ProyectoFinal Kevin García 2490-19-20597, Cristopher Luna 2490-17-

Kevin García

31/5/2021

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

require(stats)  
library(tidyverse)

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

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

## v ggplot2 3.3.3 v purrr 0.3.4  
## v tibble 3.1.1 v dplyr 1.0.5  
## v tidyr 1.1.3 v stringr 1.4.0  
## v readr 1.4.0 v forcats 0.5.1

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

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

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

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

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

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

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

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

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(dplyr)  
library(ggplot2)  
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(magrittr)

##   
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':  
##   
## set\_names

## The following object is masked from 'package:tidyr':  
##   
## extract

#DATASETS UTILIZADOS  
Covidworld <- read.csv('D:/Estadistica2/covid19Complete.csv')  
  
covidata <- read.csv('D:/Estadistica2/Covid-data.csv')

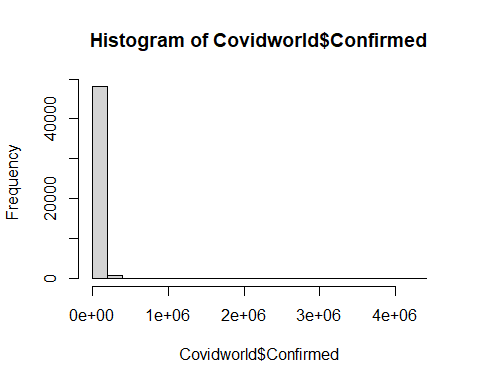
summary(Covidworld)

## Province.State Country.Region Lat Long   
## Length:49068 Length:49068 Min. :-51.796 Min. :-135.00   
## Class :character Class :character 1st Qu.: 7.873 1st Qu.: -15.31   
## Mode :character Mode :character Median : 23.634 Median : 21.75   
## Mean : 21.434 Mean : 23.53   
## 3rd Qu.: 41.204 3rd Qu.: 80.77   
## Max. : 71.707 Max. : 178.06   
## Date Confirmed Deaths Recovered   
## Length:49068 Min. : 0 Min. : 0.0 Min. : 0   
## Class :character 1st Qu.: 4 1st Qu.: 0.0 1st Qu.: 0   
## Mode :character Median : 168 Median : 2.0 Median : 29   
## Mean : 16885 Mean : 884.2 Mean : 7916   
## 3rd Qu.: 1518 3rd Qu.: 30.0 3rd Qu.: 666   
## Max. :4290259 Max. :148011.0 Max. :1846641   
## Active WHO.Region   
## Min. : -14 Length:49068   
## 1st Qu.: 0 Class :character   
## Median : 26 Mode :character   
## Mean : 8085   
## 3rd Qu.: 606   
## Max. :2816444

summary(covidata)

