EXPLORACION DE DATOS

María Alejandra Osorio Alvarado 2490-19-2838 / Samuel Rolando Martinez Lopez 2490-19-17930

1/6/2021

#ANALISIS DE DATOS DE LA APPLE STORE Y DE COCHES USADOS

#CARGAR LIBRERIAS

require(stats)  
library(dplyr)

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

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)

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

library(corrplot)

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

## corrplot 0.84 loaded

library(lubridate)

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

##   
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':  
##   
## date, intersect, setdiff, union

library(forecast)

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

## Registered S3 method overwritten by 'quantmod':  
## method from  
## as.zoo.data.frame zoo

library(tidyverse)

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

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v tibble 3.1.0 v purrr 0.3.4  
## v tidyr 1.1.3 v stringr 1.4.0  
## v readr 1.4.0 v forcats 0.5.1

## Warning: package 'tibble' was built under R version 4.0.4

## Warning: package 'tidyr' was built under R version 4.0.4

## Warning: package 'readr' was built under R version 4.0.4

## Warning: package 'purrr' was built under R version 4.0.4

## Warning: package 'stringr' was built under R version 4.0.4

## Warning: package 'forcats' was built under R version 4.0.4

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x lubridate::as.difftime() masks base::as.difftime()  
## x lubridate::date() masks base::date()  
## x dplyr::filter() masks stats::filter()  
## x lubridate::intersect() masks base::intersect()  
## x dplyr::lag() masks stats::lag()  
## x lubridate::setdiff() masks base::setdiff()  
## x lubridate::union() masks base::union()

#CARGAR DATASETS

AppleStore <- read.csv("~/PROYECTO DE ESTADISTICA 2/AppleStore.csv", stringsAsFactors=TRUE)  
coches.de.segunda.mano.sample <- read.csv("~/PROYECTO DE ESTADISTICA 2/coches-de-segunda-mano-sample.csv", stringsAsFactors=TRUE)

#ANALISIS DE LA APPLE STORE

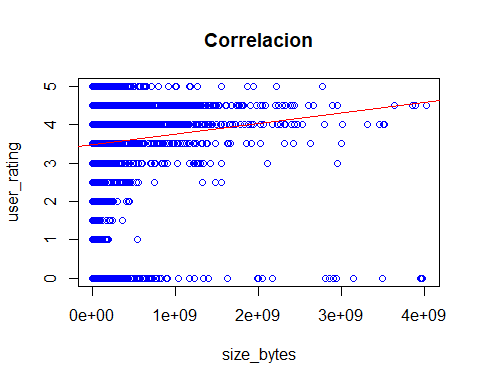
#ANALISIS 1: Existe una posible relación entre las variables peso de la aplicacion (size\_bytes) y la calificacion de los usuarios (user\_rating).

#2 VARIABLES CUANTITATIVAS #size\_bytes #user\_rating

#GRAFICO DE CORRELACION

#H0: existe una buena correlacion entre las variables #H1. no existe una buena correlacion entre las dos variables

plot(AppleStore$size\_bytes, AppleStore$user\_rating,col="blue",main="Correlacion",xlab="size\_bytes",ylab="user\_rating")  
modelo1 <- lm(AppleStore$user\_rating ~ AppleStore$size\_bytes)  
  
abline(modelo1,col="red")



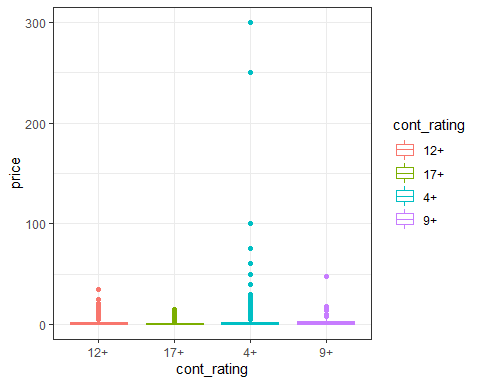
#CONCLUSION: se puede observar que tiene una mala correlacion entre el peso de la app y la calificacion del usuario esto se da a que mientras mas pesada sea la app el usuario consumira mas espacio en su dispositivo esto afecta en la calificacion que se le da a la app, asi que por ende rechazamos la hipotesis nula y afirmamos la alternativa.

#ANALISIS 2: Existe una diferencia con el precio y cont\_rating.

#“cont\_rating” #CUALITATIVA #“price” #CUANTITATIVA

#H0: si existe una diferencia en el precio y cont\_rating #H1: no existe diferencia en el precio y cont\_rating

ggplot(data= AppleStore,aes(x=cont\_rating,y=price,color=cont\_rating))+geom\_boxplot()+theme\_bw()



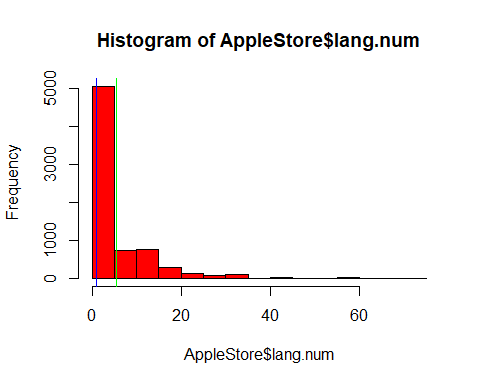
anova<-aov(AppleStore$price~ AppleStore$cont\_rating)  
summary(anova)

## Df Sum Sq Mean Sq F value Pr(>F)   
## AppleStore$cont\_rating 3 566 188.82 5.56 0.000829 \*\*\*  
## Residuals 7193 244270 33.96   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#CONCLUSION: podemos ver en el modelo que si existe una diferencia entre las variables precio y cont\_rating ya que el 4+ es mas alto que los demas por ende no se rechaza la hipotesis nula y no se afirma la alternativa.

#ANALISIS 3: Verificar la media del lang.num de las APPS, esto para verificar el promedio que las APPS tienden a tener.

hist(AppleStore$lang.num,col="red")  
mediap <- mean(AppleStore$lang.num)  
medianap <-median(AppleStore$lang.num)  
abline(v=mediap,col="green")  
abline(v=medianap,col="blue")



mediap

## [1] 5.434903

medianap

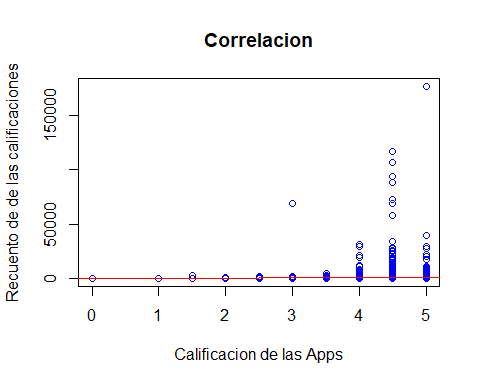
## [1] 1

#CONCLUSION: podemos ver que la media del lang.num de la AppleStore el tamaño de la media es de 5.434903 lo cual es el promedio de puntuacion que reciben las apps.

