1000 genomas

Haydeé

2022-09-24

A.Paulina Perez-Gonzalez paulinapglz.99@gmail.com

SCRIPT PARA BioFreelancer

Proyecto 1000 genomas

Este script genera un grafico tipo donut con datos basicos de 1000 Genomas https://www.kaggle.com/datasets/daiearth22/1000-genome-data

```
setwd("~/R_sites/R-ladies/1000Genome_coordpolar")
```

Las librerias

El paquete `pacman` sirve para instalar y cargar paquetes: https://www.rdocumentation.org/packages/pacman/versions/0.5.1

```
library("pacman")

p_load("vroom",
       "dplyr",
       "ggplot2",
       "RColorBrewer")
```

Visualizacion de datos

`vroom` es equivalente a read.csv o similar, detecta automaticamente los header `.df` es solo para identificar nosotras que es un dataframe

```
genomes.df <- vroom(file = "1000genomesinfo.csv")
```

```
## Rows: 3500 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (4): Sample, Population, Center, seq_platform
## dbl (4): AlignedNonDuplicatedCoverage, percentajeTargetsCoveredto20x_or_grea...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Quiero hacer una cuenta de cuantas muestras por pais hay, para ello

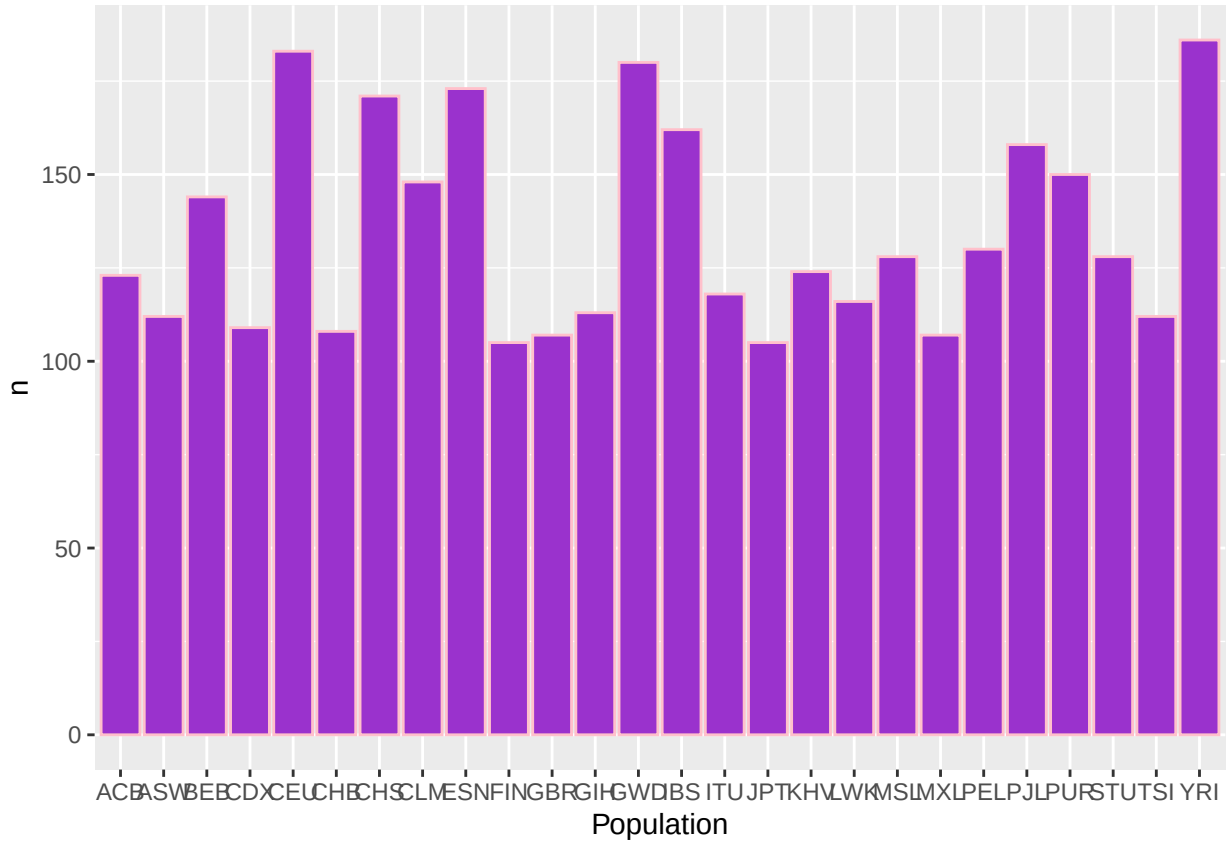
```
counts_per_country <- genomes.df %>%   #Primero llamo a los datos. Uso el símbolo %>% para conectar las
  group_by(Population) %>%   #Luego indico que quiero un grupo por cada valor único en
  tally() #tally cuenta cuantos elementos hay en cada grupo
```

Barplot

Vamos a graficar un barplot sencillo. Los colores predefinidos por R están en: https://r-charts.com/colors/