## iso\_code continent location date   
## Length:91026 Length:91026 Length:91026 Length:91026   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## total\_cases new\_cases new\_cases\_smoothed total\_deaths   
## Min. : 1 Min. :-74347 Min. : -6223.0 Min. : 1   
## 1st Qu.: 1055 1st Qu.: 2 1st Qu.: 7.1 1st Qu.: 48   
## Median : 11611 Median : 70 Median : 87.3 Median : 333   
## Mean : 915156 Mean : 6034 Mean : 6046.6 Mean : 24794   
## 3rd Qu.: 123217 3rd Qu.: 783 3rd Qu.: 819.7 3rd Qu.: 3229   
## Max. :167316360 Max. :905992 Max. :826374.3 Max. :3473036   
## NA's :2690 NA's :2691 NA's :3698 NA's :12542   
## new\_deaths new\_deaths\_smoothed total\_cases\_per\_million  
## Min. :-1918.0 Min. : -232.143 Min. : 0.0   
## 1st Qu.: 0.0 1st Qu.: 0.000 1st Qu.: 220.9   
## Median : 2.0 Median : 1.286 Median : 1514.1   
## Mean : 141.6 Mean : 126.434 Mean : 11220.1   
## 3rd Qu.: 18.0 3rd Qu.: 14.000 3rd Qu.: 11074.4   
## Max. :17906.0 Max. :14436.286 Max. :175616.4   
## NA's :12384 NA's :3698 NA's :3163   
## new\_cases\_per\_million new\_cases\_smoothed\_per\_million total\_deaths\_per\_million  
## Min. :-2153.437 Min. :-276.825 Min. : 0.001   
## 1st Qu.: 0.202 1st Qu.: 1.213 1st Qu.: 7.297   
## Median : 7.846 Median : 10.428 Median : 43.718   
## Mean : 75.756 Mean : 75.893 Mean : 245.708   
## 3rd Qu.: 70.164 3rd Qu.: 78.714 3rd Qu.: 267.625   
## Max. :18293.675 Max. :4083.500 Max. :3059.931   
## NA's :3164 NA's :4166 NA's :13002   
## new\_deaths\_per\_million new\_deaths\_smoothed\_per\_million reproduction\_rate  
## Min. :-76.445 Min. :-10.921 Min. :-0.01   
## 1st Qu.: 0.000 1st Qu.: 0.000 1st Qu.: 0.85   
## Median : 0.124 Median : 0.148 Median : 1.01   
## Mean : 1.516 Mean : 1.354 Mean : 1.01   
## 3rd Qu.: 1.329 3rd Qu.: 1.271 3rd Qu.: 1.18   
## Max. :218.329 Max. : 63.140 Max. : 5.81   
## NA's :12844 NA's :4166 NA's :17659   
## icu\_patients icu\_patients\_per\_million hosp\_patients   
## Min. : 0.0 Min. : 0.00 Min. : 0   
## 1st Qu.: 32.0 1st Qu.: 4.71 1st Qu.: 118   
## Median : 186.0 Median : 17.87 Median : 703   
## Mean : 1099.5 Mean : 27.08 Mean : 4842   
## 3rd Qu.: 730.2 3rd Qu.: 42.22 3rd Qu.: 2862   
## Max. :28889.0 Max. :192.55 Max. :133214   
## NA's :81850 NA's :81850 NA's :79611   
## hosp\_patients\_per\_million weekly\_icu\_admissions  
## Min. : 0.00 Min. : 0.00   
## 1st Qu.: 23.48 1st Qu.: 8.99   
## Median : 85.89 Median : 50.82   
## Mean : 176.02 Mean : 281.46   
## 3rd Qu.: 261.70 3rd Qu.: 226.75   
## Max. :1532.57 Max. :4002.46   
## NA's :79611 NA's :90196   
## weekly\_icu\_admissions\_per\_million weekly\_hosp\_admissions  
## Min. : 0.00 Min. : 0.00   
## 1st Qu.: 1.95 1st Qu.: 51.62   
## Median : 9.70 Median : 340.83   
## Mean : 21.40 Mean : 3753.21   
## 3rd Qu.: 23.93 3rd Qu.: 1873.08   
## Max. :278.76 Max. :116938.00   
## NA's :90196 NA's :89580   
## weekly\_hosp\_admissions\_per\_million new\_tests total\_tests   
## Min. : 0.00 Min. :-239172 Min. : 0   
## 1st Qu.: 9.42 1st Qu.: 1509 1st Qu.: 137476   
## Median : 45.18 Median : 5726 Median : 690273   
## Mean : 115.23 Mean : 45208 Mean : 6661465   
## 3rd Qu.: 134.02 3rd Qu.: 22561 3rd Qu.: 2898434   
## Max. :2656.91 Max. :2945871 Max. :441017808   
## NA's :89580 NA's :49840 NA's :50153   
## total\_tests\_per\_thousand new\_tests\_per\_thousand new\_tests\_smoothed  
## Min. : 0.00 Min. :-23.01 Min. : 0   
## 1st Qu.: 12.33 1st Qu.: 0.13 1st Qu.: 1567   
## Median : 61.55 Median : 0.58 Median : 5934   
## Mean : 259.96 Mean : 1.99 Mean : 43073   
## 3rd Qu.: 251.57 3rd Qu.: 1.86 3rd Qu.: 24942   
## Max. :7617.47 Max. :327.09 Max. :2016058   
## NA's :50153 NA's :49840 NA's :43344   
## new\_tests\_smoothed\_per\_thousand positive\_rate tests\_per\_case   
## Min. : 0.00 Min. :0.00 Min. : 1.3   
## 1st Qu.: 0.13 1st Qu.:0.02 1st Qu.: 7.6   
## Median : 0.59 Median :0.06 Median : 17.5   
## Mean : 1.90 Mean :0.09 Mean : 159.3   
## 3rd Qu.: 1.92 3rd Qu.:0.13 3rd Qu.: 54.0   
## Max. :92.10 Max. :0.74 Max. :44258.7   
## NA's :43344 NA's :46582 NA's :47176   
## tests\_units total\_vaccinations people\_vaccinated   
## Length:91026 Min. :0.000e+00 Min. : 0   
## Class :character 1st Qu.:7.945e+04 1st Qu.: 63893   
## Mode :character Median :6.258e+05 Median : 461405   
## Mean :1.977e+07 Mean : 11406894   
## 3rd Qu.:3.789e+06 3rd Qu.: 2600204   
## Max. :1.702e+09 Max. :775676935   
## NA's :78920 NA's :79673   
## people\_fully\_vaccinated new\_vaccinations new\_vaccinations\_smoothed  
## Min. : 1 Min. : 0 Min. : 0   
## 1st Qu.: 32302 1st Qu.: 3677 1st Qu.: 852   
## Median : 266555 Median : 21852 Median : 6493   
## Mean : 6227444 Mean : 499830 Mean : 247146   
## 3rd Qu.: 1515542 3rd Qu.: 114007 3rd Qu.: 38916   
## Max. :395315319 Max. :33958872 Max. :28391104   
## NA's :82259 NA's :80842 NA's :70489   
## total\_vaccinations\_per\_hundred people\_vaccinated\_per\_hundred  
## Min. : 0.00 Min. : 0.00   
## 1st Qu.: 1.43 1st Qu.: 1.24   
## Median : 7.02 Median : 5.11   
## Mean : 17.03 Mean : 11.70   
## 3rd Qu.: 22.17 3rd Qu.: 15.92   
## Max. :226.14 Max. :115.73   
## NA's :78920 NA's :79673   
## people\_fully\_vaccinated\_per\_hundred new\_vaccinations\_smoothed\_per\_million  
## Min. : 0.00 Min. : 0   
## 1st Qu.: 0.58 1st Qu.: 326   
## Median : 2.70 Median : 1461   
## Mean : 6.59 Mean : 2957   
## 3rd Qu.: 7.78 3rd Qu.: 3944   
## Max. :110.41 Max. :118759   
## NA's :82259 NA's :70489   
## stringency\_index population population\_density median\_age   
## Min. : 0.00 Min. :8.090e+02 Min. : 0.137 Min. :15.10   
## 1st Qu.: 44.44 1st Qu.:2.352e+06 1st Qu.: 36.253 1st Qu.:22.20   
## Median : 60.19 Median :9.905e+06 Median : 83.479 Median :29.90   
## Mean : 58.68 Mean :1.265e+08 Mean : 380.203 Mean :30.58   
## 3rd Qu.: 75.00 3rd Qu.:3.481e+07 3rd Qu.: 209.588 3rd Qu.:39.10   
## Max. :100.00 Max. :7.795e+09 Max. :20546.766 Max. :48.20   
## NA's :13646 NA's :604 NA's :6364 NA's :9273   
## aged\_65\_older aged\_70\_older gdp\_per\_capita extreme\_poverty  
## Min. : 1.144 Min. : 0.526 Min. : 661.2 Min. : 0.10   
## 1st Qu.: 3.466 1st Qu.: 2.043 1st Qu.: 4466.5 1st Qu.: 0.60   
## Median : 6.614 Median : 3.871 Median : 12951.8 Median : 2.20   
## Mean : 8.801 Mean : 5.573 Mean : 19296.9 Mean :13.37   
## 3rd Qu.:14.312 3rd Qu.: 9.167 3rd Qu.: 27216.4 3rd Qu.:21.20   
## Max. :27.049 Max. :18.493 Max. :116935.6 Max. :77.60   
## NA's :10197 NA's :9727 NA's :9054 NA's :35547   
## cardiovasc\_death\_rate diabetes\_prevalence female\_smokers male\_smokers   
## Min. : 79.37 Min. : 0.990 Min. : 0.10 Min. : 7.70   
## 1st Qu.:167.29 1st Qu.: 5.310 1st Qu.: 1.90 1st Qu.:21.60   
## Median :242.65 Median : 7.110 Median : 6.30 Median :31.40   
## Mean :258.05 Mean : 7.877 Mean :10.54 Mean :32.67   
## 3rd Qu.:329.63 3rd Qu.:10.080 3rd Qu.:19.30 3rd Qu.:41.10   
## Max. :724.42 Max. :30.530 Max. :44.00 Max. :78.10   
## NA's :8863 NA's :7007 NA's :26688 NA's :27633   
## handwashing\_facilities hospital\_beds\_per\_thousand life\_expectancy  
## Min. : 1.19 Min. : 0.100 Min. :53.28   
## 1st Qu.: 20.86 1st Qu.: 1.300 1st Qu.:67.92   
## Median : 49.84 Median : 2.400 Median :74.62   
## Mean : 50.88 Mean : 3.029 Mean :73.23   
## 3rd Qu.: 83.24 3rd Qu.: 3.861 3rd Qu.:78.74   
## Max. :100.00 Max. :13.800 Max. :86.75   
## NA's :49738 NA's :16092 NA's :4594   
## human\_development\_index  
## Min. :0.394   
## 1st Qu.:0.602   
## Median :0.748   
## Mean :0.728   
## 3rd Qu.:0.848   
## Max. :0.957   
## NA's :8675

