodo(Correlacion)  
cor(Facebook$www\_likes,Facebook$friend\_count)

## [1] 0.2298027

# 0.2298027  
  
#En este caso se nota que no tienen mucha correlacion ya que es muy poco el valor correlativo que significa que no importa  
#de que plataforma (mobile o PC) se este siempre tendra diferenties amigos totales ya que no afectaria por cuestiones de acceisbilidad

# No.11

#El genero afecta en los likes recibidos   
  
#Grafico (ggplot)  
ggplot(data=Facebook, aes(x=gender,y=likes\_received, color=gender ))+ geom\_boxplot()+theme\_bw()



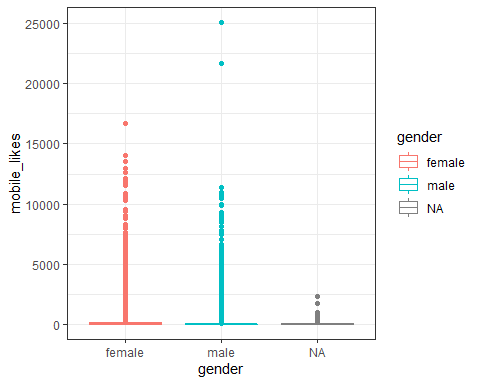
#Metodo (Chi2)  
anova<- aov(Facebook$friendships\_initiated~Facebook$gender)  
summary(anova)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Facebook$gender 1 2.800e+06 2799565 78.55 <2e-16 \*\*\*  
## Residuals 98826 3.522e+09 35640   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 175 observations deleted due to missingness

#P-Value = 2e-16  
#En este caso el P-values es menor a 0.05 entonce se rechaza la hipotesis nula que significa que el genero  
#si afecta en los likes recividos   
  
#En donde se not que las mujeres tiene un mayor pico de likes recividos e igual un poco mas constantes a diferencia  
#que los hombres en donde tiene un menor pico de likes recibidos y menos constantes segun los datos atipicos y los   
#usuairos ue no definieron su genero tienen menos likes recividos tanto en el pico maximo como en la constancia  
  
#esto se explica gracias a la constante de edades que es de 18 a 22 a?os que es en donde segun estudiso tien mas curiosisdad  
#y atraccion el sus gustos y las estadisticas muestran que los hombres tiene mas intereses en las mujeres y por eso reciven mas   
#likes por y las mujeres tieden a compartir ese tipo de informacion ya que tiene un mayor valor a la privacidad y tambien es un dato  
#importante en el sexismo o violencia en internet en donde las mujeres reciven mas atencion ya sea deceada o no

# No.12

#El genero afecta si tiene mas likes hechos desde la aplicacion mobile de facebook  
  
#H0: El genero no afecta (Genero = Likes desde Mobile)  
#H1: El genero si afecta (Genero =/= Likes desde Mobile)  
  
#Grafico (ggplot)  
ggplot(data=Facebook, aes(x=gender,y=mobile\_likes, color=gender ))+ geom\_boxplot()+theme\_bw()



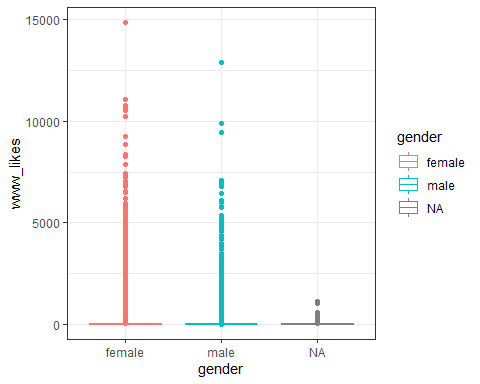
#Metodo (Anova)  
anova<- aov(Facebook$mobile\_likes~Facebook$gender)  
summary(anova)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Facebook$gender 1 3.028e+08 302767078 1549 <2e-16 \*\*\*  
## Residuals 98826 1.931e+10 195400   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 175 observations deleted due to missingness

#P-value = 2e-16  
  
#Se nota que el P-value es el minimo y por ende es menor a 0.05 que quiere decir que se rechaza la hipotesis nula  
#en dondoe da a etender que dependiedo el genero afecta si usa mas mobile o no  
  
#En la grafica se nota que los hombres tiene un mayor pico de likes desde el telefono pero tiene menos constancia  
#en las mujeres tiene menos pico de likes mobiles pero una mayor constancia

# No.13

#El genero afecta si tiene mas likes hechos desde la pagina web de facebook  
  
#H0: El genero no afecta (Genero = likes desde la pagina web)  
#H1: El genero si afecta (Genero =/= likes desde la pagina web)  
  
#Grafico (ggplot)  
ggplot(data=Facebook, aes(x=gender,y=www\_likes, color=gender ))+ geom\_boxplot()+theme\_bw()



#Metodo (Chi2)  
anova<- aov(Facebook$www\_likes~Facebook$gender)  
summary(anova)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Facebook$gender 1 9.386e+07 93857765 1163 <2e-16 \*\*\*  
## Residuals 98826 7.976e+09 80703   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 175 observations deleted due to missingness

