Entrega Final: Proyecto programación

31/05/2021

Buen desempeño económico, la clave del éxito para un rendimiento sobresaliente en los juegos olímpicos.

El ojetivo de este trabajo es determinar los efectos de las principales variables macroeconómicas en la cantidad de medallas olímpicas obtenidas en promedio.

Integrantes de grupo.

- *Joan Galeano R*
- *Nicolás González J*
- *Alejandro Guevara H*

Paquetes

Para cumplir el objetivo del trabajo hacemos uso de RStudio como herramienta metodológica y de análisis. Se utilizan varios paquetes para el analisis, para el manejo de datos, graficos y modelos entre otros.

```
library(tidyverse)
library(rvest)
library(haven)
library(wbstats)
library(dplyr)
library(naniar)
library(knitr)
library(ggthemes)
library(readxl)
library(GGally)
library(Hmisc)
library(corrplot)
library(PerformanceAnalytics)
library(stargazer)
```

Manejo de la base de datos

Variable explicada

Organizamos la base de datos, eliminamos datos faltantes y renombramos algunos países para evitar problemas (p.e: The Bahamas -> Bahamas, The) y de esta manera todas los nombres coinciden en las diferentes bases de datos sin afectar los datos de interés.

```
pagina <-"http://www.olympedia.org/statistics/medal/country"
pagina_desc <- read_html(pagina)

paises <- pagina_desc %>% html_nodes("td:nth-child(1)") %>% html_text()

medallas <- pagina_desc %>% html_nodes("td:nth-child(6)") %>% html_text()
medallas <- as.integer(medallas)
medallas_por_pais <- tibble(paises, medallas)

medallas_por_pais[4,1] <- "United Kingdom"
medallas_por_pais[7,1] <- "China"
medallas_por_pais[18,1] <- "Korea, Rep."
medallas_por_pais[43,1] <- "Iran, Islamic Rep."
medallas_por_pais[50,1] <- "Slovak Republic"
medallas_por_pais[59,1] <- "Egypt, Arab Rep."
medallas_por_pais[62,1] <- "Bahamas, The"
medallas_por_pais[83,1] <- "Venezuela, RB"
medallas_por_pais[84,1] <- "Serbia"
medallas_por_pais[97,1] <- "Cote d'Ivoire"
medallas_por_pais[98,1] <- "Hong Kong SAR, China"
medallas_por_pais[113,1] <- "Moldova"
medallas_por_pais[117,1] <- "Tanzania"
medallas_por_pais[121,1] <- "Kyrgyz Republic"
medallas_por_pais[122,1] <- "Saudi Arabia"
```

Variables explicativas

En este caso, se presenta la Tasa de crecimiento del PIB como primera variable explicativa.

```
growth_gdp <- wb_data("NY.GDP.PCAP.KD.ZG", start_date = 1950, end_date = 2016)
growth_gdp <- tibble(growth_gdp$country,growth_gdp$date,growth_gdp$NY.GDP.PCAP.KD.ZG)
growth_gdp <- growth_gdp %>%
  rename(paises = `growth_gdp$country`,
    fecha = `growth_gdp$date`,
    growth = `growth_gdp$NY.GDP.PCAP.KD.ZG`
      )
growth_gdp <- drop_na(growth_gdp)

growth_gdp_prom <- aggregate(growth_gdp$growth, list(growth_gdp$paises), FUN=mean)
growth_gdp_prom <- growth_gdp_prom %>%
  rename(
    paises = Group.1, GDP=x
    ) %>%
  mutate(GDP2=GDP^2)
```

A continuación, se usará la Tasa de crecimiento de la población.

```
growth_pob <- wb_data("SP.POP.GROW", start_date = 1950, end_date = 2016)
growth_pob <- tibble(growth_pob$country,growth_pob$date,growth_pob$SP.POP.GROW)
growth_pob <- growth_pob %>%
  rename(
    paises = "growth_pob$country",
```

```
    fecha ="growth_pob$date",
    growth_p ="growth_pob$SP.POP.GROW"
      )
growth_pob <-drop_na(growth_pob)

growth_pob_prom <- aggregate(growth_pob$growth_p, list(growth_pob$paises), FUN=mean)
growth_pob_prom <- growth_pob_prom %>%
  rename(
    paises = Group.1, POB = x
    ) %>%
  mutate(POB2=POB^2)
```

También se decidió incluir la Tasa promedio de paro.

```
desempleo_total <- wb_data("SL.UEM.TOTL.ZS", start_date = 1950, end_date = 2016)
desempleo_total <- tibble(desempleo_total$country,desempleo_total$date,desempleo_total$SL.UEM.TOTL.ZS)
desempleo_total <-desempleo_total%>%
  rename(
    paises = "desempleo_total$country",
    fecha ="desempleo_total$date",
    desempleo_t ="desempleo_total$SL.UEM.TOTL.ZS"
      )
desempleo_total <-drop_na(desempleo_total)
desempleo_total_prom <- aggregate(desempleo_total$desempleo_t, list(desempleo_total$paises), FUN=mean)
desempleo_total_prom <- desempleo_total_prom %>%
  rename(
    paises = Group.1, DESP=x
    )
```