#----------------------------------------------------------------------------  
  
#CUESTIION 1: Dataset COVIDWORLD  
#Existe afección distinta y superior de casos confirmados segun las regiones de procedencia.  
hist(Covidworld$Confirmed)



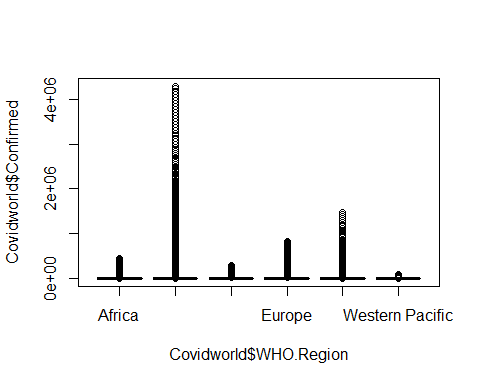
summary(Covidworld$Confirmed)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 4 168 16885 1518 4290259

anova<-aov(Covidworld$Confirmed~Covidworld$WHO.Region)  
summary.aov(anova)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Covidworld$WHO.Region 5 1.190e+13 2.381e+12 149.1 <2e-16 \*\*\*  
## Residuals 49062 7.832e+14 1.596e+10   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

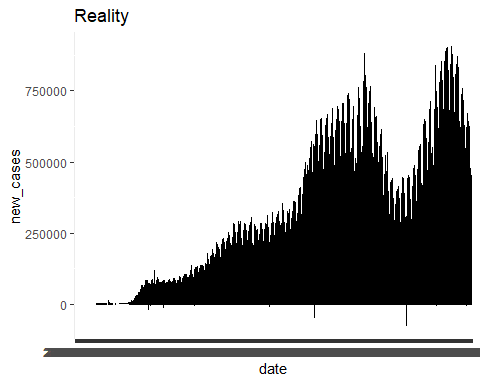
boxplot(Covidworld$Confirmed~Covidworld$WHO.Region,col=c("blue","yellow"))



#Y = 1millón, 2millones, 3millones, 4millones.  
  
#X = AFRICA, AMERICA, MEDITERRANEO ORIENTAL, EUROPA, SURESTE DE ASIA, PACIFICO OESTE.  
  
  
#H0: La region de procedencia no tiene influencia en los casos de coronavirus confirmados.  
  
#H1: La region de procedencia tiene una marcada influencia en los casos de coronavirus confirmados.  
  
#Conclusion: Tras evaluar las hipotesis por medio del anova, se determina que: rechazamos la H0, dejando a entender que la region si tiene que ver con la afluencia de casos. Se podría considerar que es debido al desarrollo de los distintos continentes.  
#----------------------------------------------------------------------------

#----------------------------------------------------------------------------  
  
#CUESTION 2: Dataset COVIDATA  
#Determinar un pronostico de un mes con relacion a los nuevos casos. Usando series de tiempo con ARIMA.  
ggplot(covidata, aes(x = date, y = new\_cases))+ geom\_line() + labs(title = 'Reality')

## Warning: Removed 106 row(s) containing missing values (geom\_path).



#En la grafica se observa el incremento de nuevos casos comprendidos entre 24-02-2020 hasta 24-05-2021. Se nota que aproximadamente durante los primeros 6 o 7 meses hay un incremente suavizado, mientras que el resto de los meses se notan incrementos exorbitantes. Puede apreciarse que durante el mes 15 empieza a decrecer un poco.  
  
nuevoscasos\_ts <- ts(covidata$new\_cases, start = 1, frequency = 30)

nuevoscasos\_ts

ajuste <- auto.arima(y = nuevoscasos\_ts)  
  
summary(ajuste)

## Series: nuevoscasos\_ts   
## ARIMA(5,1,1)(0,0,1)[30] with drift   
##   
## Coefficients:  
## ar1 ar2 ar3 ar4 ar5 ma1 sma1 drift  
## -0.1447 -0.2127 -0.2854 -0.2827 -0.2644 -0.0694 -0.0734 1.6786  
## s.e. 0.0062 0.0033 0.0032 0.0034 0.0036 0.0057 0.0037 7.1774  
##   
## sigma^2 estimated as 30242969: log likelihood=-886172.6  
## AIC=1772363 AICc=1772363 BIC=1772448  
##   
## Training set error measures:  
## ME RMSE MAE MPE MAPE MASE ACF1  
## Training set -1.032293 5499.081 869.0697 NaN Inf 0.2513392 -0.01004519

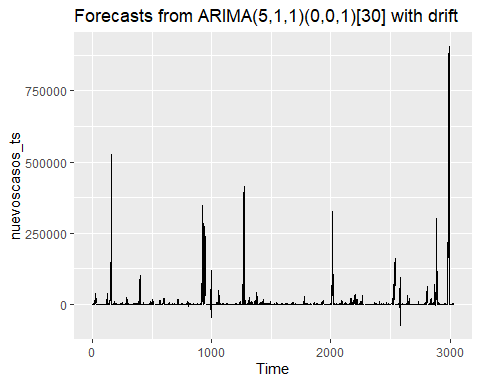
prediccion <- forecast(ajuste)  
min(prediccion[['lower']])

## [1] -36197.56

max(prediccion[['upper']])