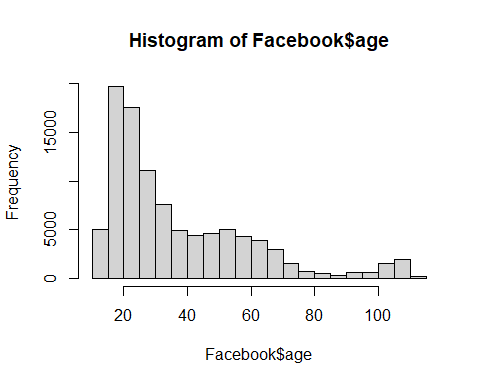
#P-alue = 2e-16   
  
#Al ser el P-value el minimo quiere decir que es menor a 0.05 y por ende no se acepta la hipotesis  
#nula lo que significa que el genero si afecta en los likes realizados desde la pagina web de facebook  
  
#En esta grafica se nota que el genero femenino tiende a dar mas likes desde la pagina web con constancia   
#y un pico maximo mayor al genero masculino, esto puede explicarse a la privasidad o seguridad que se tiene desde   
# la pagina web, los hombres tienden a querer todo con facilidad sin importar la seguiridad que se tenga ya que se a probado  
#que desde aplicaciones es mas accesible a ataques o robo de informacion a diferencia de las mujeres quienes aprecian  
#mas la privacidad por eso utilizan paginas web. Aunque en los ultimos estudios se nota que hay un crecimiento de uso  
#de telefonos en mujeres esto puede ocacionar un aumento en este tipo de dato que afectaria en un dataset mas actualizado

# No.14

#¿Cual es la edad con mas promedio que usa la red social Facebook?  
  
#Shapiro  
summary(Facebook)

## userid age dob\_day dob\_year   
## Min. :1000008 Min. : 13.00 Min. : 1.00 Min. :1900   
## 1st Qu.:1298806 1st Qu.: 20.00 1st Qu.: 7.00 1st Qu.:1963   
## Median :1596148 Median : 28.00 Median :14.00 Median :1985   
## Mean :1597045 Mean : 37.28 Mean :14.53 Mean :1976   
## 3rd Qu.:1895744 3rd Qu.: 50.00 3rd Qu.:22.00 3rd Qu.:1993   
## Max. :2193542 Max. :113.00 Max. :31.00 Max. :2000   
##   
## dob\_month gender tenure friend\_count   
## Min. : 1.000 Length:99003 Min. : 0.0 Min. : 0.0   
## 1st Qu.: 3.000 Class :character 1st Qu.: 226.0 1st Qu.: 31.0   
## Median : 6.000 Mode :character Median : 412.0 Median : 82.0   
## Mean : 6.283 Mean : 537.9 Mean : 196.4   
## 3rd Qu.: 9.000 3rd Qu.: 675.0 3rd Qu.: 206.0   
## Max. :12.000 Max. :3139.0 Max. :4923.0   
## NA's :2   
## friendships\_initiated likes likes\_received mobile\_likes   
## Min. : 0.0 Min. : 0.0 Min. : 0.0 Min. : 0.0   
## 1st Qu.: 17.0 1st Qu.: 1.0 1st Qu.: 1.0 1st Qu.: 0.0   
## Median : 46.0 Median : 11.0 Median : 8.0 Median : 4.0   
## Mean : 107.5 Mean : 156.1 Mean : 142.7 Mean : 106.1   
## 3rd Qu.: 117.0 3rd Qu.: 81.0 3rd Qu.: 59.0 3rd Qu.: 46.0   
## Max. :4144.0 Max. :25111.0 Max. :261197.0 Max. :25111.0   
##   
## mobile\_likes\_received www\_likes www\_likes\_received   
## Min. : 0.00 Min. : 0.00 Min. : 0.00   
## 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.: 0.00   
## Median : 4.00 Median : 0.00 Median : 2.00   
## Mean : 84.12 Mean : 49.96 Mean : 58.57   
## 3rd Qu.: 33.00 3rd Qu.: 7.00 3rd Qu.: 20.00   
## Max. :138561.00 Max. :14865.00 Max. :129953.00   
##

hist(Facebook$age)



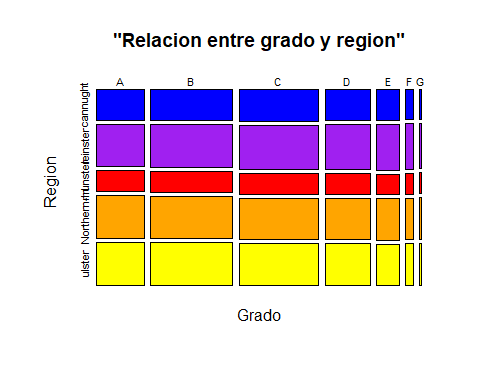
# shapiro.test(Facebook$age) (el test por medio de shapiro no es posible realizarlo por tener un limite  
#de 3 a 5000 datos en donde nosotros excedemos la cantidad maxima)  
  
#Como se nota en la grafica la edad con mayor promedio de uso de facebook es entre 18 a 22 años  
#esto se explica que las nuevas generaciones ya crecen con la tecnologia al alcanze de la mano   
#esto ocaciona que se sientan mas comodos o familiarizado de uso de redes sociales y tienen un mayor aumento  
#a este rango de edades a la accesibilidad que tiene en esta edad en donde son adolecentes y quiren conocer   
#personas que compartan sus gustos para poder compartir o sentirse adentro de una comunidad

# No.15

#¿habra dependencia entre grado segun la region?  
   
