Practica 2: Tipologia y Ciclo de vida de los datos

Gilberto Jose Martinez

Enero 2019

---

1 Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?.

Este dataset contiene la publicación del ranking de las mejores universidades del mundo, en donde se mide la calidad de la educacion y entrenamiento de los estudiantes, el rango de esta información es del año 2012 al 2015.

Es importante conocer, a la hora de tomar la decisión de estudiar en la Universidad, nosotros o nuestros hijos, cuales son las mas prestigiosas, donde estan ubicadas, cuales son los paises con mayor calidad de educacion universitaria en el mundo segun el CWUR (The Center for World University Rankings), y que universidades en nuestro pais estan incluidas en este ranking.

---

2 Integración y selección de los datos de interés a analizar.

Este dataset ha sido descargado de la URL https://www.kaggle.com/mylesoneill/world-university-rankings/version/3#_=_ (https://www.kaggle.com/mylesoneill/world-university-rankings/version/3#_=_)

Vamos a importar y obtener un resumen general del dataset para conocer con que atributos contamos y como estan compuestas las observaciones.

```
library(psych)
setwd("C:/Users/PS/Desktop/UOC/2do Semestre/Tipologia del ciclo de vida de los datos/Bloque4/Solucion/")
datos <- read.csv("cwurData.csv",sep = ",")
dim(datos)
```

```
## [1] 2200   14
```

```
describe(datos)
```

```
##                      vars    n    mean     sd median trimmed    mad
## world_rank              1 2200  459.59 304.32  450.5  452.80 407.71
## institution*            2 2200  517.30 295.26  522.0  518.43 377.32
## country*                3 2200   36.08  20.27   34.0   37.17  34.10
## national_rank           4 2200   40.28  51.74   21.0   28.37  25.20
## quality_of_education    5 2200  275.10 121.94  355.0  294.02  17.79
## alumni_employment       6 2200  357.12 186.78  450.5  370.98 172.72
## quality_of_faculty      7 2200  178.89  64.05  210.0  192.83  11.86
## publications            8 2200  459.91 303.76  450.5  453.21 406.97
## influence               9 2200  459.80 303.33  450.5  453.32 407.71
## citations              10 2200  413.42 264.37  406.0  412.01 354.34
## broad_impact           11 2000  496.70 286.92  496.0  496.34 363.24
## patents                12 2200  433.35 274.00  426.0  431.58 395.11
## score                  13 2200   47.80   7.76   45.1   45.96   1.19
## year                   14 2200 2014.32   0.76 2014.0 2014.44   1.48
##                         min  max   range  skew kurtosis   se
## world_rank             1.00 1000  999.00  0.11    -1.28 6.49
## institution*           1.00 1024 1023.00 -0.03    -1.20 6.30
## country*               1.00   59   58.00 -0.20    -1.50 0.43
## national_rank          1.00  229  228.00  1.97     3.25 1.10
## quality_of_education   1.00  367  366.00 -1.00    -0.64 2.60
## alumni_employment      1.00  567  566.00 -0.51    -1.24 3.98
## quality_of_faculty     1.00  218  217.00 -1.54     0.82 1.37
## publications           1.00 1000  999.00  0.11    -1.28 6.48
## influence              1.00  991  990.00  0.11    -1.28 6.47
## citations              1.00  812  811.00  0.06    -1.28 5.64
## broad_impact           1.00 1000  999.00  0.01    -1.19 6.42
## patents                1.00  871  870.00  0.00    -1.35 5.84
## score                 43.36  100   56.64  4.18    19.77 0.17
## year                2012.00 2015    3.00 -1.22     1.60 0.02
```

```
names(datos)
```

```
##  [1] "world_rank"          "institution"          "country"
##  [4] "national_rank"       "quality_of_education" "alumni_employment"
##  [7] "quality_of_faculty"  "publications"         "influence"
## [10] "citations"           "broad_impact"         "patents"
## [13] "score"               "year"
```