## [1] 36450.18

predict <- autoplot(prediccion)  
predict

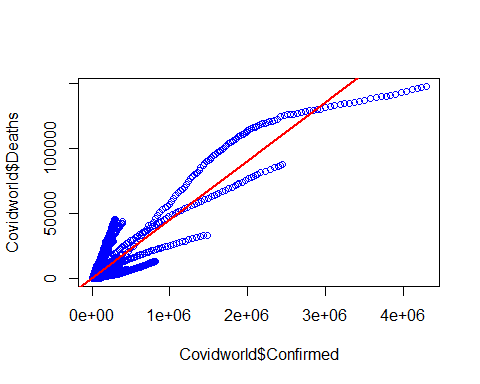


# Conclusión: En la nueva gráfica se aprecian los incremento durante un próximo mes, se haya más variación y empieza a disminuir, esto pasa segun el comportamiento final de la grafica original. Tambien se consideran el inicio de vacunacion en algunas regiones.  
#----------------------------------------------------------------------------

#----------------------------------------------------------------------------  
  
#CUESTION 3: DATASET COVIDWORLD  
#Grafique y estudie la relacion entre casos confirmados y muertes confirmadas.  
cor(Covidworld$Confirmed,Covidworld$Deaths)

## [1] 0.912361

plot(Covidworld$Confirmed,Covidworld$Deaths,col="blue")  
abline(lm(Covidworld$Deaths~Covidworld$Confirmed),col="red",lwd=2)

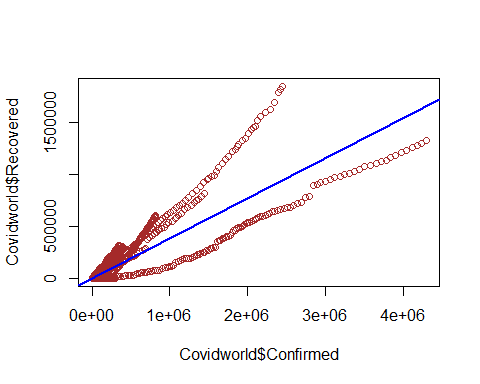


#H0: Los casos de muertes irán en incremento si y solo si los casos confirmados aumentan.  
  
#H1: Los casos de muertes no se relacionan con los casos confirmados.  
  
#Conclusion: Sucede que entre más casos se confirmen, asi mismo será el auge de muertes, obviamente no serán paralelos los resultados, pero si habrá más muertes debido a más casos confirmados. No se rechaza la hipotesis nula H0.  
#----------------------------------------------------------------------------

#----------------------------------------------------------------------------  
  
#CUESTION 4: DATASET COVIDWORLD  
#Grafique y estudie la relacion entre casos confirmados y los recuperados.  
cor(Covidworld$Confirmed,Covidworld$Recovered)

## [1] 0.895506

plot(Covidworld$Confirmed,Covidworld$Recovered,col="brown")  
abline(lm(Covidworld$Recovered~Covidworld$Confirmed),col="blue",lwd=2)

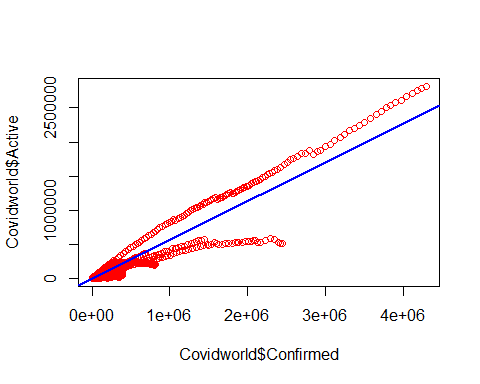


#H0: Las recuperaciones no se comparan con el incremento de nuevos casos de covid.  
  
#H1: Las recuperaciones destacan notoriamente equivalente a los nuevos casos de covid.  
  
#Conclusion: La gráfica muestra que si bien hay recuperaciones, no se comparan con el gran incremento de nuevos casos. No se rechaza la H0.  
#----------------------------------------------------------------------------

#----------------------------------------------------------------------------  
#CUESTION 5: DATASET COVIDWORLD  
#Grafique y estudie la relacion entre casos confirmados y los casos activos que existen.  
cor(Covidworld$Confirmed,Covidworld$Active)

## [1] 0.9502548

plot(Covidworld$Confirmed,Covidworld$Active,col="red")  
abline(lm(Covidworld$Active~Covidworld$Confirmed),col="blue",lwd=2)

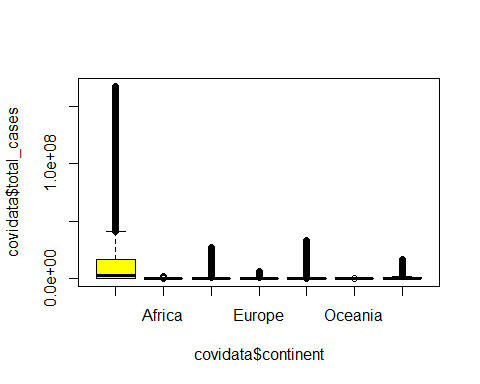


#H0: Los casos activos disminuyen al haber muchas recuperaciones de los casos confirmados.  
#H1: Los casos activos no presentan disminucion.  
#Conclusion: No se rechaza la H0. A través de la grafica se comprende que casos activos van en disminucion, ya que son los terceros a ser contados, despues de las muertes y los casos recuperados.  
#----------------------------------------------------------------------------

#----------------------------------------------------------------------------  
  
#CUESTION 6: DATASET COVIDATA  
#El total de los casos interfiere con respecto al continente en que se dan.  
anova<-aov(covidata$total\_cases~covidata$continent)  
summary.aov(anova)

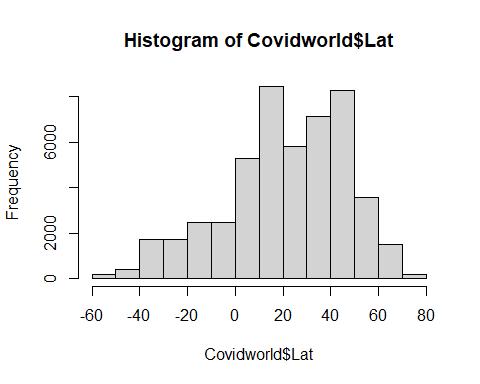
## Df Sum Sq Mean Sq F value Pr(>F)   
## covidata$continent 6 6.494e+17 1.082e+17 3317 <2e-16 \*\*\*  
## Residuals 88329 2.882e+18 3.263e+13   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 2690 observations deleted due to missingness

boxplot(covidata$total\_cases~covidata$continent,col=c("yellow","red"))



#H0: El total de casos es totalmente independiente al continente en el que se da.  
  