#h0 = Las variables son Independientes, El grado no depende de la region  
#h1 = Las variables no son Independientes , El grado depende de la region  
  
  
RG <- table(Banco$grade, Banco$region, dnn=c("Grado", "Region"))  
RG

## Region  
## Grado cannught leinster munster Northern-Irl ulster  
## A 25513 34485 17836 34980 35388  
## B 44553 60041 30311 58732 60898  
## C 43664 59985 28768 56037 57406  
## D 24120 34864 16339 31845 32374  
## E 12320 18015 8031 16260 16079  
## F 3941 5889 2638 5253 5325  
## G 918 1367 651 1292 1261

#Grafico (Mosaico)  
mosaicplot(RG, main=deparse("Relacion entre grado y region"), col=c("blue","purple", "red", "orange", "yellow"))



#Metodo (Chi2)  
chisq.test(RG)

##   
## Pearson's Chi-squared test  
##   
## data: RG  
## X-squared = 335.45, df = 24, p-value < 2.2e-16

#2.2e-16  
2.2e-16 < 0.05

## [1] TRUE

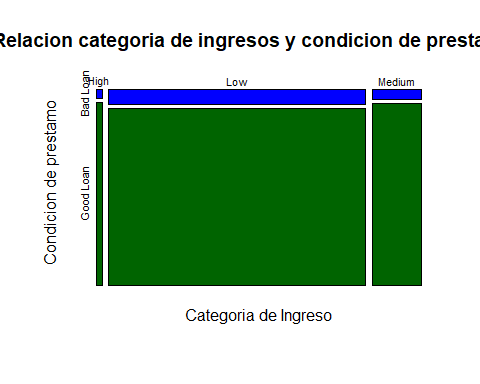
#El pvalue es menor a 0.05  
#El grado depende de la region  
#se rechaza la hipotesis nula   
  
#---analisis---#  
  
#-Considero que existen diferencias de rango, grado o categoría dependiendo el sector o región, porque considero que en algunas regiones/sectores hay mayor nivel y calidad de vida que en otro sector/región.  
#Nos podemos fijar en sector urbano y otro rural, donde Estratos sociales influyen en un territorio   
#“la región presenta características específicas que reclaman adaptar y generar una teoría sobre la estratificación desde la casa.”  
#(Maria Marinho, VeronicaQuiroz, pág. 20).  
#Estratificación social: una propuesta metodológica multidimensional para la subregión norte de América Latina y el Caribe (cepal.org)

# No.16

#¿Las variables son independientes entre condicion de pretamos y categoria de ingresos?  
  
#h0 = Las variables son independientes, la condicion de prestamo no depende de la categoria de ingreso  
#h1 = Las variables no son independientes, la condicion de prestamo depende de la categoria de ingreso  
  
Loip <- table(Banco$income\_category, Banco$loan\_condition, dnn=c("Categoria de Ingreso", "Condicion de prestamo"))  
Loip

## Condicion de prestamo  
## Categoria de Ingreso Bad Loan Good Loan  
## High 808 15978  
## Low 58865 670751  
## Medium 7756 133221

#Grafico (Mosaico)  
mosaicplot(Loip, main=deparse("Relacion categoria de ingresos y condicion de prestamos"), col=c("blue","dark green"))



#Metodo (Chi2)  
chisq.test(Loip)

##   
## Pearson's Chi-squared test  
##   
## data: Loip  
## X-squared = 1297.3, df = 2, p-value < 2.2e-16

#2.2e-16  
2.2e-16 < 0.05

## [1] TRUE

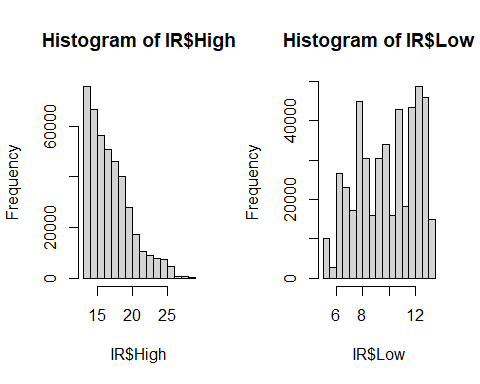
#El pvalue es menor a 0.05  
#la condicion de prestamo depende de la categoria de ingreso  
#se rechaza la hipotesis nula  
  
#---analisis---#  
  
#Hay dependencia entre el ingreso y la condición de préstamo, pongamos como ejemplo el préstamo de un local: #\*una persona con Ingresos bajos, tiene más necesidad de pedir un préstamo para poder poner un local #\*una persona con Ingresos Medios, puede tener o no tener la necesidad de pedir un préstamo #\*una persona con Ingresos Altos, tiene poca necesidad de adquirir un préstamo #Mientras tanto en la condición de préstamo indica que hay mayor aceptación a un buen préstamo que un mal préstamo #“No todas las personas que solicitan un crédito lo consiguen ni obtienen los mismos términos y condiciones: existen algunos factores que son tomados en consideración por los prestadores para determinar su solvencia, como por ejemplo, el nivel de ingresos, gastos, deudas y antecedentes de crédito.”  
  
#Fuente: https://www.consumidor.ftc.gov/articulos/s0347-sus-derechos-de-igualdad-de-oportunidad-de-credito

# No.17

#¿3. los datos de taza de interes y pagos de interés estan en normalidad?  
  