Como podemos observar el dataset esta compuesto de 2200 observaciones y 14 atributos que son:

- **world_rank:** Ranking mundial de la universidad
- **institution:** Nombre de la universidad
- **country:** País de cada universidad
- **national_rank:** Ranking de la universidad en su propio país
- **quality_of_education:** Ranking por calidad de educación
- **alumni_employment:** Ranking por empleabilidad de sus alumnos
- **quality_of_faculty:** Ranking por calidad de facultad
- **publications:** Ranking por publicaciones
- **influence:** Ranking por influencia
- **citations:** Número de estudiantes de la universidad
- **broad_impact:** Ranking por amplio impacto (disponible solo para los años 2014 y 2015)
- **patents:** Ranking por patentes
- **score:** Puntuación total, utilizada para determinar la clasificación mundial
- **year:** Año del ranking (del 2012 al 2015)

Para nuestro analisis, vamos a seleccionar los atributos: world_rank, institution, country, national_rank, quality_of_education, alumni_employment, score y year, y los vamos a almacenar en un nuevo dataset al que llamaremos **datos2**

```
datos2 <- as.data.frame(c(datos[2:6],datos[1],datos[13:14]))
```

3 Limpieza de los datos.

3.1 ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

Vamos a generar un resumen del nuevo dataset que nos dira si el mismo contiene datos vacios o ceros.

```
summary(datos2)
```

```
##                              institution             country
##  Ã‰cole normale supÃ©rieure - Paris:   4   USA           :573
##  Ã‰cole Polytechnique              :   4   China         :167
##  Arizona State University         :   4   Japan         :159
##  Boston University                :   4   United Kingdom:144
##  Brown University                 :   4   Germany       :115
##  California Institute of Technology:   4   France        :109
##  (Other)                          :2176   (Other)       :933
##  national_rank    quality_of_education alumni_employment   world_rank
##  Min.   :  1.00   Min.   :  1.0        Min.   :  1.0     Min.   :   1.0
##  1st Qu.:  6.00   1st Qu.:175.8        1st Qu.:175.8     1st Qu.: 175.8
##  Median : 21.00   Median :355.0        Median :450.5     Median : 450.5
##  Mean   : 40.28   Mean   :275.1        Mean   :357.1     Mean   : 459.6
##  3rd Qu.: 49.00   3rd Qu.:367.0        3rd Qu.:478.0     3rd Qu.: 725.2
##  Max.   :229.00   Max.   :367.0        Max.   :567.0     Max.   :1000.0
##
##      score             year
##  Min.   : 43.36   Min.   :2012
##  1st Qu.: 44.46   1st Qu.:2014
##  Median : 45.10   Median :2014
##  Mean   : 47.80   Mean   :2014
##  3rd Qu.: 47.55   3rd Qu.:2015
##  Max.   :100.00   Max.   :2015
##
```

Podemos observar que los atributos seleccionados no contienen datos en cero, ya que aquellos de tipo numerico contienen como valor minimo uno (1), y no existen datos vacios, de ser así se nos mostraria como (NA's)

3.2 Identificación y tratamiento de valores extremos.

La mejor manera de poder visualizar si existen datos extremos en nuestro dataset es a traves de un diagrama de caja, crearemos uno a partir de los atributos numericos de nuestro dataset, agrupando aquellos cuyo valor se asemejen, para identificar si se presenta esta situación.