#ANALISIS 4: Existe una posible relación entre las variables entre la calificacion de la apps (user\_rating\_ver) y el recuento de las calificaciones (rating\_count\_ver).

#H0: existe una buena correlacion entre las variables #H1. no existe una buena correlacion entre las dos variables

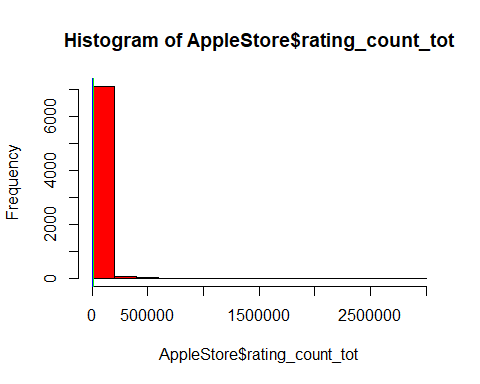
plot(AppleStore$user\_rating\_ver, AppleStore$rating\_count\_ver,col="blue",main="Correlacion",xlab="Calificacion de las Apps",ylab="Recuento de de las calificaciones")  
modelo1 <- lm(AppleStore$rating\_count\_ver ~ AppleStore$user\_rating\_ver)  
  
abline(modelo1,col="red")



#CONCLUSION: podemos ver que no existe una buena correlacion en los datos ya que teneindo todo el recuento de las apps podemos deducir que no se obtuvieron buenas calificaciones por ende se rechaza la hipotesis nula y se afirma la hipotesis alternativa.

#ANALISIS 5: Verificar la media del recuento total de las calificaciones (rating\_count\_tot) de las apps de la Applestore para verificar cual es el promedio de las calificaciones.

hist(AppleStore$rating\_count\_tot,col="red")  
mediar <- mean(AppleStore$rating\_count\_tot)  
medianar <-median(AppleStore$rating\_count\_tot)  
abline(v=mediar,col="green")  
abline(v=medianar,col="blue")



mediar

## [1] 12892.91

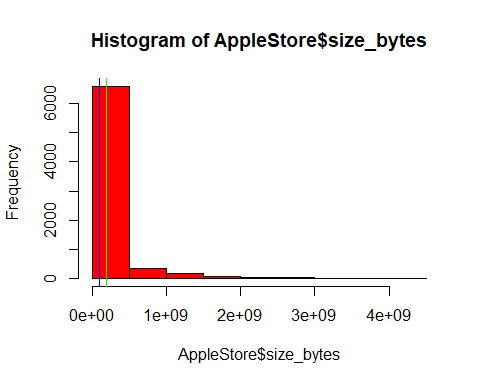
medianar

## [1] 300

#CONCLUSION: podemos ver que la media de las apps de la AppleStore el tamaño de la media es de 12892.91 podemos deducir que se tiene un buen promedio de calificaciones total de la AppleStore.

#ANALISIS 6: Verificar la media del tamaño de las APPS, esto para verificar el tamaño promedio que las APPS tienden a tener.

hist(AppleStore$size\_bytes,col="red")  
medides <- mean(AppleStore$size\_bytes)  
mediana <-median(AppleStore$size\_bytes)  
abline(v=medides,col="green")  
abline(v=mediana,col="blue")



medides

## [1] 199134454

mediana

## [1] 97153024

#CONCLUSION: Se observa que la media de las apps en relacion a tamaña es de 199134454 bytes, lo que significa que en mb es un total de 199.13 lo cual se puede decir que en la appstores encontraremos apps con un buen tamaño.

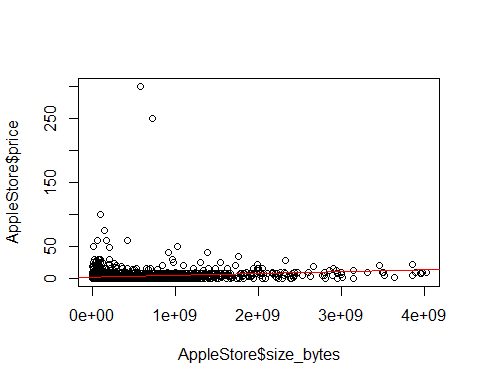
#ANALSIS 7: VERIFICAR SI EL TAMAÑO DE LAS APPS AUMENTA SI ESTAS SON DE PAGA, ESTO PARA VER SI ES PREMIUM TENDRA MEJORES GRAFICOS O MEJORES OPNIONES DE USO

#2 VARIABLES CUANTITATIVAS #size\_bytes #price

#GRAFICO DE CORRELACION #MODELO PEARSON

#H0= LAS APPS DE PAGA PESAN MAS #H1= LAS APPS DE PAGA PESAN MENOS

plot(AppleStore$size\_bytes,AppleStore$price) #variables numericas  
# crear un modelo lineal  
modelo1 <- lm(AppleStore$price ~ AppleStore$size\_bytes)  
# graficar la linea  
abline(modelo1,col="red")



#modelo para ver si existe correlacion  
cor(AppleStore$size\_bytes, AppleStore$price, method = c("pearson"))

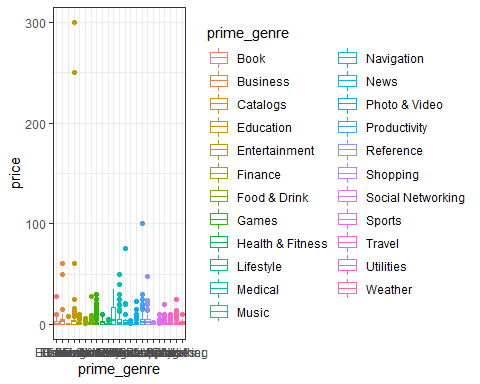
## [1] 0.1823919

#CONCLUCION: GRAFICAMENTE SE OBSERVA QUE ES POCO LA RELACION ENTRE LAS VARIAVLES, EN EL MODELO QUE UTILIZAMOS AFIRMA QUE LA CORRELACION NO ES BUENA ENTONCES Se rechaza la h0 y se afirma la hipotesis alternativa

#ANALSIS 8: VERIFICAR QUE GENERO PRINCIPAL TIENE PRECIOS MAS ALTOS DENTRO DE LA STORE

#“prime\_genre”: género principal #CUALITATIVA #“price”: importe del precio #CUANTITATIVA

ggplot(data= AppleStore,aes(x=prime\_genre,y=price,color=prime\_genre))+geom\_boxplot()+theme\_bw()