Se incluirá la Tasa promedio de inflación para explicar su efecto parcial sobre el desempeño deportivo en los juegos olímpicos.

```
inflacion <- wb_data("NY.GDP.DEFL.KD.ZG", start_date = 1950, end_date = 2016)
inflacion<- tibble(inflacion$country,inflacion$date,inflacion$NY.GDP.DEFL.KD.ZG)
inflacion <- inflacion %>%
  rename(
    paises = "inflacion$country",
    fecha = "inflacion$date",
   inflacion_t ="inflacion$NY.GDP.DEFL.KD.ZG"
      )
inflacion <- drop_na(inflacion)

inflacion_prom <- aggregate(inflacion$inflacion_t, list(inflacion$paises), FUN=mean)
inflacion_prom <-inflacion_prom %>%
  rename(
    paises = Group.1,INF=x
    ) %>%
  mutate(INF2=INF^2)
```

Finalmente, se hace un proceso de adjunción de todas las variables en un mismo objeto.

```
datos <- left_join(medallas_por_pais, growth_pob_prom,by = "paises")
datos1 <- left_join(datos,growth_gdp_prom,by = "paises")
datos2 <- left_join(datos1, desempleo_total_prom,by = "paises")
tidy_data<- left_join(datos2, inflacion_prom,by = "paises")

tidy_data <- drop_na(tidy_data)
tidy_data <- tidy_data %>%
  relocate(paises,medallas,GDP,POB,DESP,INF)
head(tidy_data)
```

```
## # A tibble: 6 x 9
##   paises         medallas   GDP   POB  DESP   INF   POB2  GDP2  INF2
##   <chr>             <int> <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1 United States      2847  2.00 1.04   6.07  3.40 1.08    4.00 11.6
## 2 Germany            1019  1.88 0.220  7.72  2.55 0.0483  3.52  6.51
## 3 United Kingdom      898  2.03 0.402  6.74  5.61 0.161   4.13 31.4
## 4 France              863  2.18 0.640  9.88  4.28 0.410   4.74 18.3
## 5 Italy               716  2.11 0.337  9.90  6.65 0.114   4.47 44.3
## 6 China               608  6.85 1.30   3.88  3.60 1.68   46.9  13.0
```

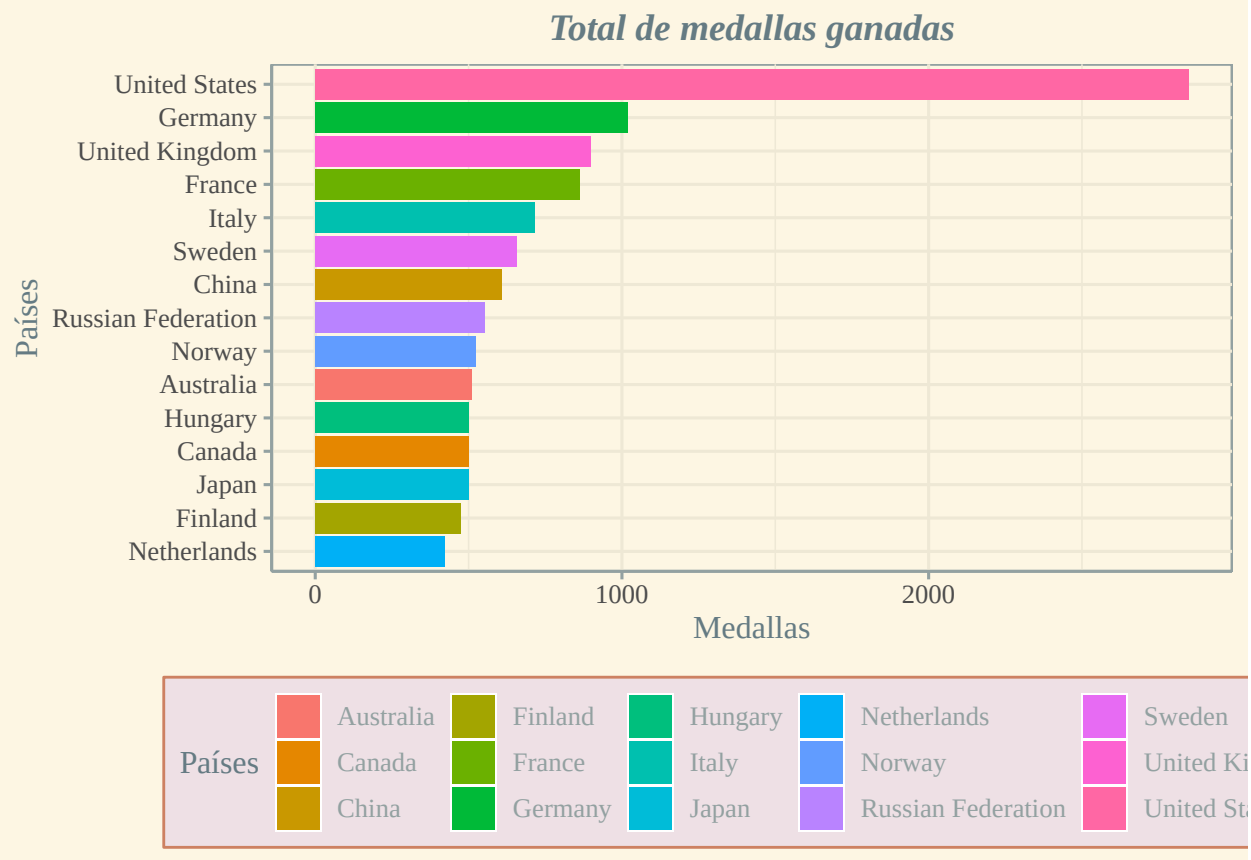
Estadísticas descriptivas

```
resumen_paises <- tidy_data %>%
  summary()
resumen_paises
```