Los primeros graficos que crearemos son basados en el **ranking mundial**

```
library(dplyr)
```

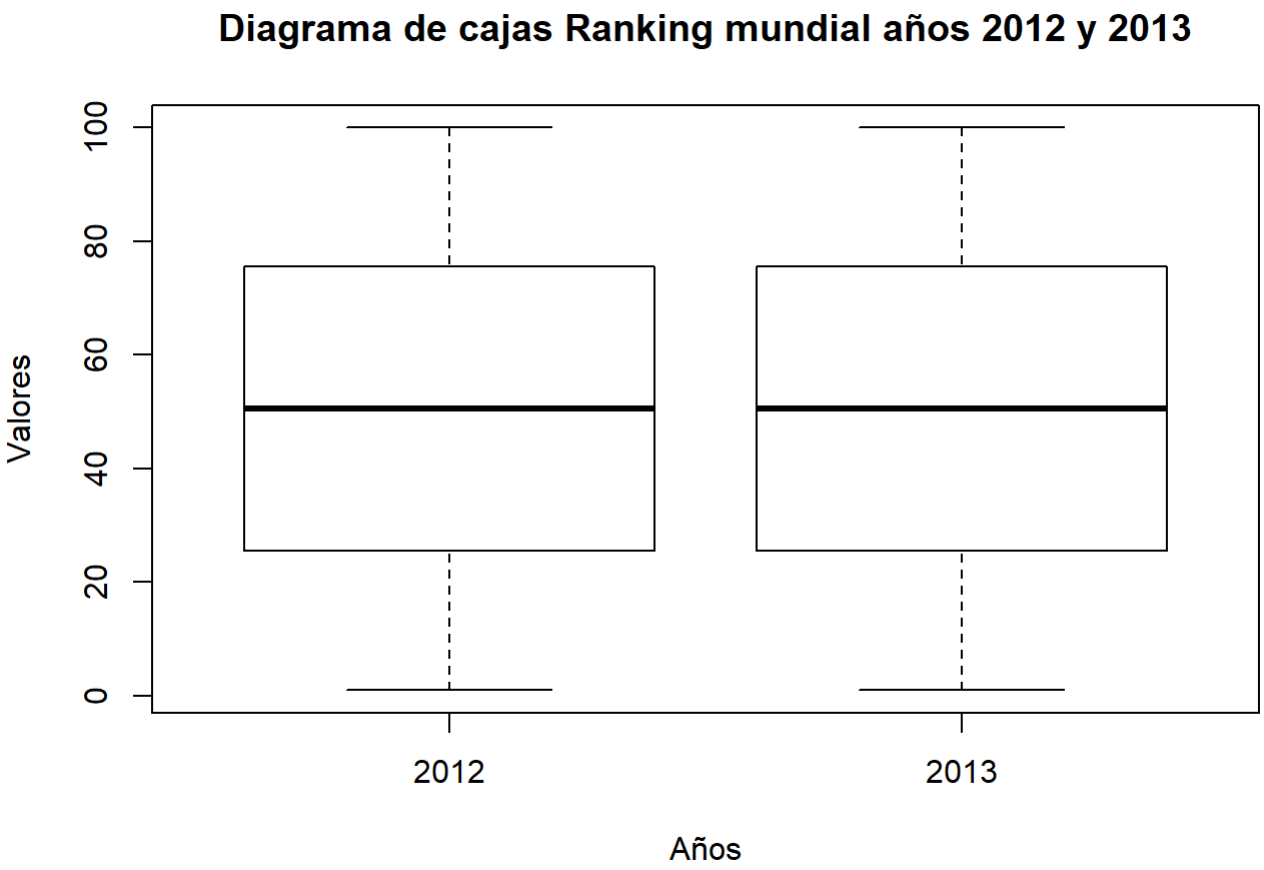
```
##
## Attaching package: 'dplyr'
```