#CONCLUCION: EL PRECIO MAS ALTO EN RELACION A PRECIO ES DEL GENERO productivity ESTE GENERO SE RELACIONA A JUEGOS GAMING, ENTORNOS DE DESARROLLO, PRECIO MAXIMO ES DE 100 $ EN ESTE CASO EN ESA MONEDA

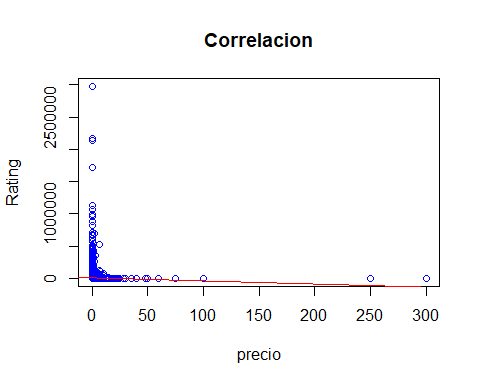
#ANALISIS 9: Verifique si a mayor precio la calificacion del usuario mejora, esto suponiendo que la aplicacion de paga es mucho mejor que las free

#rating\_count\_tot CUANTITATIVA #price CUANTITATIVA

#grafico  
#modelo

#h0= A mayor precio la calificacion de los usuarios es alta #h1= A mayor precio la calificacion de los usuarios es baja

plot(AppleStore$price,AppleStore$rating\_count\_tot,col="blue",main="Correlacion",xlab="precio",ylab="Rating")  
# crear un modelo lineal  
modelo1 <- lm(AppleStore$rating\_count\_tot ~ AppleStore$price)  
# graficar la linea  
abline(modelo1,col="red")



cor(AppleStore$rating\_count\_tot, AppleStore$price, method = c("pearson"))

## [1] -0.03904419

#CONCLUCION: Graficamente se ve que si existe correlacion pero no es nada buena y en el modelo utilizado(pearson) nos da una correlacion de 0.40 lo cual nos hace que se rechace la h0 y se afirme la h1, esto dando a entender que las apps de paga no aran que los usuarios esten satisfechos del todo

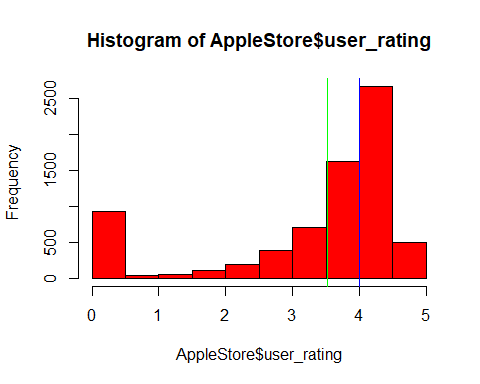
#ANALISIS 10: ANALICE GRAFICAMENTE LA VARIVABLE DE VALOR PROMEDIO DE CALIFICACIÓN DE USUARIOS, VERIFICAR SI EXISTE NORMALIDAD EN LOS DATOS DE ESTA TAMBIEN VERIFIQUE LA MEDIA Y MEDIANA DE LA MISMA.

#rating\_count cuantitativa

#grafico

#user\_rating

hist(AppleStore$user\_rating,col="red")  
medides <- mean(AppleStore$user\_rating)  
mediana <-median(AppleStore$user\_rating)  
abline(v=medides,col="green")  
abline(v=mediana,col="blue")



mediana

## [1] 4

medides

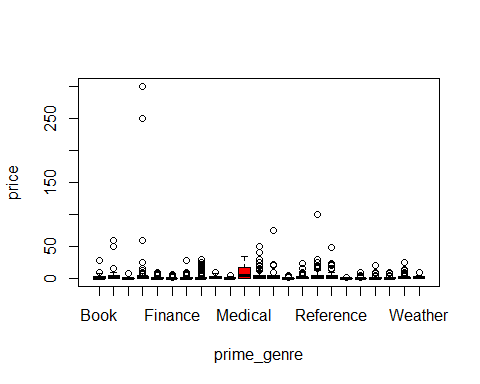
## [1] 3.526956

#CONCLUSION: en el histograma utilizado se observa que no existe normalidad en los datos, luego se observa que esta sesgado a la derecha(los datos), se ve que la valoracion de 2 a 3 no tiene normalidad por lo que la mediana se encuentra en 4 por lo que se pude afirmar que la mayoria de usuarios estan satisfechos con el contenido de la appstore

#ANALISIS 11: EXPLORE TANTO GRÁFICA COMO ANALÍTICAMENTE SI EXISTE DIFERENCIA ENTRE EL TIPO DE APPS Y EL PRECIO QUE ESTE TENGA, ESTO SUPONIENDO SI UNA APP ES DE ALGUN GENERO, SI ESTE VARIA SU PRECIO O SIMPLEMENTE TODAS VALEN POR IGUAL

#price= CUANTITATIVA #prime\_genre= CUALITATIVA #grafico= BOXPLOT #MODELO= anova #H0=EL PRECIO DEPENDE DEL GENERO #H1=EL PRECIO NO DEPENDE DEL GENERO

cajas <- boxplot(price ~ prime\_genre, data=AppleStore, col=c("green","red","blue"))



str(cajas)

## List of 6  
## $ stats: num [1:5, 1:23] 0 0 0 2.99 6.99 0 0 2.99 4.99 9.99 ...  
## $ n : num [1:23] 112 57 10 453 535 ...  
## $ conf : num [1:2, 1:23] -0.446 0.446 1.946 4.034 0 ...  
## $ out : num [1:714] 9.99 27.99 14.99 59.99 49.99 ...  
## $ group: num [1:714] 1 1 2 2 2 3 4 4 4 4 ...  
## $ names: chr [1:23] "Book" "Business" "Catalogs" "Education" ...

anova<-aov(AppleStore$price ~ AppleStore$prime\_genre)  
summary(anova)

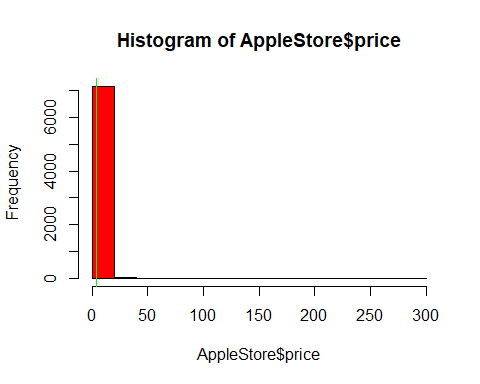
## Df Sum Sq Mean Sq F value Pr(>F)   
## AppleStore$prime\_genre 22 9537 433.5 13.22 <2e-16 \*\*\*  
## Residuals 7174 235299 32.8   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

2e-16<0.05 #CONCLUCION: Graficamente se observa que efectivamente en algunos generos de contenido, el precio depende y mucho tal es el caso entretenimiento, es algo que me asombra porque el precio de este contenido es bajo a comparacion de los demas generos, y en el modelo no se rechaza la h0 por lo tanto el precio si depende del genero #Como el pvalue es mayor a 0.5 no se rechaza la hipotesis nula.