```
#crear un barplot
genomes.p <- ggplot(counts_per_country,
                    aes(x = Population,
                        y = n)) +
  geom_col(position = "stack",
           fill = "darkorchid3",
           color = "pink")
```

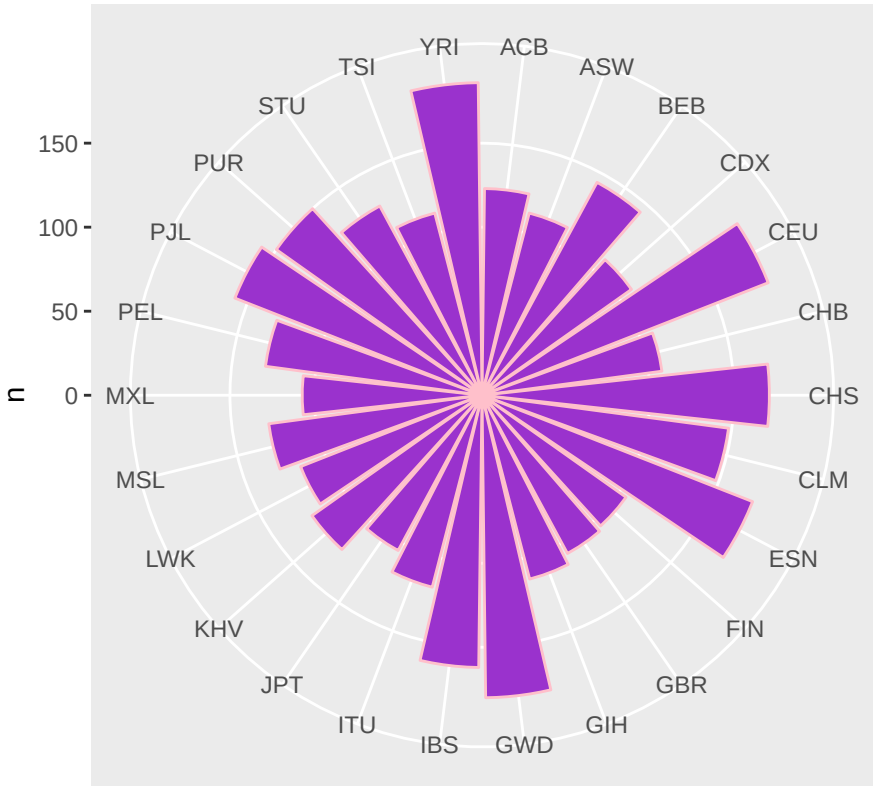
```
#visualizar el plot
genomes.p
```



Vamos a darle formato.