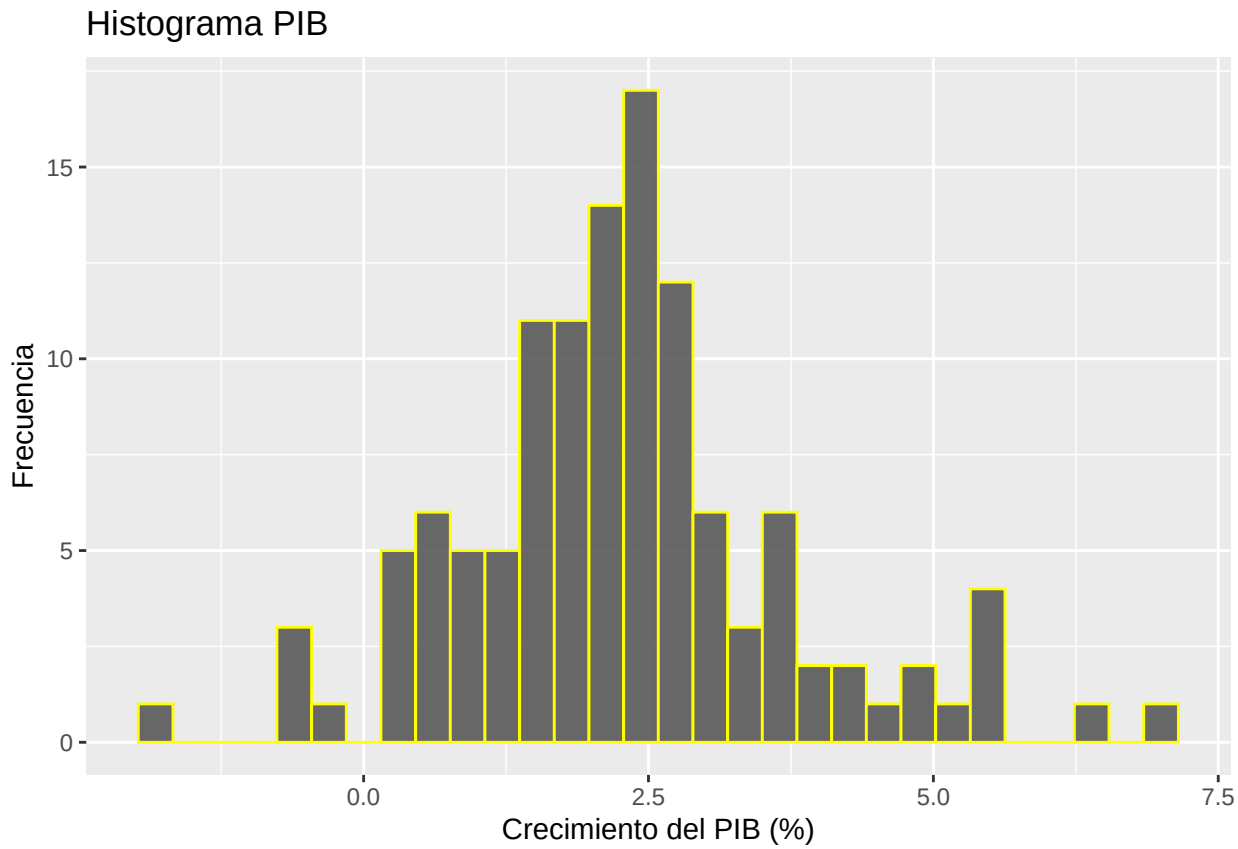
```
##     paises             medallas            GDP              POB
##  Length:120         Min.   :   1.00   Min.   :-1.976   Min.   :-0.1763
##  Class :character   1st Qu.:   2.00   1st Qu.: 1.453   1st Qu.: 0.5564
##  Mode  :character   Median :  14.50   Median : 2.210   Median : 1.6006
##                     Mean   : 136.15   Mean   : 2.274   Mean   : 1.6837
##                     3rd Qu.:  97.75   3rd Qu.: 2.866   3rd Qu.: 2.3548
##                     Max.   :2847.00   Max.   : 6.848   Max.   : 8.2464
##       DESP              INF              POB2              GDP2
##  Min.   : 0.4885   Min.   :  1.201   Min.   : 0.00022   Min.   : 0.02327
##  1st Qu.: 4.0865   1st Qu.:  4.682   1st Qu.: 0.30972   1st Qu.: 2.13591
##  Median : 7.2510   Median :  7.670   Median : 2.56410   Median : 4.88603
##  Mean   : 8.3105   Mean   : 26.795   Mean   : 4.69046   Mean   : 7.23571
##  3rd Qu.:11.2228   3rd Qu.: 17.474   3rd Qu.: 5.54519   3rd Qu.: 8.21238
##  Max.   :33.1550   Max.   :455.599   Max.   :68.00310   Max.   :46.89291
##       INF2
##  Min.   :     1.44
##  1st Qu.:    21.92
##  Median :    58.85
##  Mean   :  4315.70
##  3rd Qu.:   305.50
##  Max.   :207570.06
```