#ANALISIS 12: GRAFICAR EL PRECIO DE LAS APPS, TRAZAR LA MEDIA DE ESTA, LUEGO VERIFICAR SI LAS APPS DE PAGA SON LO SUFICIENTEMENTE BUENAS PARA PAGAR POR ELLAS ESTO BASANDOSE POR LA VALORACION MEDIA DE LA VALORACIÓN DEL USUARIO

#price CUANTITATIVA

hist(AppleStore$price,col="red")  
media <- mean(AppleStore$price)  
abline(v=medides,col="green")



media

## [1] 1.726218

#CONCLUCION: la media del precio de las apps es de 1.726218 acordemonos que son en dolares esto dependera de la alta o baja del dolar.

#ANALISIS 13: CON RELACION AL ANALISIS ANTERIOR, VERIFICAR SI LAS APPS DE PAGAN VALEN LA PENA ESTO EN RELACION A LA VALORACION QUE LOS USUARIOS LE DAN A LAS APPS

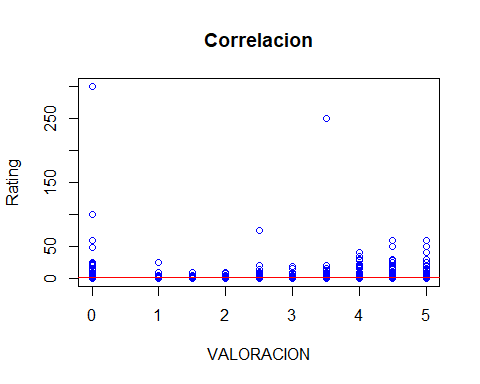
#price CUANTITATIVA #user\_rating\_ver CUANTITATIVA

#grafico PLOT #modelo PEARSON

#h0= las apps de paga valen la pnea #h1= las apps de paga no valen la pena

#Verificar si valen la pena

plot(AppleStore$user\_rating\_ver,AppleStore$price,col="blue",main="Correlacion",xlab="VALORACION",ylab="Rating")  
# crear un modelo lineal  
modelo1 <- lm(AppleStore$price ~ AppleStore$user\_rating\_ver)  
# graficar la linea  
abline(modelo1,col="red")



cor(AppleStore$user\_rating\_ver, AppleStore$price, method = c("pearson"))

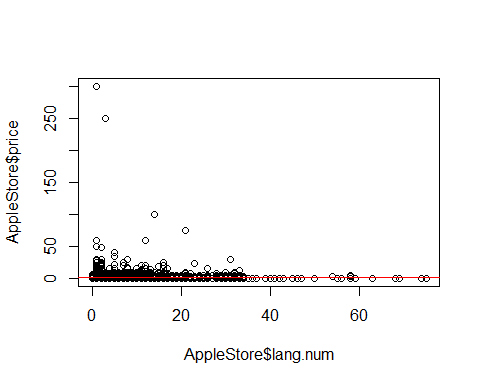
## [1] 0.02517333

#CONCLUCION: En el grafico se observa que la valoracion a las apps de paga es exactamente igual, esto se afirma con el modelo utilizado el cual nos dice que la correlacion con las variables se da muy poco el cual es de 0.02 el cual es casi inexitente, entonces se rechaza la h0 y entonces no vale la pena pagar por una app.

#ANALISIS 14: VERIFICAR SI EXISTE CORRELACION ENTRE LAS VARIABLES PRICE Y lang.num, ESTO PARA VER SI LAS APPS DE PAGA TIENEN MAS IDIOMAS ADMITIDOS QUE LAS GRATUITAS

#H0=EXISTE CORRELACION #H1=NO EXISTE CORRELACION #price #lang.num

plot(AppleStore$lang.num,AppleStore$price) #variables numericas  
# crear un modelo lineal  
modelo1 <- lm(AppleStore$price ~ AppleStore$lang.num)  
# graficar la linea  
abline(modelo1,col="red")



cor(AppleStore$lang.num, AppleStore$price, method = c("pearson"))

## [1] -0.006713362

#CONCLUCION: LAS VARIBLES NO TIENEN CORRELACION ESTO BASADO A LA GRAFICA Y AL MODELO EL CUAL NOS DA UN RESULTADO DE -0.006 DE CORRELACION LA CUAL ES CASI INEXISTENTE LA RELACION DE LAS VARIABLES, ESTO ARROJANDONOS QUE LA VARIABLE PRICE NO TIENE CORRELACION CON LA MEJORA DE LAS APPS.

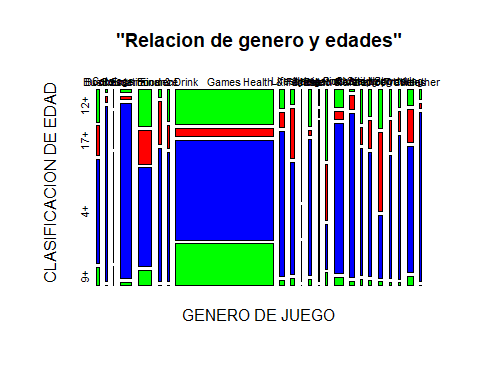
#ANALISIS 15: ESTUDIE SI EXISTE RELACION ENTRE LA VIARIABLE (prime\_genre ) Y LA VARIABLE (cont\_rating), Esto para saber si el genero tiene que ver con las edades admitidas por los terminos y condiciones

#ho= son variables independientes no hay relacion #h1= son variables independientes (si hay relacion)

tabla <- table(AppleStore$prime\_genre, AppleStore$cont\_rating, dnn=c("GENERO DE JUEGO","CLASIFICACION DE EDAD"))  
tabla