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

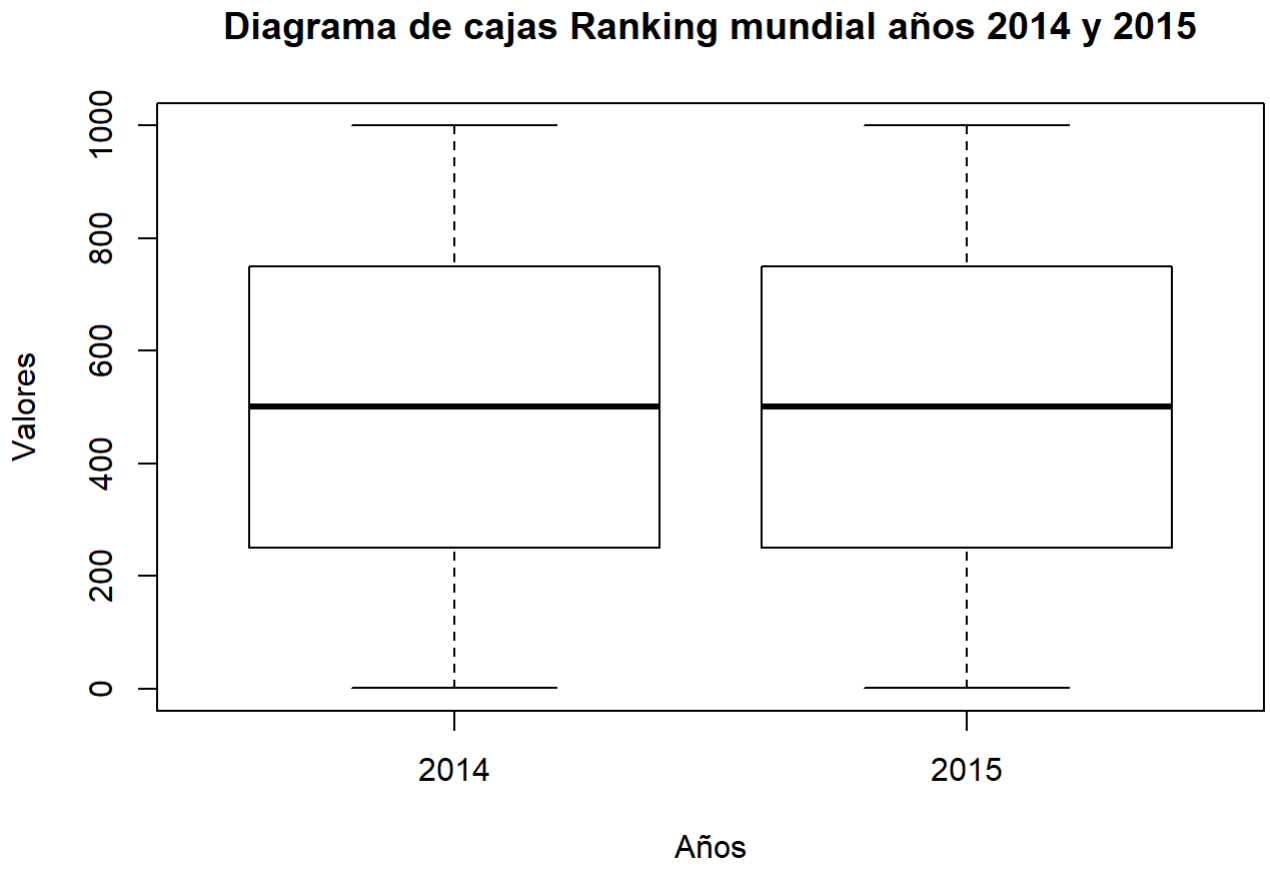
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
rk_2012 <- filter(datos2,year==2012)
rk_2013 <- filter(datos2,year==2013)
rk_2014 <- filter(datos2,year==2014)
rk_2015 <- filter(datos2,year==2015)