Medallas y PIB

```
tidy_data %>%
  group_by(paises,medallas) %>%
  head(15)%>%
ggplot(aes(x=reorder(paises,medallas),y=medallas, fill=paises)) +
geom_col() + coord_flip() +  labs(title = "Total de medallas ganadas", x= "Países", y= "Medallas")+
theme_solarized(light = T)  + scale_colour_solarized('green')+theme(text = element_text(family = "serif
plot.title = element_text(face = "bold.italic", hjust = 0.5))+  theme(legend.position = "bottom")+theme
```



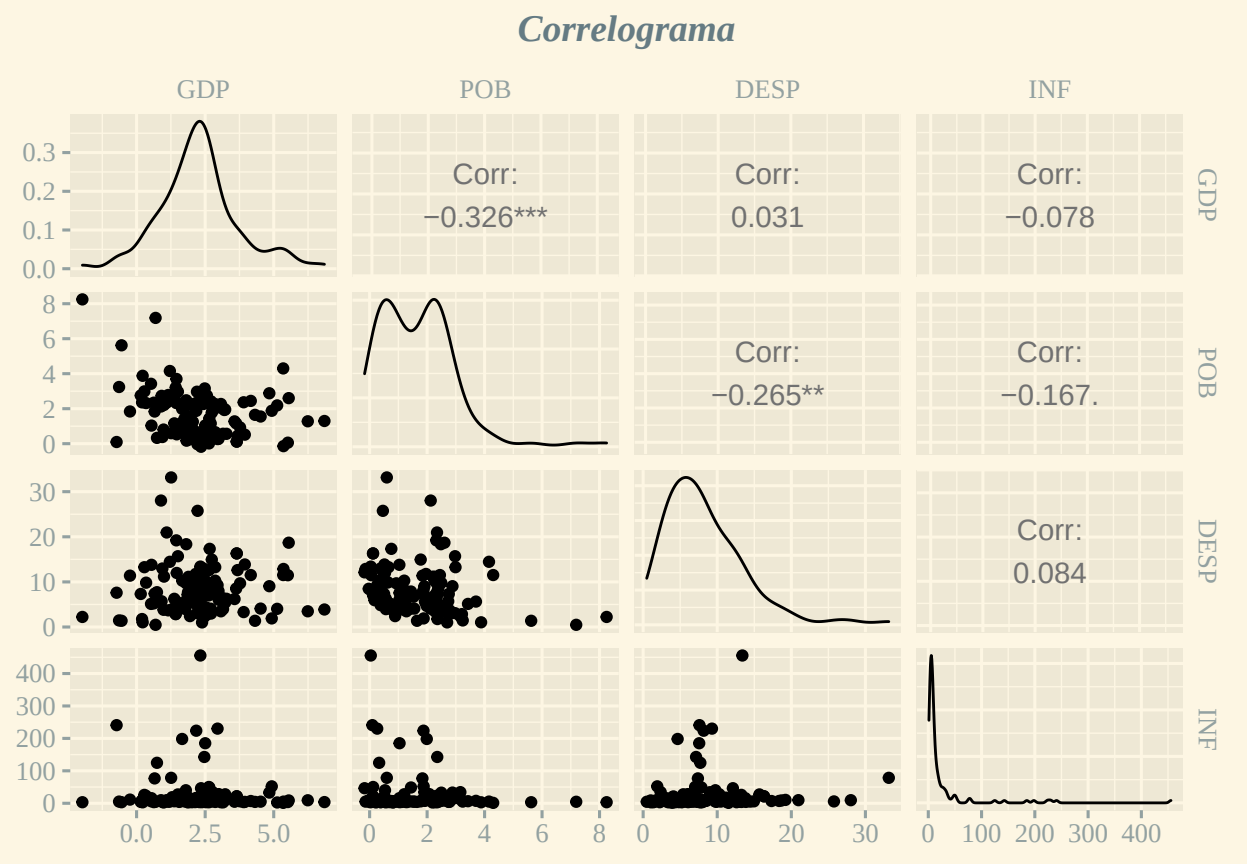

```
ggplot(tidy_data, aes(x = tidy_data$GDP)) +
  geom_histogram(position = "identity", color = "Yellow",alpha = 0.9)+
  theme(legend.position = "top") +
  ggtitle("Histograma PIB") +
  ylab("Frecuencia") +
  xlab("Crecimiento del PIB (%)")
```



Al analizar los diagramas de dispersión, se observa que la gran mayoría de pares ordenados presentan correlaciones negativas entre sí, a excepción de la relación entre la inflación y el desempleo. Se encuentra además que la matriz de correlación expone 3 ejemplares significativos. La primera es el desempleo y la población, la cual llega a ser significativa a al 5%, y es de carácter negativa. La segunda es la relación entre el desempleo y el PIB, la cual es significativa al 10% e inversa. La tercera es la relación entre la inflación y el PIB, la cual es significativa al 10% y presenta una relación inversa entre sí. Finalmente, se encuentra que la relación entre la inflación y el desempleo es de carácter positiva, por lo que se establece que mayores niveles de inflación están asociados con mayores niveles de desempleo, tal y como afirma la teoría económica al describir el postulado de Curva de Phillips.