## CLASIFICACION DE EDAD  
## GENERO DE JUEGO 12+ 17+ 4+ 9+  
## Book 20 18 63 11  
## Business 1 2 54 0  
## Catalogs 1 2 6 1  
## Education 8 7 432 6  
## Entertainment 108 98 285 44  
## Finance 5 24 75 0  
## Food & Drink 11 8 44 0  
## Games 741 177 2079 865  
## Health & Fitness 19 15 141 5  
## Lifestyle 12 39 88 5  
## Medical 14 4 5 0  
## Music 28 4 102 4  
## Navigation 4 2 40 0  
## News 29 23 18 5  
## Photo & Video 35 15 285 14  
## Productivity 2 14 158 4  
## Reference 12 6 44 2  
## Shopping 18 19 82 3  
## Social Networking 36 71 57 3  
## Sports 17 19 76 2  
## Travel 8 9 63 1  
## Utilities 22 44 170 12  
## Weather 4 2 66 0

mosaicplot(tabla, main=deparse("Relacion de genero y edades"),col=c("green","red","blue"))



chisq.test(tabla)

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

##   
## Pearson's Chi-squared test  
##   
## data: tabla  
## X-squared = 1532.4, df = 66, p-value < 2.2e-16

2.2e-16>0.05 #CONCLUCION: GRAFICAMENTE Y ANALITAICAMENTE SE OBSERVA QUE LAS VARIABLES SI TIENEN RELACION POR ENDE NO SE RECHAZA LA H0, ESTO DANDO A ENTENDER QUE CADA GENERO TIENE LA EDAD DEFINIDA POR CADA GENERO Y EL GENERO QUE DEJA OBTENER LA APP CON MENOS EDAD ES DE GAMES (JUEGOS) ESTO POR EL ALTO CONTENIDO EDUCATIVO QUE OFRECEN.

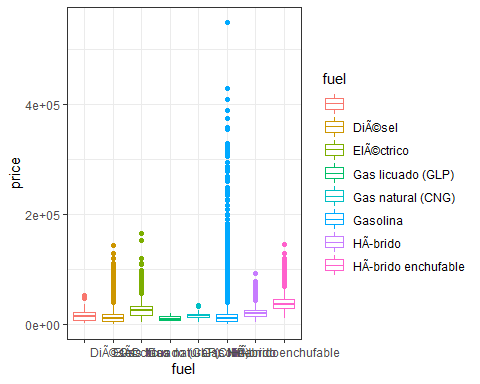
#ANALISIS DE COCHES USADOS

#ANALISIS 1: analisar graficamente y analiticamente si el precio del vehiculo depende al tipo de combustible que este utilice, formule hiptesis

#h0=El precio del vehiculo depende del tipo de combustible #h1=El precio del vehiculo no depende del tipo de combustible

#price CUANTITATIVA #fuel CUALITATIVA

ggplot(data= coches.de.segunda.mano.sample,aes(x=fuel,y=price,color=fuel))+geom\_boxplot()+theme\_bw()



anova<-aov(coches.de.segunda.mano.sample$price~ coches.de.segunda.mano.sample$fuel)  
summary(anova)

## Df Sum Sq Mean Sq F value Pr(>F)   
## coches.de.segunda.mano.sample$fuel 7 5.410e+11 7.729e+10 296.1 <2e-16 \*\*\*  
## Residuals 49992 1.305e+13 2.610e+08   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

2e-16<0.05

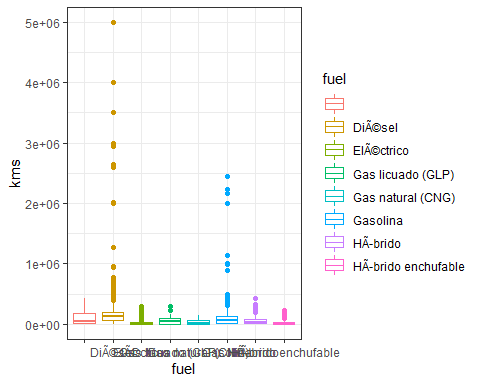
#CONCLUCION: GRAFICAMENTE SE OBSERVA QUE EL PRECIO SI DEPENDE DEL TIPO DE COMBUSTIBLE QUE ESTE UTILICE, EL PRECIO MAS ALTO LO TIENE EL DE GASOLINA ESTO SE AFIRMA CON EL MODELO UTILIZADO EL CUAL QUE NOS DICE QUE LA H0 NO SE RECHAZA, ESTO PODRIA SER POR EL FACTOR QUE LA GASOLINA SIEMPRE SE MANTIENEN POR DEBAJO DE LOS PRECIOS DE OTROS COMBUSTIBLES, Y QUE ESTE TIPO DE VEHICULOS NO GASTAN MUCHOS GALONES POR MILLAS

#ANALISIS 2: VERIFICAR SI QUE VEHICULOS TIENE MAS KILOMETRAJE, ESTO PARA VER SI LOS AUTOS DE DISEL O GASOLINA SON BASTANTE USADOS, BASANDONOS EN EL ANALISIS ANTERIOR SON CON LOS PRECIOS MAS ALTOS, PERO SERAN LOS MAS USADOS?

#H0= LOS AUTOS MAS CAROS SON LOS MAS UTILIZADOS #H1= LOS AUTOS MAS BARATOS SON MENOS UTILIZADOS

#kms CUANTITATIVA #fuel CUALITATIVA

ggplot(data= coches.de.segunda.mano.sample,aes(x=fuel,y=kms,color=fuel))+geom\_boxplot()+theme\_bw()



anova<-aov(coches.de.segunda.mano.sample$kms~ coches.de.segunda.mano.sample$fuel)  
summary(anova)

## Df Sum Sq Mean Sq F value Pr(>F)   
## coches.de.segunda.mano.sample$fuel 7 4.549e+13 6.499e+12 612.9 <2e-16 \*\*\*  
## Residuals 49992 5.301e+14 1.060e+10   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

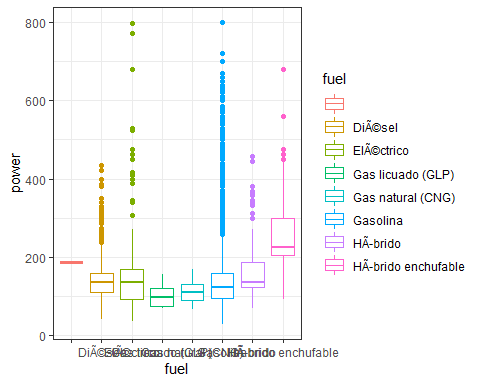
2e-16<0.05 #CONCLUCION: GRAFICAMENTE OBSERVAMOS QUE EFECTIVAMENTE LOS CARROS MAS CAROS SON MAS UTILIZADOS ESTO CON RELACION A LOS CARROS DE DISEL Y GASOLINA ESTO SE AFIRMO EN EL ANALISIS ANTERIOR, POR LO TANTO LA H0 NO SE RECHAZA