```
genomes_1.p <- genomes.p +
  coord_polar()

genomes_1.p
```

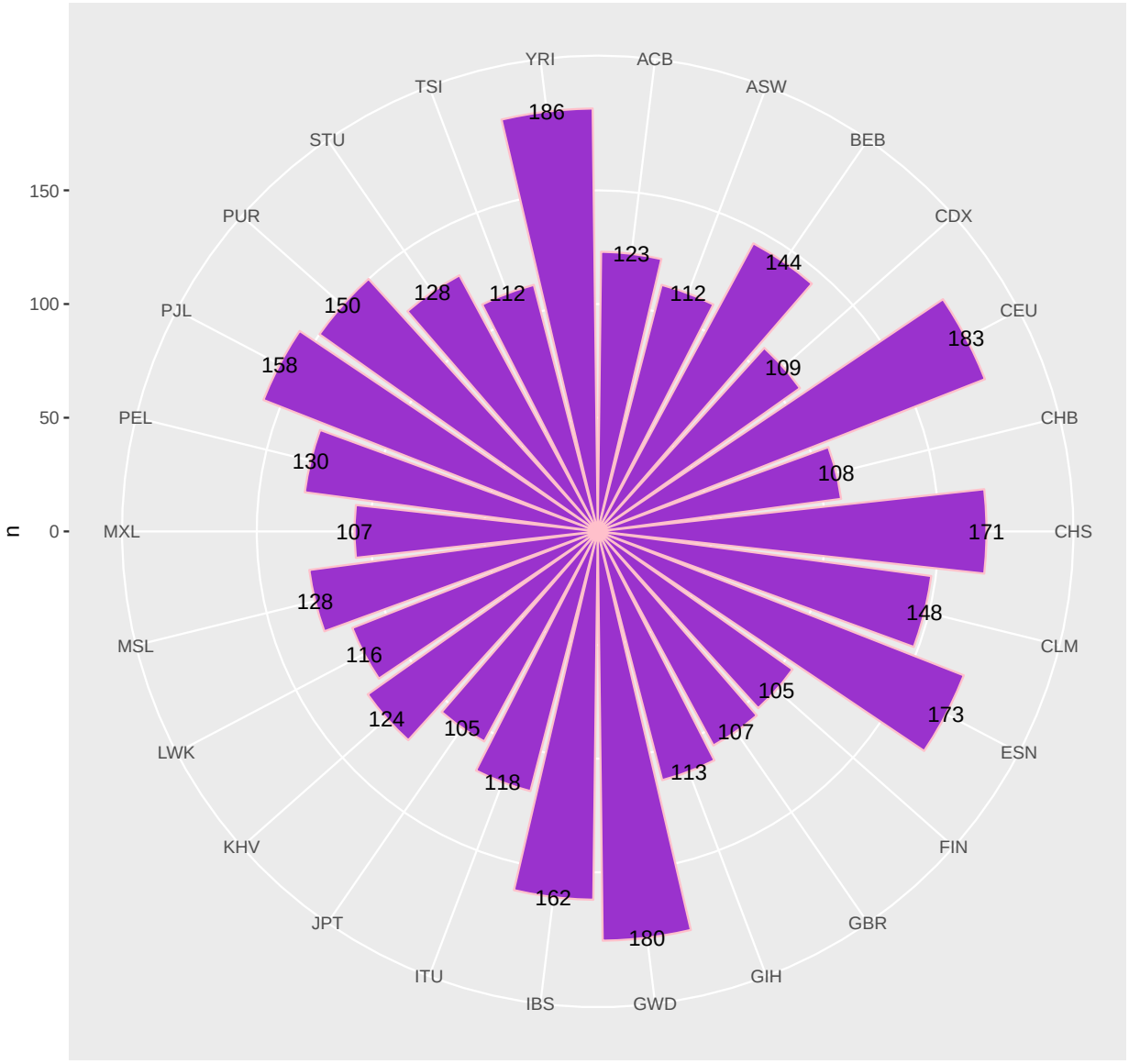


Plot 2

Ponerle una capa de texto y que no se sobre encimen los textos.