#H1: El total de casos si tiene relacion dependiendo del continente donde se da.  
  
#Conclusion: Rechazamos la H0 (hipotesis nula) lo que nos deja al descubierto que si existen razones varias por las que algunos continentes presentan cantidades notables y diferentes a otros continentes que suelen tener mejor desarrollo. Por ejemplo en una parte de america es donde se ve el gran incremento.  
#----------------------------------------------------------------------------

#----------------------------------------------------------------------------  
  
#CUESTION 7: DATASET COVIDWORLD  
#Evaluar si la latitud de la region tiene influencia sobre el numero de casos confirmados.  
hist(Covidworld$Lat)



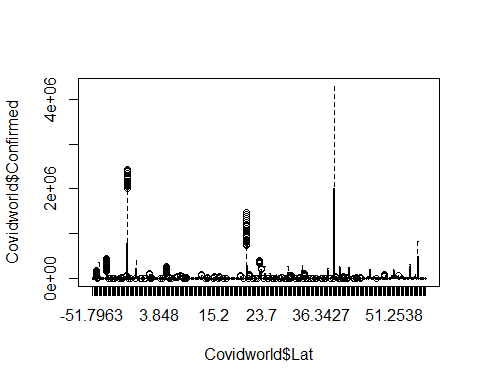
summary(Covidworld$Confirmed)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 4 168 16885 1518 4290259

anova<-aov(Covidworld$Confirmed~Covidworld$Lat)  
summary.aov(anova)

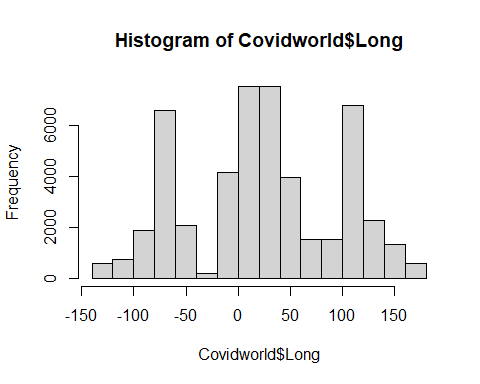
## Df Sum Sq Mean Sq F value Pr(>F)   
## Covidworld$Lat 1 1.069e+12 1.069e+12 66.05 4.5e-16 \*\*\*  
## Residuals 49066 7.941e+14 1.618e+10   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

boxplot(Covidworld$Confirmed~Covidworld$Lat,col=c("yellow","red"))



#H0: No influye la latitud sobre los casos confirmados.  
  
#H1: Si influye la latitud sobre los casos confirmados.  
  
#Conclusion: El valor de value es menor al de significancia. Por esto rechazamos la H0. Lo que da a entender que si existe alguna influencia en la latitud con relacion a los casos confirmados de covid.  
#----------------------------------------------------------------------------

#----------------------------------------------------------------------------  
#CUESTION 8: COVIDWORLD  
#//Relacion existente entre la longitud con los datos confirmados.  
  
hist(Covidworld$Long)



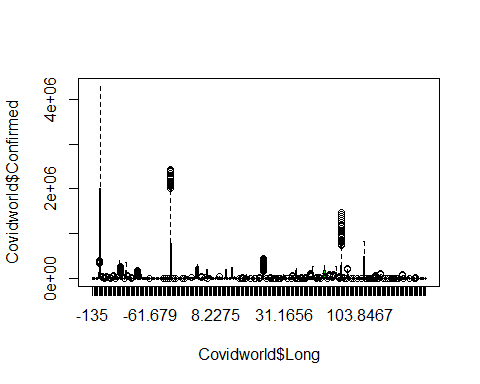
summary(Covidworld$Confirmed)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 4 168 16885 1518 4290259

anova<-aov(Covidworld$Confirmed~Covidworld$Long)  
summary.aov(anova)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Covidworld$Long 1 4.951e+12 4.951e+12 307.4 <2e-16 \*\*\*  
## Residuals 49066 7.902e+14 1.610e+10   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

boxplot(Covidworld$Confirmed~Covidworld$Long,col=c("blue","green"))

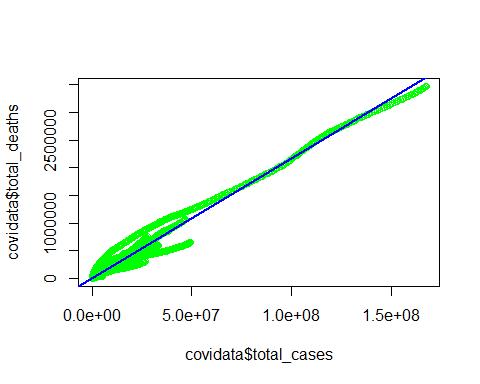


#H0: Depende de la longitud que hayan incremento en los casos de covid.  
  
#H1: No existe relacion.  
  
#CONCLUSION: Aparentemente, al rechazar la H0, se deja al descubierto que no hay relacion entre la longitud con el incremento de casos. Esto apoya hasta cierto punto las creencias sobre que el virus se debilitaba en áreas cercanas al ecuador. Ya que la longitud con mas frecuencia es aprox. -135 y esto en latitud 0 da al oceano, sin embargo al buscar tierra se puede llegar hasta el norte de america, un lugar conocidamente frio.  
#----------------------------------------------------------------------------

#----------------------------------------------------------------------------  
#CUESTION 9: COVIDATA  
#//Estudiar la correlacion entre total\_cases con total\_deaths, detallar hipotesis.  
  
cor(covidata$total\_cases,covidata$total\_deaths)

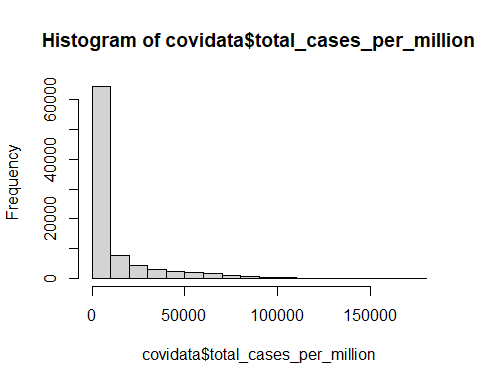
## [1] NA

plot(covidata$total\_cases,covidata$total\_deaths,col="green")  
abline(lm(covidata$total\_deaths~covidata$total\_cases),col="blue",lwd=2)