#los datos si presentan normalidad  
#los datos no presentan normalidad  
summary(Banco)

## id year issue\_d final\_d   
## Min. : 54734 Min. :2007 Length:887379 Min. :1012008   
## 1st Qu.: 9206643 1st Qu.:2013 Class :character 1st Qu.:1012016   
## Median :34433267 Median :2014 Mode :character Median :1012016   
## Mean :32465133 Mean :2014 Mean :1047089   
## 3rd Qu.:54908135 3rd Qu.:2015 3rd Qu.:1092015   
## Max. :68617057 Max. :2015 Max. :1122015   
## emp\_length\_int home\_ownership home\_ownership\_cat income\_category   
## Min. : 0.500 Length:887379 Min. :1.000 Length:887379   
## 1st Qu.: 3.000 Class :character 1st Qu.:1.000 Class :character   
## Median : 6.050 Mode :character Median :3.000 Mode :character   
## Mean : 6.051 Mean :2.099   
## 3rd Qu.:10.000 3rd Qu.:3.000   
## Max. :10.000 Max. :6.000   
## annual\_inc income\_cat loan\_amount term   
## Min. : 0 Min. :1.000 Min. : 500 Length:887379   
## 1st Qu.: 45000 1st Qu.:1.000 1st Qu.: 8000 Class :character   
## Median : 65000 Median :1.000 Median :13000 Mode :character   
## Mean : 75028 Mean :1.197 Mean :14755   
## 3rd Qu.: 90000 3rd Qu.:1.000 3rd Qu.:20000   
## Max. :9500000 Max. :3.000 Max. :35000   
## term\_cat application\_type application\_type\_cat purpose   
## Min. :1.0 Length:887379 Min. :1.000 Length:887379   
## 1st Qu.:1.0 Class :character 1st Qu.:1.000 Class :character   
## Median :1.0 Mode :character Median :1.000 Mode :character   
## Mean :1.3 Mean :1.001   
## 3rd Qu.:2.0 3rd Qu.:1.000   
## Max. :2.0 Max. :2.000   
## purpose\_cat interest\_payments interest\_payment\_cat loan\_condition   
## Min. : 1.000 Length:887379 Min. :1.000 Length:887379   
## 1st Qu.: 3.000 Class :character 1st Qu.:1.000 Class :character   
## Median : 6.000 Mode :character Median :1.000 Mode :character   
## Mean : 4.875 Mean :1.476   
## 3rd Qu.: 6.000 3rd Qu.:2.000   
## Max. :14.000 Max. :2.000   
## loan\_condition\_cat interest\_rate grade grade\_cat   
## Min. :0.00000 Min. : 5.32 Length:887379 Min. :1.000   
## 1st Qu.:0.00000 1st Qu.: 9.99 Class :character 1st Qu.:2.000   
## Median :0.00000 Median :12.99 Mode :character Median :3.000   
## Mean :0.07599 Mean :13.25 Mean :2.798   
## 3rd Qu.:0.00000 3rd Qu.:16.20 3rd Qu.:4.000   
## Max. :1.00000 Max. :28.99 Max. :7.000   
## dti total\_pymnt total\_rec\_prncp recoveries   
## Min. : 0.00 Min. : 0 Min. : 0 Min. : 0.00   
## 1st Qu.: 11.91 1st Qu.: 1915 1st Qu.: 1201 1st Qu.: 0.00   
## Median : 17.65 Median : 4895 Median : 3215 Median : 0.00   
## Mean : 18.16 Mean : 7559 Mean : 5758 Mean : 45.92   
## 3rd Qu.: 23.95 3rd Qu.:10617 3rd Qu.: 8000 3rd Qu.: 0.00   
## Max. :9999.00 Max. :57778 Max. :35000 Max. :33520.27   
## installment region   
## Min. : 15.67 Length:887379   
## 1st Qu.: 260.70 Class :character   
## Median : 382.55 Mode :character   
## Mean : 436.72   
## 3rd Qu.: 572.60   
## Max. :1445.46

IR <- split(Banco$interest\_rate, Banco$interest\_payments)  
par(mfrow=c(1,2))  
  
#histogramas  
hist(IR$High)  
  
#histogramas  
hist(IR$Low)



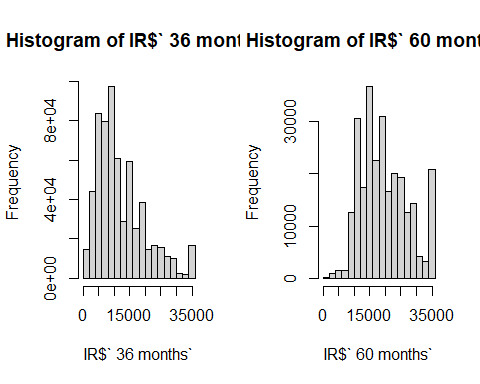
par(mfrow=c(1,1))  
  
#los datos no presentan normalidad segun la grafica  
#se rechaza la hipotesis nula

# No.18

#¿los datos de monto del préstamo por Plazo presentan normalidad?  
  