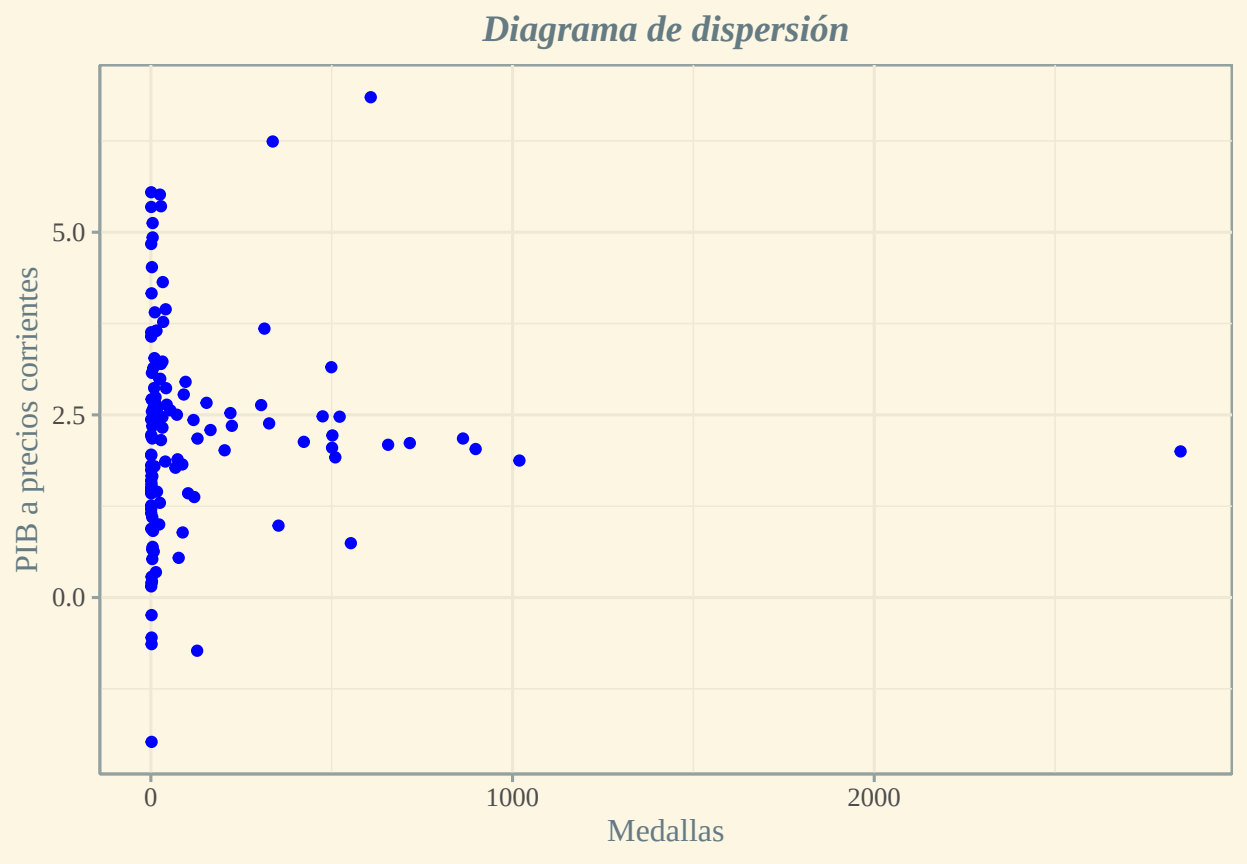
Correlaciones

```
ggpairs(tidy_data, columns = 3:6, method = c("everything", "pearson"),title="Correlograma",color="red")+
plot.title = element_text(face = "bold.italic", hjust = 0.5))+  theme(legend.position = "bottom")+theme
```