#ANALISIS 3: SEGUIMOS ANALIZANDO LA VARIABLE FUEL(Tipo de combustible del veh?culo), analizar SI EL TIPO DE COMBUSTIBLE TIENE QUE VER CON LA POTENCIA QUE ESTE TENGA, ESPERO UNA POTENCIA BAJA EN RELACION A DISEL

#power #fuel

ggplot(data= coches.de.segunda.mano.sample,aes(x=fuel,y=power,color=fuel))+geom\_boxplot()+theme\_bw()

## Warning: Removed 8528 rows containing non-finite values (stat\_boxplot).



anova<-aov(coches.de.segunda.mano.sample$power~ coches.de.segunda.mano.sample$fuel)  
summary(anova)

## Df Sum Sq Mean Sq F value Pr(>F)   
## coches.de.segunda.mano.sample$fuel 7 7090727 1012961 206.3 <2e-16 \*\*\*  
## Residuals 41464 203640158 4911   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 8528 observations deleted due to missingness

#CONCLUCION: EFECTIVAMNETE SE OBSERVA UNA BAJA DE POTENCIA EN RELACION AL VEHICULO DE DISEL E INCLUSO EL AUTO HIBRIDO TIENE UN MAXIMO DE POTENCIA MAS QUE DISEL ESTO TOMANDO EN CUENTA QUE LOS CARROS ELECTRICOS, NO TIENEN POTENCIA ALTA, PERO LOS AUTOS ELECTRICOS SI TIENE BUENA POTENCIA AL IGUAL QUE LOS VEHICULOS A GASOLINA

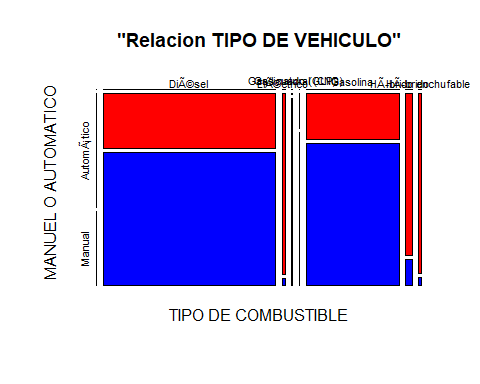
#ANALISIS 4: ANALISAR LA VARIABLE FUEL, EN RELACION A LA VARIABLE shift, ESTO PARA VER QUE VEHICULOS EN SU MAYORIA SON MANUALES O AUTOMATICOS.

#fuel CUALITATIVA #shift CUANTITATIVA

tabla <- table(coches.de.segunda.mano.sample$fuel, coches.de.segunda.mano.sample$shift, dnn=c("TIPO DE COMBUSTIBLE","MANUEL O AUTOMATICO"))  
tabla

## MANUEL O AUTOMATICO  
## TIPO DE COMBUSTIBLE AutomÃ¡tico Manual  
## 0 28 18  
## DiÃ©sel 77 8999 21790  
## ElÃ©ctrico 2 535 21  
## Gas licuado (GLP) 0 1 113  
## Gas natural (CNG) 0 17 74  
## Gasolina 28 4122 12489  
## HÃ­brido 2 1012 160  
## HÃ­brido enchufable 2 488 22

mosaicplot(tabla, main=deparse("Relacion TIPO DE VEHICULO"),col=c("green","red","blue"))



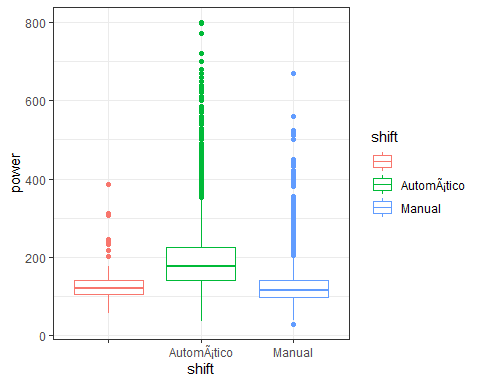
#CONCLUCION: SE OBSERVA EN LA GRAFICA CLARAMENTE LOS TIPOS DE VEHICULOS QUE SON EN SU MAYORIA MANUALES LOS CUALES SON DISEL Y GASOLINA ESTO ES UN CLARO EJEMPLO DE GUATEMALA EN SU MAYORIA LOS AUTOS SON MECANICOS POR EL HECHO QUE LOS DISEL Y GASOLINA VIENEN PARA TRABAJO DURO Y LAS CAJAS AUTOMATICAS IGUALMENTE SON MAS POTENTES , ESTO LO VERIFICAREMOS EN EL SIGUIENTE ANALISIS, LOS DEMAS LA MAYORIA SON AUTOMATICOS POR EL HECHO QUE SON PARA LA CIUDAD, PARA TRABAJO COTIDIANO

#ANALISIS 5: VERIFICAR QUE TIPO DE TRANSMICION TIENE MAS POTENCIA (MANUAL O AUTOMATICO)

#power CUANTITATIVA #shift CUALITATIVA

ggplot(data= coches.de.segunda.mano.sample,aes(x=shift,y=power,color=shift))+geom\_boxplot()+theme\_bw()

## Warning: Removed 8528 rows containing non-finite values (stat\_boxplot).



anova<-aov(coches.de.segunda.mano.sample$power~ coches.de.segunda.mano.sample$shift)  
summary(anova)

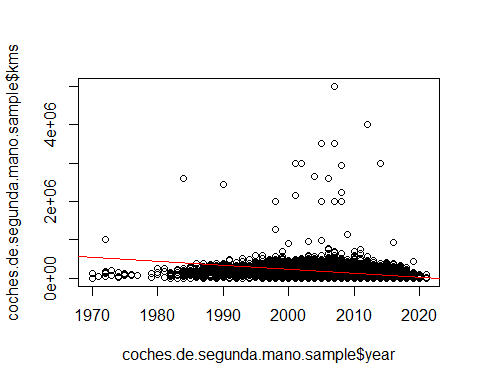
## Df Sum Sq Mean Sq F value Pr(>F)   
## coches.de.segunda.mano.sample$shift 2 55645534 27822767 7440 <2e-16 \*\*\*  
## Residuals 41469 155085350 3740   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 8528 observations deleted due to missingness

#CONCLUCION: SE VE QUE SI, LAS CAJAS AUTOMATICAS TIENEN MAS POTENCIAS, UN FACTOR ES QUE LAS REVOLUCIONES DE LA CAJA SUBEN MAS QUE LA MANUEL, LA CUAL DEPENDE DE NOSOTROS EN CAMBIO UNA CAJA AUTOMATICA VA REVOLUCIONANDO SEGUN LA PRECION QUE LE TENGAMOS, AUNQUE LAS CAJAS MANUALES NO SE QUEDAN ATRAS

#ANALISIS 6: ANALICE si existe correlacion entre la variable year(a?o del vehiculo) y la variable kms.