#los datos si presentan normalidad  
#los datos no presentan normalidad  
summary(Banco)

## id year issue\_d final\_d   
## Min. : 54734 Min. :2007 Length:887379 Min. :1012008   
## 1st Qu.: 9206643 1st Qu.:2013 Class :character 1st Qu.:1012016   
## Median :34433267 Median :2014 Mode :character Median :1012016   
## Mean :32465133 Mean :2014 Mean :1047089   
## 3rd Qu.:54908135 3rd Qu.:2015 3rd Qu.:1092015   
## Max. :68617057 Max. :2015 Max. :1122015   
## emp\_length\_int home\_ownership home\_ownership\_cat income\_category   
## Min. : 0.500 Length:887379 Min. :1.000 Length:887379   
## 1st Qu.: 3.000 Class :character 1st Qu.:1.000 Class :character   
## Median : 6.050 Mode :character Median :3.000 Mode :character   
## Mean : 6.051 Mean :2.099   
## 3rd Qu.:10.000 3rd Qu.:3.000   
## Max. :10.000 Max. :6.000   
## annual\_inc income\_cat loan\_amount term   
## Min. : 0 Min. :1.000 Min. : 500 Length:887379   
## 1st Qu.: 45000 1st Qu.:1.000 1st Qu.: 8000 Class :character   
## Median : 65000 Median :1.000 Median :13000 Mode :character   
## Mean : 75028 Mean :1.197 Mean :14755   
## 3rd Qu.: 90000 3rd Qu.:1.000 3rd Qu.:20000   
## Max. :9500000 Max. :3.000 Max. :35000   
## term\_cat application\_type application\_type\_cat purpose   
## Min. :1.0 Length:887379 Min. :1.000 Length:887379   
## 1st Qu.:1.0 Class :character 1st Qu.:1.000 Class :character   
## Median :1.0 Mode :character Median :1.000 Mode :character   
## Mean :1.3 Mean :1.001   
## 3rd Qu.:2.0 3rd Qu.:1.000   
## Max. :2.0 Max. :2.000   
## purpose\_cat interest\_payments interest\_payment\_cat loan\_condition   
## Min. : 1.000 Length:887379 Min. :1.000 Length:887379   
## 1st Qu.: 3.000 Class :character 1st Qu.:1.000 Class :character   
## Median : 6.000 Mode :character Median :1.000 Mode :character   
## Mean : 4.875 Mean :1.476   
## 3rd Qu.: 6.000 3rd Qu.:2.000   
## Max. :14.000 Max. :2.000   
## loan\_condition\_cat interest\_rate grade grade\_cat   
## Min. :0.00000 Min. : 5.32 Length:887379 Min. :1.000   
## 1st Qu.:0.00000 1st Qu.: 9.99 Class :character 1st Qu.:2.000   
## Median :0.00000 Median :12.99 Mode :character Median :3.000   
## Mean :0.07599 Mean :13.25 Mean :2.798   
## 3rd Qu.:0.00000 3rd Qu.:16.20 3rd Qu.:4.000   
## Max. :1.00000 Max. :28.99 Max. :7.000   
## dti total\_pymnt total\_rec\_prncp recoveries   
## Min. : 0.00 Min. : 0 Min. : 0 Min. : 0.00   
## 1st Qu.: 11.91 1st Qu.: 1915 1st Qu.: 1201 1st Qu.: 0.00   
## Median : 17.65 Median : 4895 Median : 3215 Median : 0.00   
## Mean : 18.16 Mean : 7559 Mean : 5758 Mean : 45.92   
## 3rd Qu.: 23.95 3rd Qu.:10617 3rd Qu.: 8000 3rd Qu.: 0.00   
## Max. :9999.00 Max. :57778 Max. :35000 Max. :33520.27   
## installment region   
## Min. : 15.67 Length:887379   
## 1st Qu.: 260.70 Class :character   
## Median : 382.55 Mode :character   
## Mean : 436.72   
## 3rd Qu.: 572.60   
## Max. :1445.46

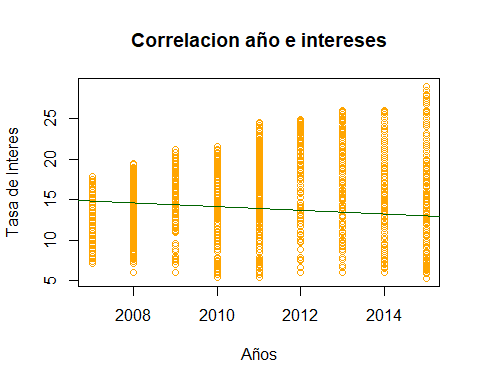
IR <- split(Banco$loan\_amount, Banco$term)  
par(mfrow=c(1,2))  
  
#histogramas  
hist(IR$` 36 months`)  
  
#histogramas  
hist(IR$` 60 months`)



par(mfrow=c(1,1))  
  
#los datos presentan normalidad segun la grafica  
#no se rechaza la hipotesis nula

# No.19

#¿tiene dependencia el interes con los años que pasen  
  
#h0=Las dos variable son independientes, el interes no tiene dependencia con los años que pasen  
#h1=las 2 variables no son independientes, el interes tiene dependencia con los años que pasen   
  
#Grafico = Plot  
plot(Banco$year, Banco$interest\_rate, col="orange", main = "Correlacion año e intereses", ylab = "Tasa de Interes", xlab = "Años")  
  
YL<-lm(Banco$interest\_rate~Banco$year)  
  
  
abline(YL,col ="dark green")



#Metodo = Correlacion  
cor(Banco$year,Banco$interest\_rate)