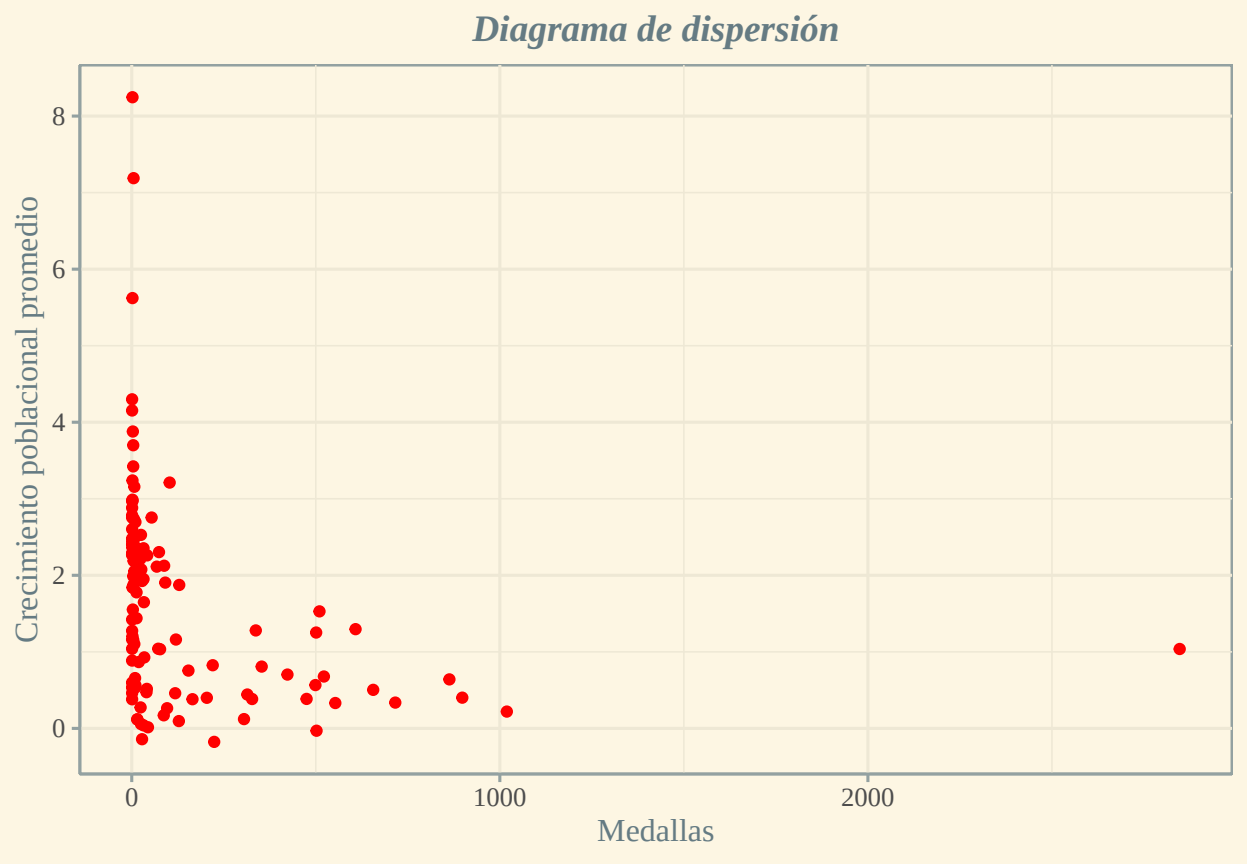


Dispersión de los datos

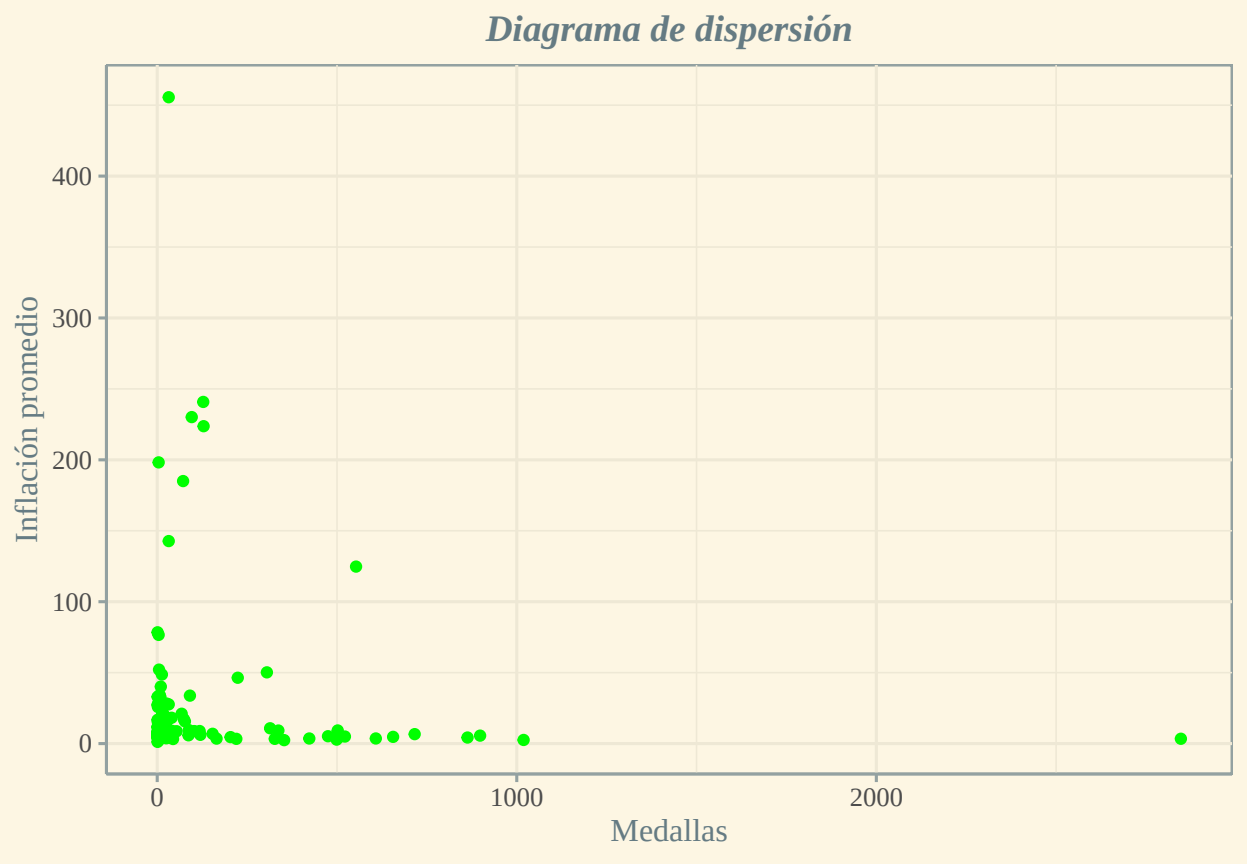
```
ggplot(data = tidy_data, aes(x =medallas,y =GDP)) +
geom_point(color="blue") +
labs(title = "Diagrama de dispersión", x= "Medallas", y= "PIB a precios corrientes")+theme_solarized(li
plot.title = element_text(face = "bold.italic", hjust = 0.5))+  theme(legend.position = "bottom")+theme
```