#H0= a?o mas reciente menos kms #H1= a?o mas reciente mas kms

plot(coches.de.segunda.mano.sample$year,coches.de.segunda.mano.sample$kms) #variables numericas  
# crear un modelo lineal  
modelo1 <- lm(coches.de.segunda.mano.sample$kms ~ coches.de.segunda.mano.sample$year)  
# graficar la linea  
abline(modelo1,col="red")



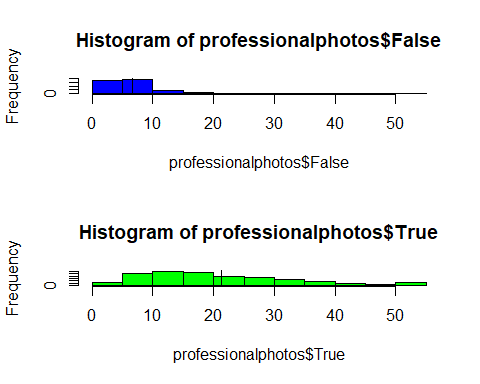
cor(coches.de.segunda.mano.sample$kms, coches.de.segunda.mano.sample$year, method = c("pearson"))

## [1] NA

#CONCLUCION: Graficamente se observa que los autos recien salidos tienen poco kilometraje, esto es por el uso cada uno de ellos lo utiliza menos.

#ANALISIS 7: construya un histograma de las fotos del coche por si es profesional o no el coche. Podría afirmar que los datos son normales?

professionalphotos <- split(coches.de.segunda.mano.sample$photos,coches.de.segunda.mano.sample$is\_professional)  
  
par(mfrow=c(2,1))  
  
hist(professionalphotos$False,col="blue")  
mediaf <- mean(professionalphotos$False)  
abline(v=mediaf,col="black")  
  
hist(professionalphotos$True,col="green")  
mediat <- mean(professionalphotos$True)  
abline(v=mediat,col="black")



#CONCLUSION: como se puede ver en los graficos se podria decir que los datos no son normales ya que existe una diferencia entre las fotos y si el auto es profesional podemos deducir de que mientras el auto tenga mas fotos se puede confirmar que es un coche profesional.

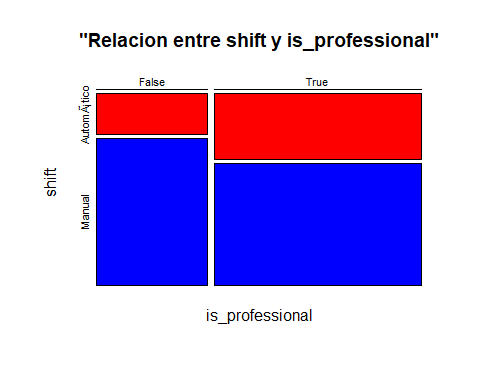
#ANALISIS 8: analizar si existe una diferencia entre la variable is\_professional y la shift.

#H0: Existe diferencia entre el is\_professional y el shift #H1: No existe diferencia entre el is\_professional y el shift

tabla1 <- table(coches.de.segunda.mano.sample$is\_professional, coches.de.segunda.mano.sample$shift, dnn=c("is\_professional","shift"))  
tabla1

## shift  
## is\_professional AutomÃ¡tico Manual  
## False 82 3730 13639  
## True 29 11472 21048

mosaicplot(tabla1, main=deparse("Relacion entre shift y is\_professional"),col=c("green","red","blue"))



chisq.test(tabla1)

##   
## Pearson's Chi-squared test  
##   
## data: tabla1  
## X-squared = 1091.1, df = 2, p-value < 2.2e-16

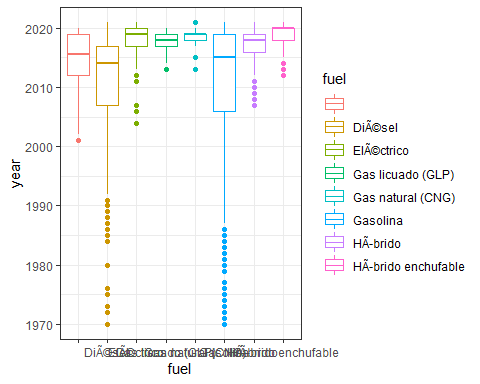
#CONCLUSION: concluimos que segun el p-value no se rechaza la hipotesis nula y no se afirma la alternativa pero como se puede ver en la grafica si existe una diferencia

#ANALISIS 9: Existe una diferencia con el año y fuel.

#H0: si existe una diferencia en el precio y cont\_rating #H1: no existe diferencia en el precio y cont\_rating

ggplot(data= coches.de.segunda.mano.sample,aes(x=fuel,y=year,color=fuel))+geom\_boxplot()+theme\_bw()

## Warning: Removed 2 rows containing non-finite values (stat\_boxplot).



anova<-aov(coches.de.segunda.mano.sample$year ~ coches.de.segunda.mano.sample$fuel)  
summary(anova)

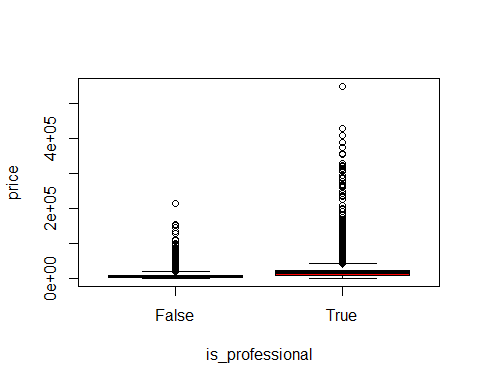
## Df Sum Sq Mean Sq F value Pr(>F)   
## coches.de.segunda.mano.sample$fuel 7 83889 11984 279.2 <2e-16 \*\*\*  
## Residuals 49990 2145601 43   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 2 observations deleted due to missingness

#CONCLUSION: podemos ver en el modelo que si existe una diferencia entre las variables año y fuel ya que la gasolina es mas alto que los demas por ende no se rechaza la hipotesis nula y no se afirma la alternativa.

#ANALISIS 10: Existe una diferencia con el precio y is professional.