## [1] -0.06363243

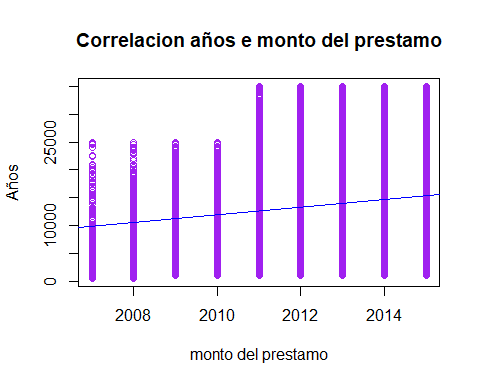
-0.06363243

## [1] -0.06363243

# la correlacion nos da un resultado de -0.06363243  
# si, si hay dependencia en los interes mientras transcuren los años  
# Se rechaza la hipotesis nula  
# ---aclaracion---#   
#En el video dije que "cor" era el pvalue  
#pero en realidad no es asi, son cosas diferentes   
#El cor, si esta mas cerca del 1 significa que hay relacion/dependencia con los elementos  
#---analisis---#  
#-Las 2 variables no son independientes, el interés tiene   
#dependencia con los años que pasen #Considero que hay   
#dependencia entre el interés y los años que pasen, considero que  
#es por la popularidad y la demanda que cada año tiene un banco #“Una de las principales razones de cambios en las tasas de  
#interés es por la demanda y oferta de crédito en el mercado.” #Fuente: https://www.afipopular.com.do/app/do/consejo\_experto.aspx?id=10545

# No.20

#¿ hay independencia en el monto del prestamo con los años que pasen  
  
#h0=las variables son independientes, el monto del prestamo no depende con los años  
#h1=las variables NO son independientes, el monto del prestamo depende con los años   
  
#Grafico = Plot  
plot(Banco$year, Banco$loan\_amount, col="purple", main = "Correlacion años e monto del prestamo", ylab = "Años", xlab = "monto del prestamo")  
LY<-lm(Banco$loan\_amount~Banco$year)  
  
  
abline(LY,col ="blue")



#Metodo = Correlacion  
cor(Banco$year,Banco$loan\_amount)

## [1] 0.1037222

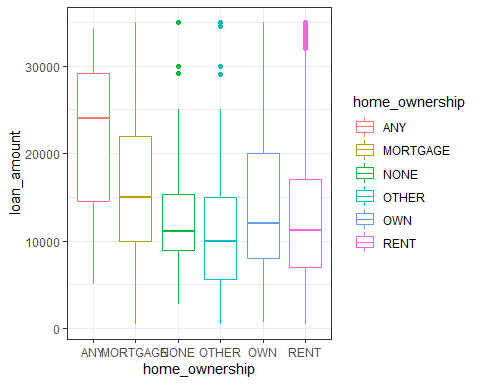
0.1037222

## [1] 0.1037222

#La correlacion nos da un resultado de 0.1037222  
#significa que si hay una gran relacion/dependencia, si, si dependen entre monto del prestamo y los años  
#Se rechaza la hipotesis nula  
  
# ---aclaracion---#   
#En el video dije que "cor" era el pvalue  
#pero en realidad no es asi, son cosas diferentes   
#El cor, si esta mas cerca del 1 significa que hay relacion/dependencia con los elementos  
  
#---analisis---#  
#Si hay dependecia en los datos, el tiempo es un factor el cual hace que el monto siga incrementando por el tiempo

# No.21

#¿hay similitud en el monto de prestamo segun la propiedad de vivienda?  
  
#Ho: no hay diferencias en el monto de prestamo por la propiedad de vivienda   
#H1: si hay diferencias en el monto de prestamo por la propiedad de vivienda  
  
#Grafico (ggplot)  
ggplot(data=Banco, aes(x=home\_ownership, y=loan\_amount, color=home\_ownership ))+ geom\_boxplot()+theme\_bw()



#Metodo (Chi2)  
anova<-aov(Banco$loan\_amount~Banco$home\_ownership)  
summary(anova)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Banco$home\_ownership 5 2.457e+12 4.913e+11 7184 <2e-16 \*\*\*  
## Residuals 887373 6.069e+13 6.839e+07   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

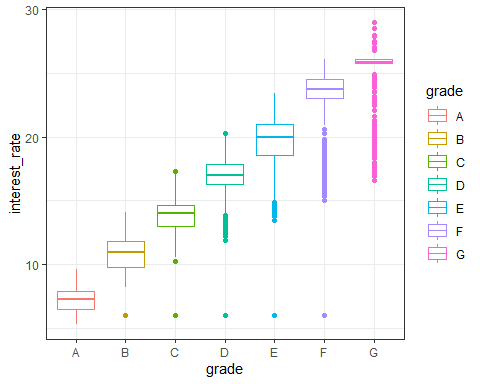
# Pr(>F)  
# 2e-16  
# (2e-16> 0.05)  
2e-16< 0.05

## [1] TRUE

#el pvalue es menor a 0.05 la grafica y el analisis indican que hay diferencias  
#Esto indica que no hay similitud de monto de prestamos por la propiedad de vivienda  
#se rechaza la hipotesis nula  
  
#---analisis---#  
  
#Pueden existir diferencias en el monto de préstamo dependiendo las necesidades #personales de la persona según la propiedad de vivienda que tenga

# No.22

#¿hay igual en la taza de intereses no dependiendo el grado?  
  