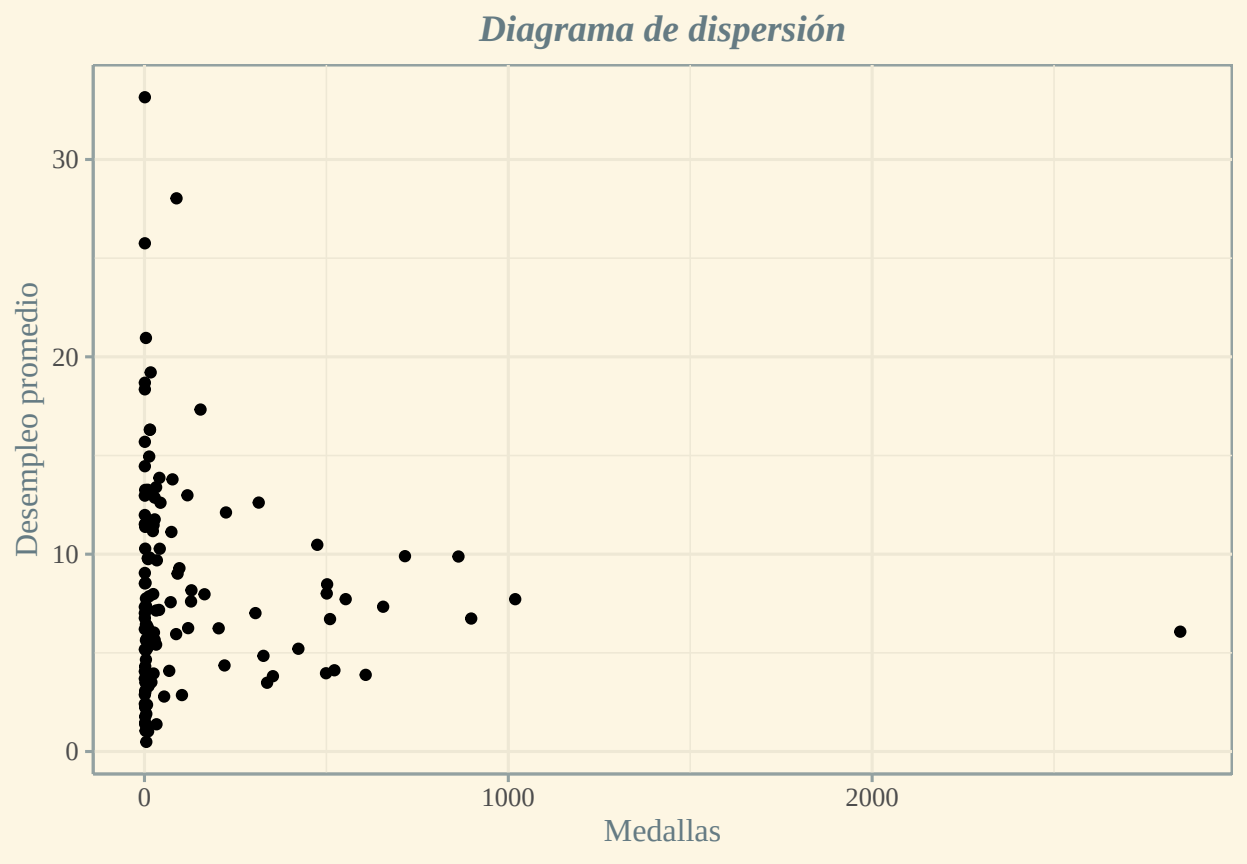

```
ggplot(data = tidy_data, aes(x =medallas,y =POB)) +
geom_point(color="red") +
labs(title = "Diagrama de dispersión", x= "Medallas", y= "Crecimiento poblacional promedio")+ theme_sola
plot.title = element_text(face = "bold.italic", hjust = 0.5))+  theme(legend.position = "bottom")+theme
```




```
ggplot(data = tidy_data, aes(x =medallas,y =INF)) +
geom_point(color="green") +
labs(title = "Diagrama de dispersión", x= "Medallas", y= "Inflación promedio")+theme_solarized(light = 
plot.title = element_text(face = "bold.italic", hjust = 0.5))+  theme(legend.position = "bottom")+theme
```




```
ggplot(data = tidy_data, aes(x =medallas,y =DESP)) +
geom_point(color = "black") +
labs(title = "Diagrama de dispersión", x= "Medallas", y= "Desempleo promedio")+theme_solarized(light = 
plot.title = element_text(face = "bold.italic", hjust = 0.5))+  theme(legend.position = "bottom")+theme
```