#H0: si existe una diferencia en el precio y cont\_rating #H1: no existe diferencia en el precio y cont\_rating

cajas <- boxplot(price ~ is\_professional, data=coches.de.segunda.mano.sample, col=c("green","red","blue"))



str(cajas)

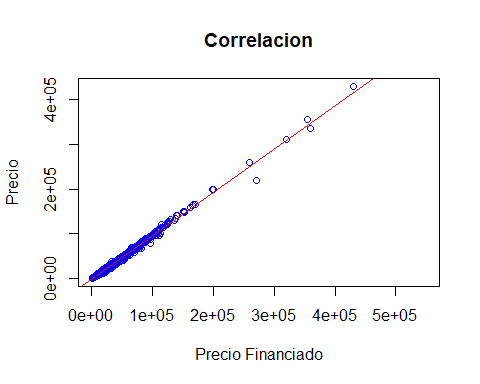
## List of 6  
## $ stats: 'integer' num [1:5, 1:2] 300 2250 4600 9300 19850 ...  
## $ n : num [1:2] 17451 32549  
## $ conf : num [1:2, 1:2] 4516 4684 15684 15916  
## $ out : num [1:3073] 48200 30000 25900 19950 20900 ...  
## $ group: num [1:3073] 1 1 1 1 1 1 1 1 1 1 ...  
## $ names: chr [1:2] "False" "True"

#CONCLUSION: Podemos ver mediante el grafico que existe una diferencia notoria entre las variables por ende concluimos que la hipotesis nula no se rechaza y la alternativa no se afirma.

#ANALISIS 11: Existe una posible relación entre las variables precio y precio finaciado de los coches.

#H0: existe una buena correlacion entre las variables #H1. no existe una buena correlacion entre las dos variables

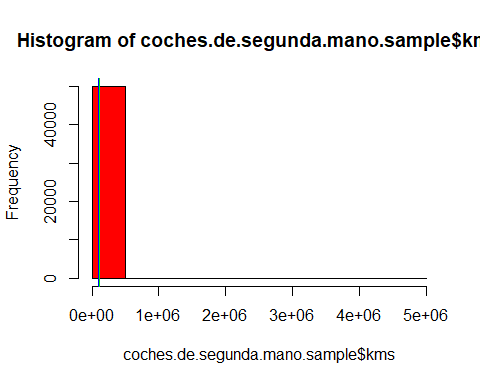
plot(coches.de.segunda.mano.sample$price, coches.de.segunda.mano.sample$price\_financed,col="blue",main="Correlacion",xlab="Precio Financiado",ylab="Precio")  
modelo1 <- lm(coches.de.segunda.mano.sample$price\_financed ~ coches.de.segunda.mano.sample$price)  
  
abline(modelo1,col="red")



#CONCLUSION: por lo que podemos apreciar en el modelo es que existe una buena correlacion en los datos del precio y precio financiado por ende concluimos que la hipotesis nula no se rechaza y la alternativa no se afirma.

#ANALISIS 12: Verificar la media de los kilometros por hora de los coches.

hist(coches.de.segunda.mano.sample$kms,col="red")  
mediak <- mean(coches.de.segunda.mano.sample$kms)  
medianak <-median(coches.de.segunda.mano.sample$kms)  
abline(v=mediak,col="green")  
abline(v=medianak,col="blue")



mediak

## [1] 116359.1

medianak

## [1] 99999

#CONCLUSION: podemos ver que la media de los kilometros por hora de los coches el tamaño de la media es de 116359.1 lo cual es buena.

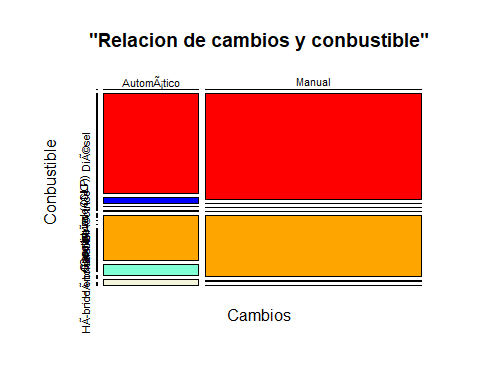
#ANALISIS 13: Estudie si existe relacion entre la variable (shift) y la variable (fuel).

#H0: Existe diferencia entre el los cambios y el combustible #H1: No existe diferencia entre el is\_professional y el shift

tabla <- table(coches.de.segunda.mano.sample$shift, coches.de.segunda.mano.sample$fuel, dnn=c("Cambios","Conbustible"))  
tabla

## Conbustible  
## Cambios DiÃ©sel ElÃ©ctrico Gas licuado (GLP) Gas natural (CNG)  
## 0 77 2 0 0  
## AutomÃ¡tico 28 8999 535 1 17  
## Manual 18 21790 21 113 74  
## Conbustible  
## Cambios Gasolina HÃ­brido HÃ­brido enchufable  
## 28 2 2  
## AutomÃ¡tico 4122 1012 488  
## Manual 12489 160 22

mosaicplot(tabla, main=deparse("Relacion de cambios y conbustible"),col=c("green","red","blue", "pink", "yellow",'orange' ,'aquamarine1', 'beige'))



chisq.test(tabla)

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

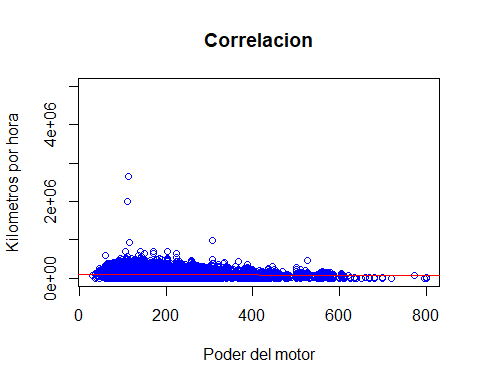
##   
## Pearson's Chi-squared test  
##   
## data: tabla  
## X-squared = 4234.9, df = 14, p-value < 2.2e-16

#CONCLUSION: como se puede apreciar en el modelo existe una diferencia notoria entre los diferentes tipos de combustible para los coches por ende no se rechaza la hipotesis nula y no se afirma la alternativa.

#ANALISIS 14: Existe una posible relación entre las variables kilometros por hora y el poder del motor de los coches.

#H0: existe una buena correlacion entre las variables #H1. no existe una buena correlacion entre las dos variables

plot(coches.de.segunda.mano.sample$power, coches.de.segunda.mano.sample$kms,col="blue",main="Correlacion",xlab="Poder del motor",ylab="Kilometros por hora")  
modelo1 <- lm(coches.de.segunda.mano.sample$kms ~ coches.de.segunda.mano.sample$power)  
  
abline(modelo1,col="red")



#CONCLUSION: por lo que se puede apreciar en el modelo existe una buena correlacion entre las variables poder del motor y kilometros por hora por ende cocnluimos que la hipotesis nula no se rechaza y la alterantiva no se afirma.