#Ho: no hay diferencias en la taza de intereses por el grado  
#H1: si hay diferencias en la taza de intereses por el grado   
  
#Grafico (ggplot)  
ggplot(data=Banco, aes(x=grade, y=interest\_rate, color=grade ))+ geom\_boxplot()+theme\_bw()



#Metodo (Chi2)  
anova<-aov(Banco$interest\_rate~Banco$grade)  
summary(anova)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Banco$grade 6 15545566 2590928 1540170 <2e-16 \*\*\*  
## Residuals 887372 1492768 2   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

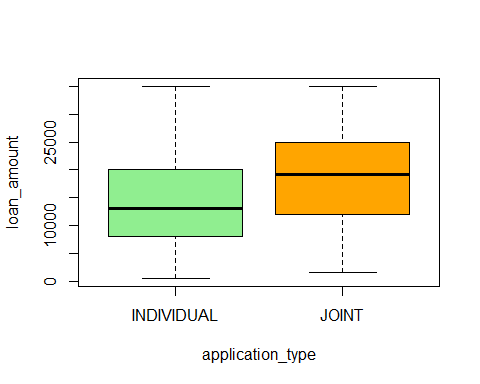
# Pr(>F)  
# 2e-16  
# (2e-16> 0.05)  
2e-16<0.05

## [1] TRUE

#el pvalue es menor a 0.05 la grafica y el analisis indican que si hay diferencias  
#Esto indica que no hay similitud en la taza de intereses por el grado  
#se rechaza la hipotesis nula  
  
#---analisis---#  
  
#Si hay diferencias, porque el nivel o jerarquía que alguien tenga, puede generarle #más beneficios, en este caso la taza de intereses para los usuarios A es menor a 10%, #a diferencia de los usuarios G que no bajan de 20%

# No.23

#¿hay diferencia en el monto del prestamo por el tipo de aplicacion  
# H0= No, no hay difenencia en el monto del prestamo por el tipo de aplicacion   
# H1= si, si hay diferencias en el monto del prestamo por el tipo de aplicacion  
  
#Grafica(Boxplot)  
boxplot(loan\_amount~application\_type, data = Banco, col = c("light green","orange"))



#Metodo(Test de Hipotesis / T.test)  
t.test(Banco$loan\_amount~Banco$application\_type, alternative = "two.sided")

##   
## Welch Two Sample t-test  
##   
## data: Banco$loan\_amount by Banco$application\_type  
## t = -11.592, df = 510.51, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -5424.75 -3852.45  
## sample estimates:  
## mean in group INDIVIDUAL mean in group JOINT   
## 14752.59 19391.19

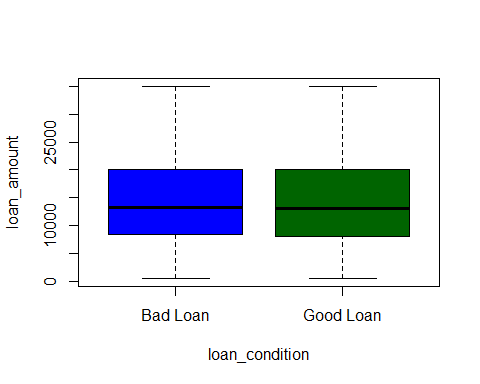
# Pr(>F)  
# 2.2e-16  
# (2.2e-16 > 0.05)  
2.2e-16<0.05

## [1] TRUE

#El P-value es menor al 0.05  
#esto quiere que si hay diferencia en el monto por el prestamo  
#se rechaza la hipotesis nula  
  
#---analisis---#  
  
#Es posible que cuando se una solicitud conjunta al momento de un #préstamo, como está asociada se debe decir mayor monto para poder repartirse en partes moderadas para cada sujeto.

# No.24

#¿hay diferencia en la condicion de prestamo con el monto de prestamo   
# H0= No, no hay difenencia en la condicion de prestamo por el monto del tiempo   
# H1= si, si hay difenencia en la condicion de prestamo por el monto del tiempo  
  
#Grafica(Boxplot)  
boxplot(loan\_amount~loan\_condition, data = Banco, col = c("blue","dark green"))



#Metodo(Test de Hipotesis / T.test)  
t.test(Banco$loan\_amount~Banco$loan\_condition, alternative = "two.sided")

##   
## Welch Two Sample t-test  
##   
## data: Banco$loan\_amount by Banco$loan\_condition  
## t = 3.2953, df = 78822, p-value = 0.0009836  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 45.30095 178.28962  
## sample estimates:  
## mean in group Bad Loan mean in group Good Loan   
## 14858.56 14746.77

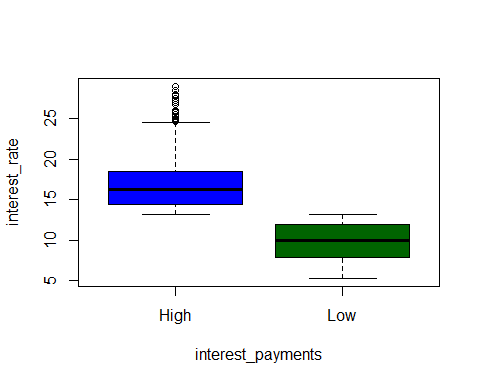
# Pr(>F)  
# 0.0009836  
# (0.0009836 > 0.05)  
0.0009836<0.05

## [1] TRUE

#El P-value es menor al 0.05  
#esto quiere que si hay diferencia en el monto de prestamo con la condicion de prestamo  
#se rechaza la hipotesis nula  
  
#---analisis---#  
#Hay diferencia ya que al momento de condicionar un préstamo se deberá de ver si a cumple,  
#en este caso, se puede ver que hay menos aceptación de préstamo

# No.25

#¿ay diferencia en pagos de interes por la tasa de interes   
# H0= No, no hay diferencia en pagos de interes por la tasa de interes   
# H1= si, si hay diferencia en pagos de interes por la tasa de interes  
  