```
genomes_2.p <- genomes_1.p +
  geom_text(aes(label=n),
            check_overlap = T)

genomes_2.p
```

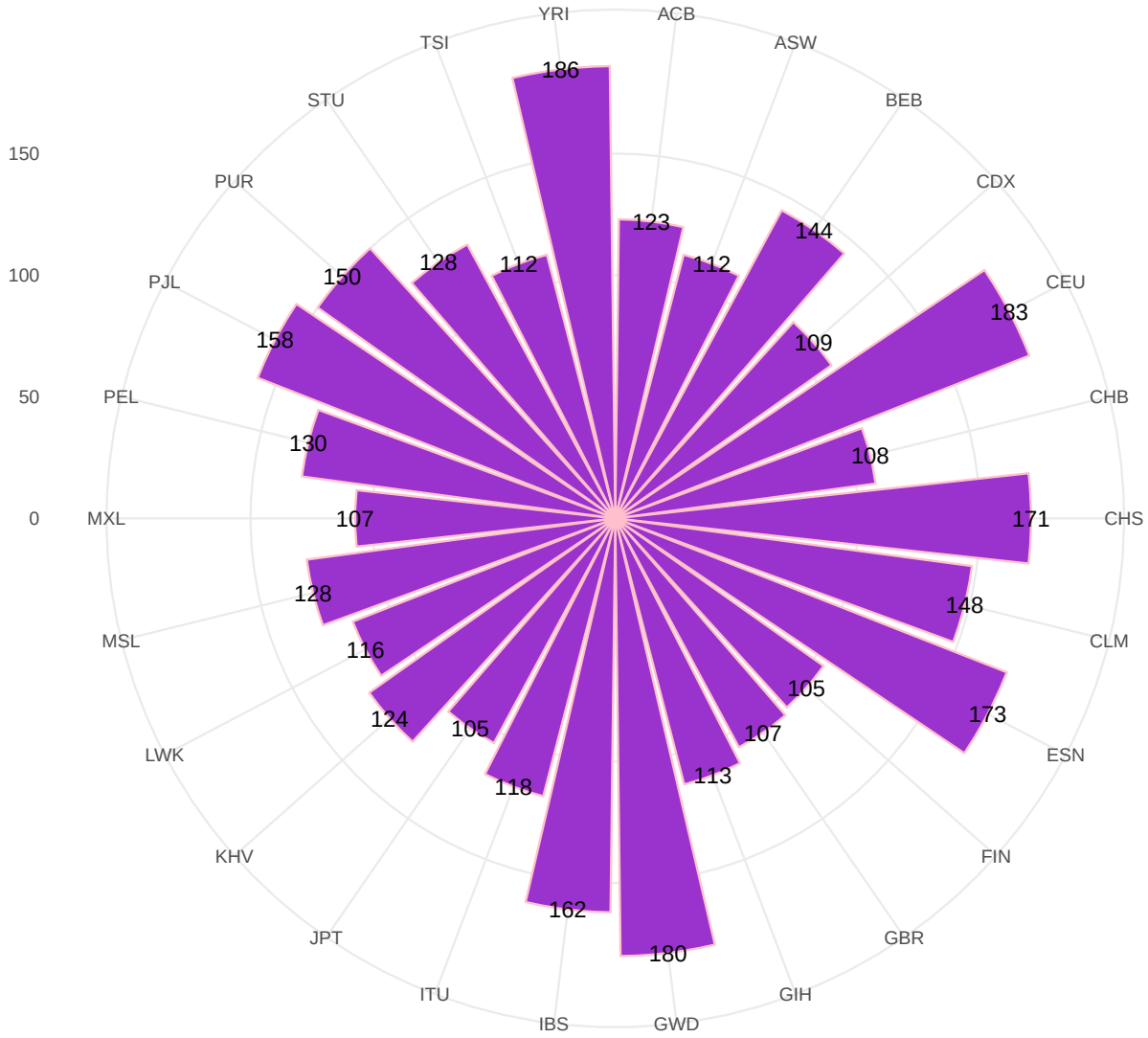


Plot 3

Más formato.