#CONCLUSION: En la grafica se nota que existe una elevacion de defunciones, sin embargo existe por encima el numero total de casos, por que se incluyen nuevos casos, casos de vacunados entre otros. Por que si fueran simetricos esto sería el caos total. Notando en la realidad nos damos cuenta que el numero de casos cada vez era mayor y así mismo, aunque no senejante, el numero de muertes a nivel mundial.  
#----------------------------------------------------------------------------

#----------------------------------------------------------------------------  
#CUESTION 10: COVIDATA  
#//Estudie el comportamietno entre la variable Continent con la variable total de casos por millon.  
hist(covidata$total\_cases\_per\_million)



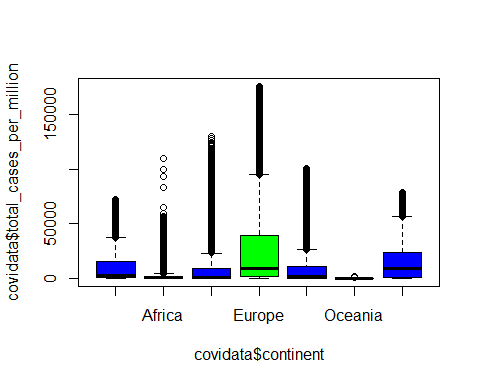
#Nos muestra que los casos por millon en su mayoria daban o, sim embargo se nota un incremento conforme avanza.  
summary(covidata$total\_cases\_per\_million)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.0 220.9 1514.1 11220.1 11074.4 175616.4 3163

anova<-aov(covidata$total\_cases\_per\_million~covidata$continent)  
summary.aov(anova)

## Df Sum Sq Mean Sq F value Pr(>F)   
## covidata$continent 6 6.071e+12 1.012e+12 2652 <2e-16 \*\*\*  
## Residuals 87856 3.352e+13 3.815e+08   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 3163 observations deleted due to missingness

boxplot(covidata$total\_cases\_per\_million~covidata$continent,col=c("blue","green"))

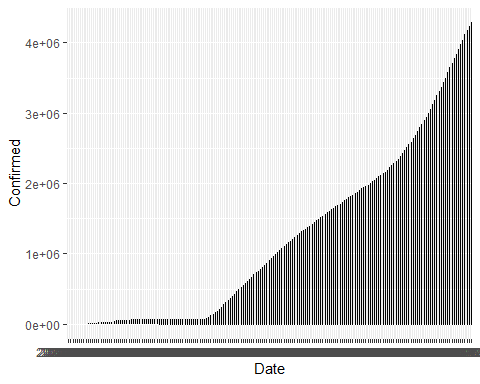


#H0: El continente con interfiere en la estadistica de casos por millon.  
  
#H1: Los casos por millos dependen del continente en cuestion.  
  
#CONCLUSION: Se rechaza la H0. Se concuerda en que dependiendo el continente se notan diferencia muy notorias. El continente con una media mayor y datos fuera del margen es Europa.   
#----------------------------------------------------------------------------

#----------------------------------------------------------------------------  
#CUESTION 11: COVIDWORLD  
#//Prediccion de casos confirmados para los siguientes 2 meses.  
library(urca)

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

ggplot(Covidworld, aes(x = Date, y = Confirmed))+ geom\_line()



confirm\_ts <- ts(Covidworld$Confirmed, start = c(2020,01,22), frequency = 60) #60 dias = 2 meses

confirm\_ts

ur.kpss(confirm\_ts) %>% summary()

##   
## #######################   
## # KPSS Unit Root Test #   
## #######################   
##   
## Test is of type: mu with 18 lags.   
##   
## Value of test-statistic is: 68.6768   
##   
## Critical value for a significance level of:   
## 10pct 5pct 2.5pct 1pct  
## critical values 0.347 0.463 0.574 0.739

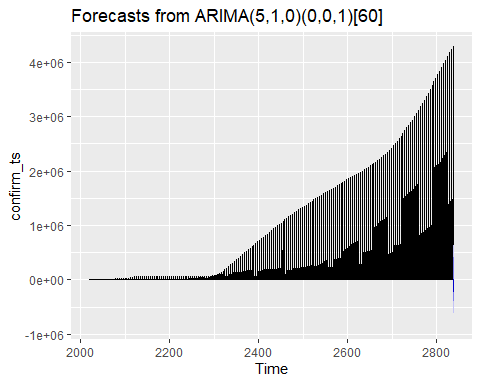
ur.kpss(diff(confirm\_ts)) %>% summary() #Calculando el Value

##   
## #######################   
## # KPSS Unit Root Test #   
## #######################   
##   
## Test is of type: mu with 18 lags.   
##   
## Value of test-statistic is: 2e-04   
##   
## Critical value for a significance level of:   
## 10pct 5pct 2.5pct 1pct  
## critical values 0.347 0.463 0.574 0.739

model <- auto.arima(y = confirm\_ts)  
summary(model)

## Series: confirm\_ts   
## ARIMA(5,1,0)(0,0,1)[60]   
##   
## Coefficients:  
## ar1 ar2 ar3 ar4 ar5 sma1  
## -0.8495 -0.6156 -0.4844 -0.3469 -0.1861 -0.0528  
## s.e. 0.0044 0.0057 0.0060 0.0057 0.0046 0.0048  
##   
## sigma^2 estimated as 1.838e+10: log likelihood=-649464.7  
## AIC=1298943 AICc=1298943 BIC=1299005  
##   
## Training set error measures:  
## ME RMSE MAE MPE MAPE MASE ACF1  
## Training set 0.7257355 135580.6 28145.09 NaN Inf 0.8972703 -0.02217374

prediction <- forecast(model)  
pronostic <- autoplot(prediction)  
pronostic



#H0: El pronostico de casos va en aumento con el pasar de los proximos dos meses.   
#H1: El pronostico de casos dismunuye.   
#CONCLUSION:Rechazamos la hipotesis nula debido al Value calculado de "2e-04". Nos permite notar que la prediccion muestra algunos niveles de disminucion (puede verse en la grafica de "pronostic"), debido a que el pico más alto ya se encuentra registrado.   
#----------------------------------------------------------------------------