#Grafica(Boxplot)  
boxplot(interest\_rate~interest\_payments, data = Banco, col = c("blue","dark green"))



#Metodo(Test de Hipotesis / T.test)  
t.test(Banco$interest\_rate~Banco$interest\_payments, alternative = "two.sided")

##   
## Welch Two Sample t-test  
##   
## data: Banco$interest\_rate by Banco$interest\_payments  
## t = 1248.7, df = 778764, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 7.027906 7.050003  
## sample estimates:  
## mean in group High mean in group Low   
## 16.93776 9.89881

# Pr(>F)  
# 2.2e-16  
# (2.2e-16 > 0.05)  
2.2e-16<0.05

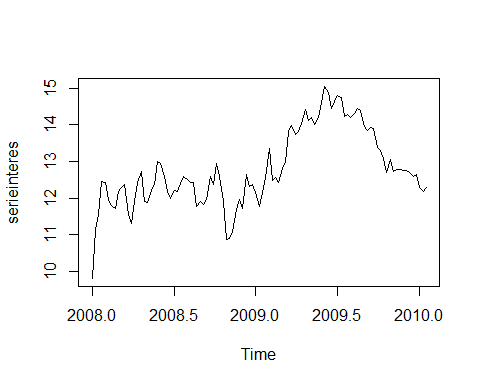
## [1] TRUE

#el pvalue ES menor a 0.05  
#esto quiere decir que hay diferencia en pagos de interes por la tasa de interes   
#se rechaza la hipotesis nula  
  
#---analisis---#  
  
#Al parecer el pago de intereses por lo regularmente es alto, puede deberse a   
#que hay mucho dinero circulando en ese banco

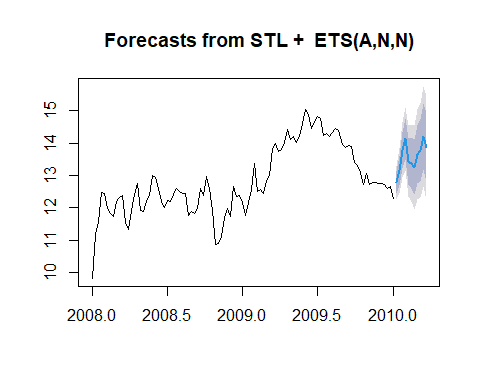
tasas <- read.csv("E:/Escritorio/Marco/Estadisitca/Proyecto/loan\_final313.csv")  
  
Banco$Fecha = as.Date(tasas$issue\_d,format="%d/%m/%Y")  
  
Banco$sem = week(Banco$Fecha)  
  
Banco$Anio = year(Banco$Fecha)  
  
interesemanal = Banco %>% group\_by(Anio,sem) %>% summarise(media=mean(interest\_rate))

## `summarise()` has grouped output by 'Anio'. You can override using the `.groups` argument.

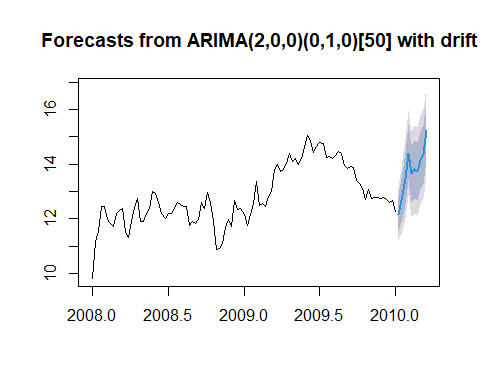
serieinteres = ts(interesemanal$media, start=c(2008,1),frequency = 50)  
  
#grafica de serie de tiempo de la tasa de interes  
plot(serieinteres)



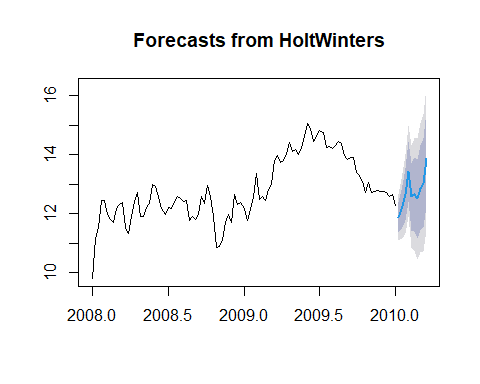
#----  
  
componente = decompose(serieinteres)  
  
tasas811 = window(serieinteres, start=2008,end=2010)  
  
primerP = forecast(tasas811,11)  
  
#Vemos la prediccion que tomara apartir de 2010  
plot(primerP)



#---  
   
modd = auto.arima(tasas811)  
  
segundoP = forecast(modd,10)  
  
plot(segundoP)



moddHT = HoltWinters(tasas811)  
  
tercerP = forecast(moddHT,10)  
  
plot(tercerP)



accuracy(primerP)

## ME RMSE MAE MPE MAPE MASE  
## Training set 0.02500998 0.2564584 0.183163 0.1851636 1.454601 0.1250011  
## ACF1  
## Training set -0.0114968

accuracy(segundoP)

## ME RMSE MAE MPE MAPE MASE  
## Training set -0.004839603 0.3438088 0.1883154 -0.08379141 1.419854 0.1285175  
## ACF1  
## Training set -0.07748538

accuracy(tercerP)

## ME RMSE MAE MPE MAPE MASE  
## Training set -0.01432816 0.3954854 0.2213174 -0.1066372 1.656105 0.1510399  
## ACF1  
## Training set -0.002238694