```
genomes_3.p <- genomes_2.p +
  theme_minimal() +
  theme(axis.title.y = element_blank() )

genomes_3.p
```



Plot 4

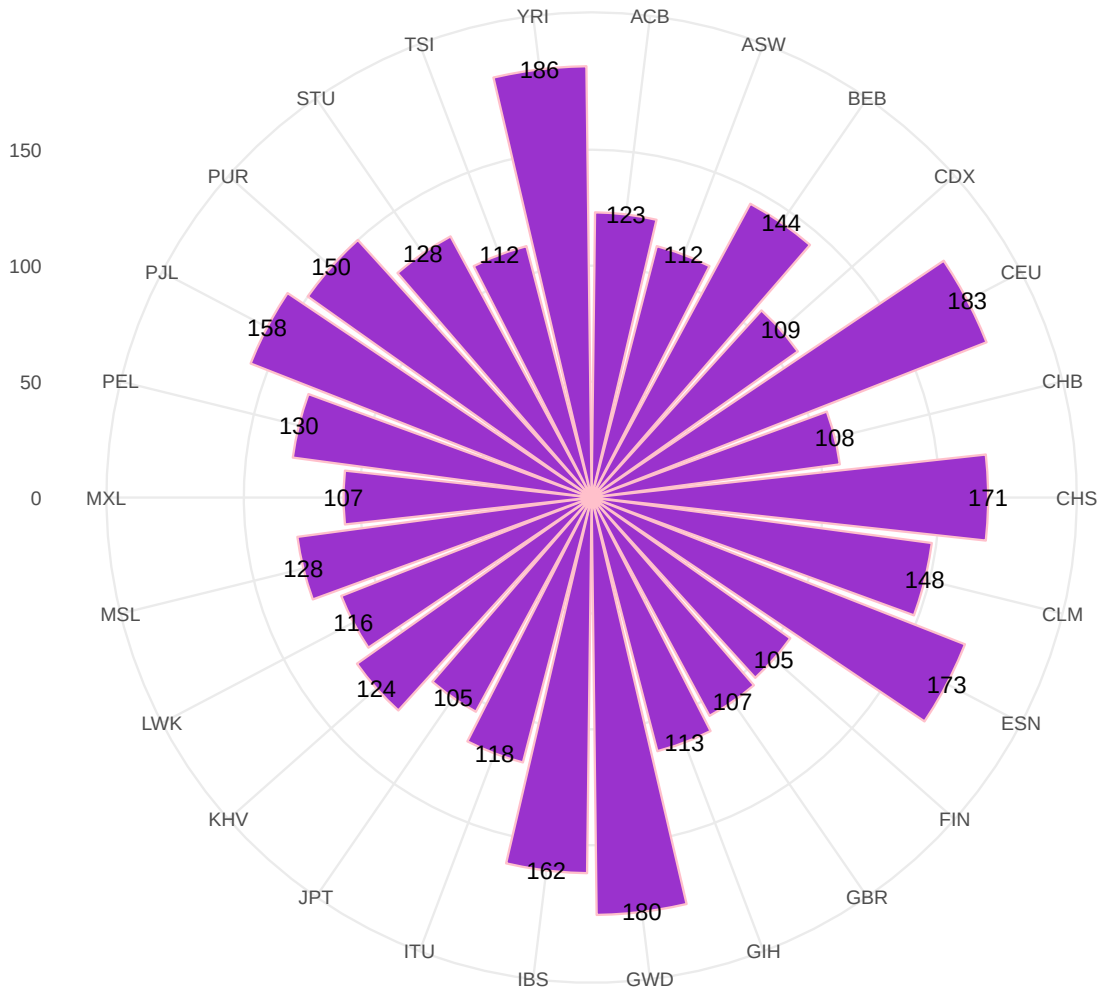
Ponemos etiquetas al título y la fuente de donde se saco.

```
genomes_4.p <- genomes_3.p +
  labs(title="Número de genomas por población",
       subtitle="Según el estudio 1000 Genomes",
       caption="source: kaggle.com") +
  xlab("Poblaciones")

genomes_4.p
```

Número de genomas por población

Según el estudio 1000 Genomes



source: kaggle.com

Guardar el plot

```
# Guardamos el plot
ggsave( filename = "genomes_per_population.png",  # el nombre del archivo de salida
        plot = genomes_4.p,           # guardamos el ultimo grafico que hicimos
        width = 8,                # ancho de 8 pulgadas
        height = 9,               # alto 7 de pulgadas
        dpi = 600 )               # resolucion de 600 puntos por pulgada
```