#/////////////////////////////////////////////////////////////////////////////  
# #  
# #  
# UTILIZANDO OTRO DATASET, AJENO A LOS ANALISIS ANTERIORES #  
# #  
# #  
#/////////////////////////////////////////////////////////////////////////////

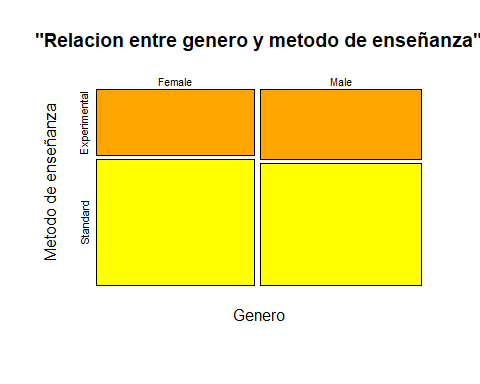
#----------------------------------------------------------------------------  
#CUESTION 12: DATASET SCORES  
Scores <- read.csv('D:/Estadistica2/test\_scores.csv')  
#Averiguar que sucede al relacionar la variable de gender con teaching\_method  
  
tabla2 <- table(Scores$gender, Scores$teaching\_method, dnn = c("Genero", "Metodo de enseñanza"))  
tabla2

## Metodo de enseñanza  
## Genero Experimental Standard  
## Female 366 690  
## Male 394 683

chisq.test(tabla2)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: tabla2  
## X-squared = 0.77875, df = 1, p-value = 0.3775

mosaicplot(tabla2, main = deparse("Relacion entre genero y metodo de enseñanza"), col=c("orange", "yellow"))



#H0: Existe cierta diferencia entre el metodo de enseñanza segun el genero.  
#H1: No existe diferencia entre el metodo de enseñanza segun el genero.  
  
#CONCLUSION: No rechazamos la H0. Por ende, se dice que si que existe una leve diferencia y es que el genero femenino recibe menos educacion experimental que los del genero masculino.   
#----------------------------------------------------------------------------

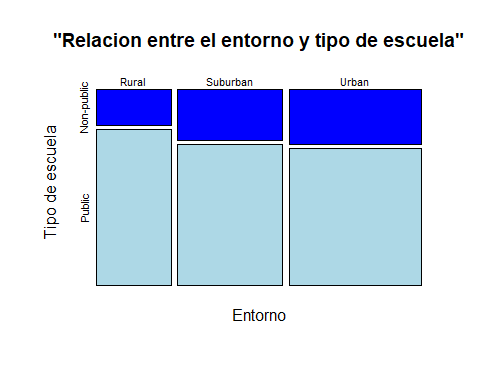
#----------------------------------------------------------------------------  
#CUESTION 13: DATASET SCORES  
#Estudiar la relacion entre si el entorno (school\_setting) tiene que ver con tipo de escuela (school\_type)  
  
tabla3 <- table(Scores$school\_setting, Scores$school\_type, dnn = c("Entorno", "Tipo de escuela"))  
tabla3

## Tipo de escuela  
## Entorno Non-public Public  
## Rural 97 413  
## Suburban 193 524  
## Urban 261 645

chisq.test(tabla3)

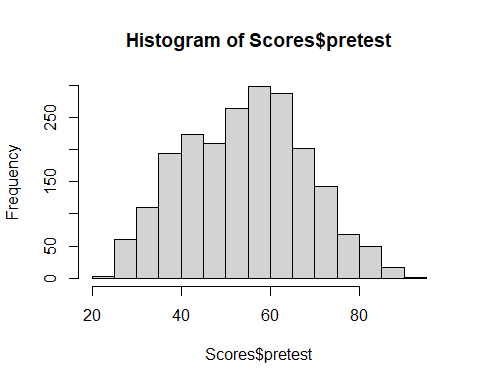
##   
## Pearson's Chi-squared test  
##   
## data: tabla3  
## X-squared = 16.983, df = 2, p-value = 0.0002052

mosaicplot(tabla3, main = deparse("Relacion entre el entorno y tipo de escuela"), col=c("blue", "lightblue"))



#H0: No existe distincion por el entorno en relacion a tipo de escuelas.  
#H1: Existe diferencia evidente segun el entorno asi habra escuelas.   
  
#CONCLUSION: El pvalue nos hace rechazar la H0. Y notando en la grafica se ve que existe una diferencia obviamente evidente y significativa del tipo de escuelas con mayor numero en los distintos entornos.  
#----------------------------------------------------------------------------

#----------------------------------------------------------------------------  
#CUESTION 14: SCORES  
#Descubrir la relacion existente entre la nota pre test y el metodo de enseñanza.  
  
hist(Scores$pretest)



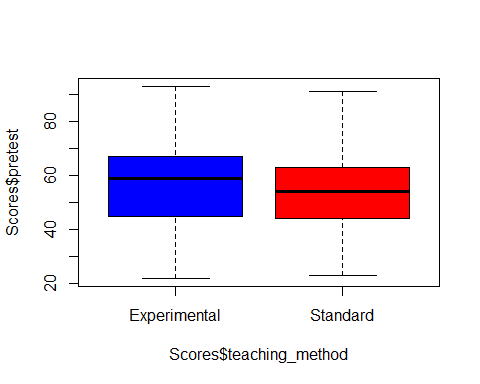
summary(Scores$pretest)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 22.00 44.00 56.00 54.96 65.00 93.00

t.test(Scores$pretest~Scores$teaching\_method, alternative = "two.sided")

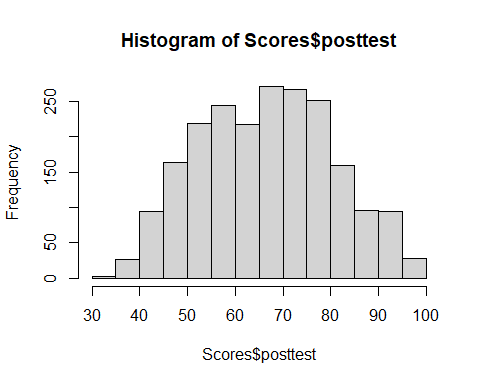
##   
## Welch Two Sample t-test  
##   
## data: Scores$pretest by Scores$teaching\_method  
## t = 5.2572, df = 1485.2, p-value = 1.676e-07  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 2.044503 4.478259  
## sample estimates:  
## mean in group Experimental mean in group Standard   
## 57.05526 53.79388

boxplot(Scores$pretest~Scores$teaching\_method, col=c("blue","red"))



#H0: No se nota distincion alguna entre el metodo de enseñanza hacia la nota preliminar.  
#H1: Se nota distincion dependiendo el metodo de enseñanza así se marcara en la nota preliminar.  
  