Análisis preliminar

Al haber analizado las estadísticas descriptas presentadas en los puntos anteriores, se logra observar sin lugar a duda que Estados Unidos es el país que mayor número de medallas (entre ellas; oro, plata y bronce) ha sumado a lo largo de la historia. Esto podría ser explicado parcialmente por el tamaño de la población, dado que es uno de los países participantes con mayor número de habitantes. Con respecto al análisis agregado junto con los demás países, se encuentra que la media de medallas ganadas por país desde 1950 hasta la actualidad corresponde a 132, sin embargo, este valor es bastante sensibles ante valores atípicos, como lo indica la participación de Estados Unidos en los juegos internacionales.

En cuanto al PIB corriente expresado en miles de millones de dólares, se evidencia que la media corresponde a 7829mm de dólares norteamericanos, además, el grado de variabilidad es bastante alto debido a que el nivel mínimo de PIB corresponde a 159mm de dólares norteamericanos, y el máximo a 40313mm. Al haber observado el histograma de la variable anterior, se encuentra que la gran mayoría de los datos están concentrados entre el mínimo y el promedio, lo cual deja en evidencia que gran parte de los países participantes poseen bajos niveles de PIB. Además, la distribución de la variable de acuerdo con el histograma llega a presentar asimetría positiva, o hacia la derecha, lo cual establece que hay diversos niveles de producto al interior las naciones partícipes de los juegos olímpicos, donde pocos países poseen ingresos muy altos, y la gran mayoría posee ingresos relativamente bajos. Por otro lado, se descarta que el comportamiento de la variable se comporte normal, debido a la asimetría positiva.

De acuerdo con los diagramas de dispersión presentados en las tablas cruzadas, se logra establecer que todas las distribuciones presentan asimetría positiva, de manera que se rechaza parcialmente el hecho de que alguna variable logre presentar un comportamiento normal. Asimismo, algunos de los cruces presentan mayor cohesión entre las observaciones, por lo que la variabilidad de los datos es más baja para aquellas gráficas cuyas observaciones estén más juntas. Finalmente, al haber analizado el diagrama de dispersión de

la variable dependiente e independiente, se establece que se logra gestar una relación positiva, donde mayores niveles de ingreso (PIB), podrían explicar mayor número de medallas ganadas con el paso del tiempo.

Modelo

```
mod1 <- lm(tidy_data$medallas~tidy_data$POB+tidy_data$POB2+tidy_data$GDP+tidy_data$GDP2+tidy_data$DESP+

mod2 <- lm(tidy_data$medallas~tidy_data$POB+tidy_data$GDP+tidy_data$DESP+tidy_data$INF)

mod3 <- lm(tidy_data$medallas~tidy_data$POB+tidy_data$DESP)

resumen <- stargazer(mod1,mod2,mod3, type="text", dep.var.labels = "Numero de medallas olímpicas obteni
```

```
##
## ================================================================================================
##                                                Dependent variable:
##                                 ---------------------------------------------------------------
##                                          Numero de medallas olímpicas obtenidas
##                                     (1)                    (2)                    (3)
## ------------------------------------------------------------------------------------------------
## POB                              -155.205***            -89.416***             -77.074***
##                                   (47.184)               (23.302)               (21.652)
##
## POB2                              11.717
##                                   (7.943)
##
## GDP                               -23.938                -21.890
##                                  (57.153)               (21.132)
##
## GDP2                               1.192
##                                   (9.425)
##
## DESP                              -9.415*                -9.384*                -9.344*
##                                   (5.345)                (5.345)                (5.344)
##
## INF                               -1.160                 -0.648
##                                   (1.210)                (0.485)
##
## INF2                               0.001
##                                   (0.003)
##
## Constant                         491.418***             431.823***             343.572***
##                                   (118.537)              (97.825)               (70.476)
##
## ------------------------------------------------------------------------------------------------
## Observations                        120                    120                    120
## R2                                 0.145                  0.122                  0.103
## Adjusted R2                        0.091                  0.091                  0.087
## Residual Std. Error  310.860 (df = 112)     310.874 (df = 115)     311.563 (df = 117)
## F Statistic         2.709** (df = 7; 112) 3.987*** (df = 4; 115) 6.686*** (df = 2; 117)
## ================================================================================================
## Note:                                                            *p<0.1; **p<0.05; ***p<0.01
```

```r
a <- AIC(mod1)
b <- AIC(mod2)
c <- AIC(mod3)

AIC <- c(a,b,c)
AIC
```

```
## [1] 1727.708 1724.891 1723.491
```

Definimos el tercer modelo como el mejor modelo pues es el que tiene menos perdida de información.

Conclusiones