#CONCLUSION: Rechazamos la hipotesis nula. Y si se fija en la grafica se deja ver que los de metodo experimental tiene mayor logro que los de metodo standard.  
#----------------------------------------------------------------------------

#----------------------------------------------------------------------------  
#CUESTION 15: DATASET SCORES  
  
hist(Scores$posttest)



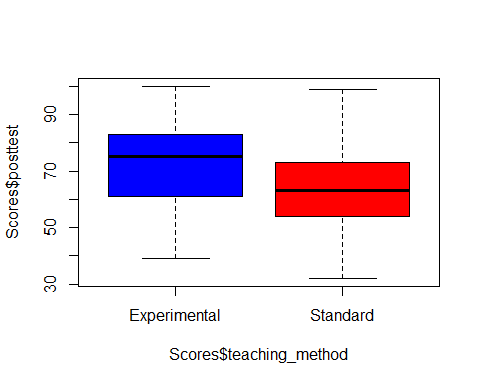
summary(Scores$posttest)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 32.0 56.0 68.0 67.1 77.0 100.0

t.test(Scores$posttest~Scores$teaching\_method, alternative = "two.sided")

##   
## Welch Two Sample t-test  
##   
## data: Scores$posttest by Scores$teaching\_method  
## t = 14.981, df = 1498.6, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 7.939642 10.332047  
## sample estimates:  
## mean in group Experimental mean in group Standard   
## 72.98289 63.84705

boxplot(Scores$posttest~Scores$teaching\_method, col=c("blue","red"))

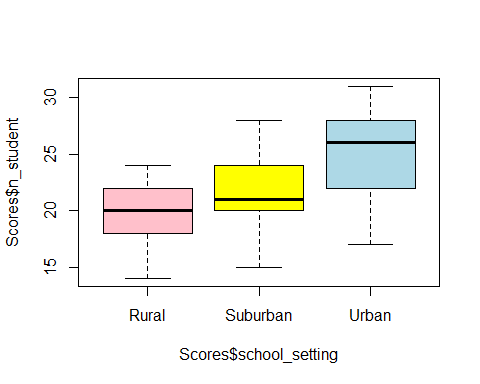


#H0: La nota posttest no es consecuente al metodo de enseñanza.  
#H1: La nota del posttest depende totalmente del metodo de enseñanza.  
  
#CONCLUSION: Basta estudiar el pvalue y la grafica para notar que la H0 es rechazada y se percibe que se destaca muy evidentemente cuando el metodo de educacion es experimental.   
#----------------------------------------------------------------------------

#----------------------------------------------------------------------------  
#CUESTION 16: DATASET SCORES  
#//Evlaue la relacion entre el numero de alumnos(n\_students) para con el entorno escolar(school\_setting)  
anova <- aov(Scores$n\_student~Scores$school\_setting)  
summary.aov(anova)

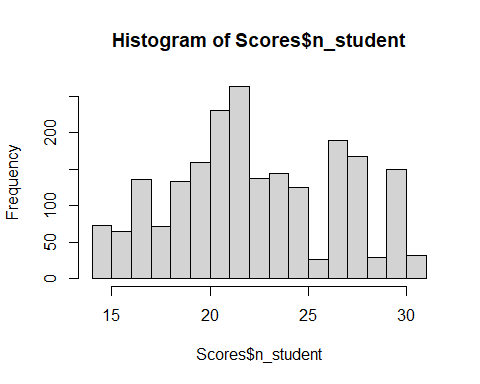
## Df Sum Sq Mean Sq F value Pr(>F)   
## Scores$school\_setting 2 10254 5127 391.8 <2e-16 \*\*\*  
## Residuals 2130 27874 13   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

boxplot(Scores$n\_student~Scores$school\_setting, col=c("pink", "yellow", "lightblue"))



#H0: En el numero de estudiantes no hay diferencia en cuanto a entorno escolar.  
#H1: Si existe mayor numero de estudiantes dependiendo el entorno escolar.  
  
#CONCLUSION: El value nos hace rechazar la H0. Se nota despues que en la grafica se refleja que en el entorno urbano existe mas demanda estudiantil y en la rural menos, evidentemente.  
#----------------------------------------------------------------------------

#----------------------------------------------------------------------------  
#CUESTION 17: DATASET SCORES  
hist(Scores$n\_student)



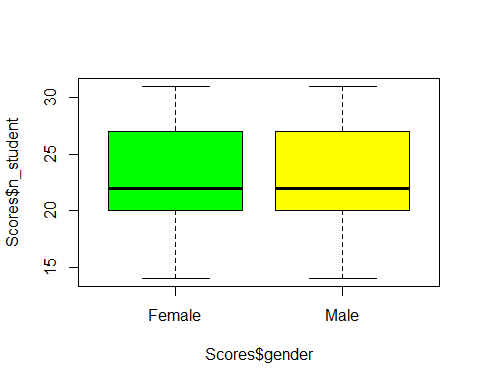
summary(Scores$n\_students)

## Length Class Mode   
## 0 NULL NULL

t.test(Scores$n\_student~Scores$gender, alternative = "two.sided")

##   
## Welch Two Sample t-test  
##   
## data: Scores$n\_student by Scores$gender  
## t = 0.88941, df = 2129.1, p-value = 0.3739  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.1962935 0.5221144  
## sample estimates:  
## mean in group Female mean in group Male   
## 22.87879 22.71588

boxplot(Scores$n\_student~Scores$gender, col=c("green","yellow"))



#H0: No hay diferencias por ser mujer u hombre en cuanto al numero de estudiantes por salon.  
#H1: Si hay diferencias por ser mujer u hombre en cuanto al numero de estudiantes por salon  
  
#CONCLUSION: No rechazamos la H0. Lo que nos indica que se estima una equitatividad en los resultados. Y ya podemos dar fe de este resultado con solo consultar la grafica, la cual nos dibuja la aparente igualdad en numero entre hombres y mujeres en cuanto al numero por salon.  
#----------------------------------------------------------------------------