# Diagrama de cajas años 2012 y 2013
rk_world_2012_2013 <- as.data.frame(c(rk_2012[6],rk_2013[6]))
names(rk_world_2012_2013) = c("2012","2013")
boxplot(rk_world_2012_2013,xlab="Años",ylab="Valores",main="Diagrama de cajas Ranking mun
dial años 2012 y 2013")
```

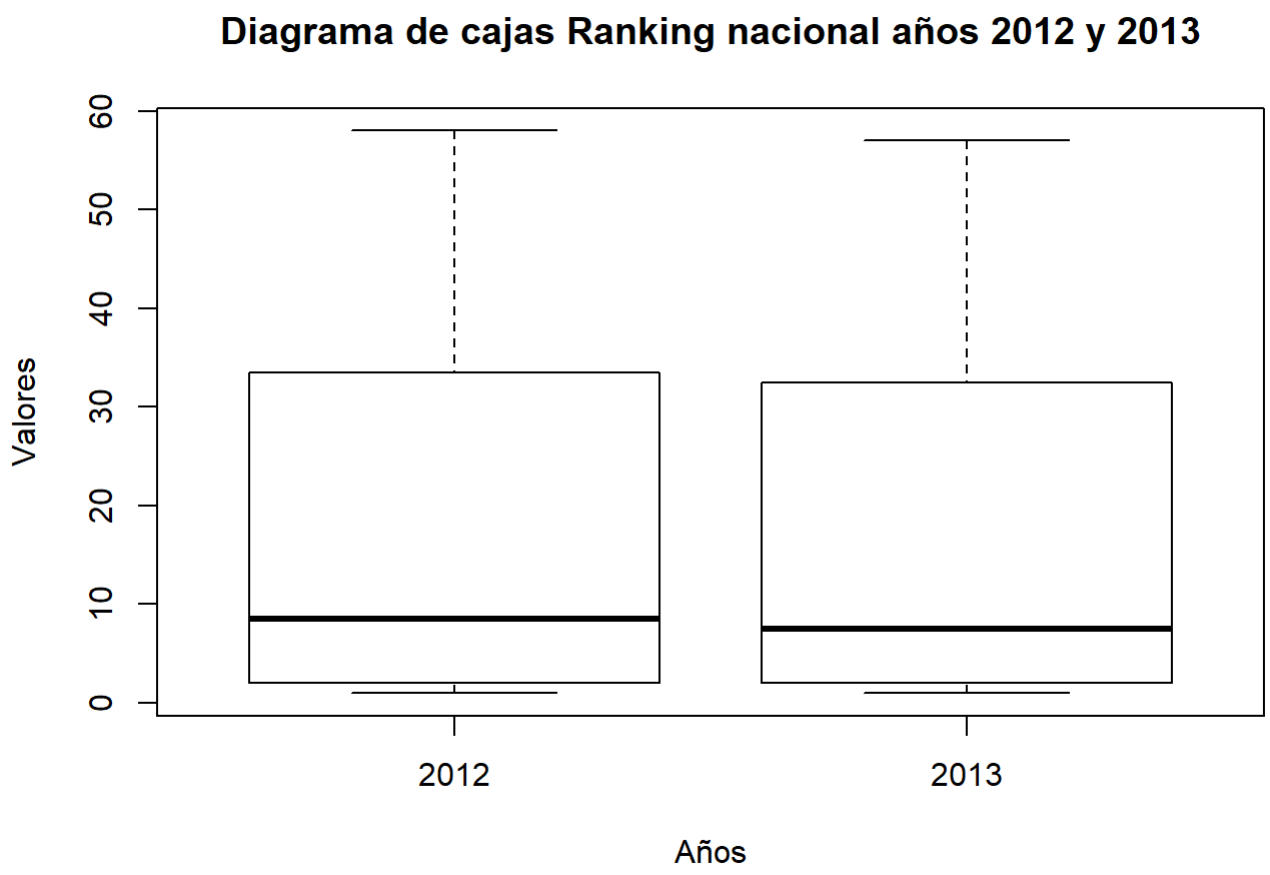


```
# Diagrama de cajas años 2014 y 2015
rk_world_2014_2015 <- as.data.frame(c(rk_2014[6],rk_2015[6]))
names(rk_world_2014_2015) = c("2014","2015")
boxplot(rk_world_2014_2015,xlab="Años",ylab="Valores",main="Diagrama de cajas Ranking mun
dial años 2014 y 2015")
```

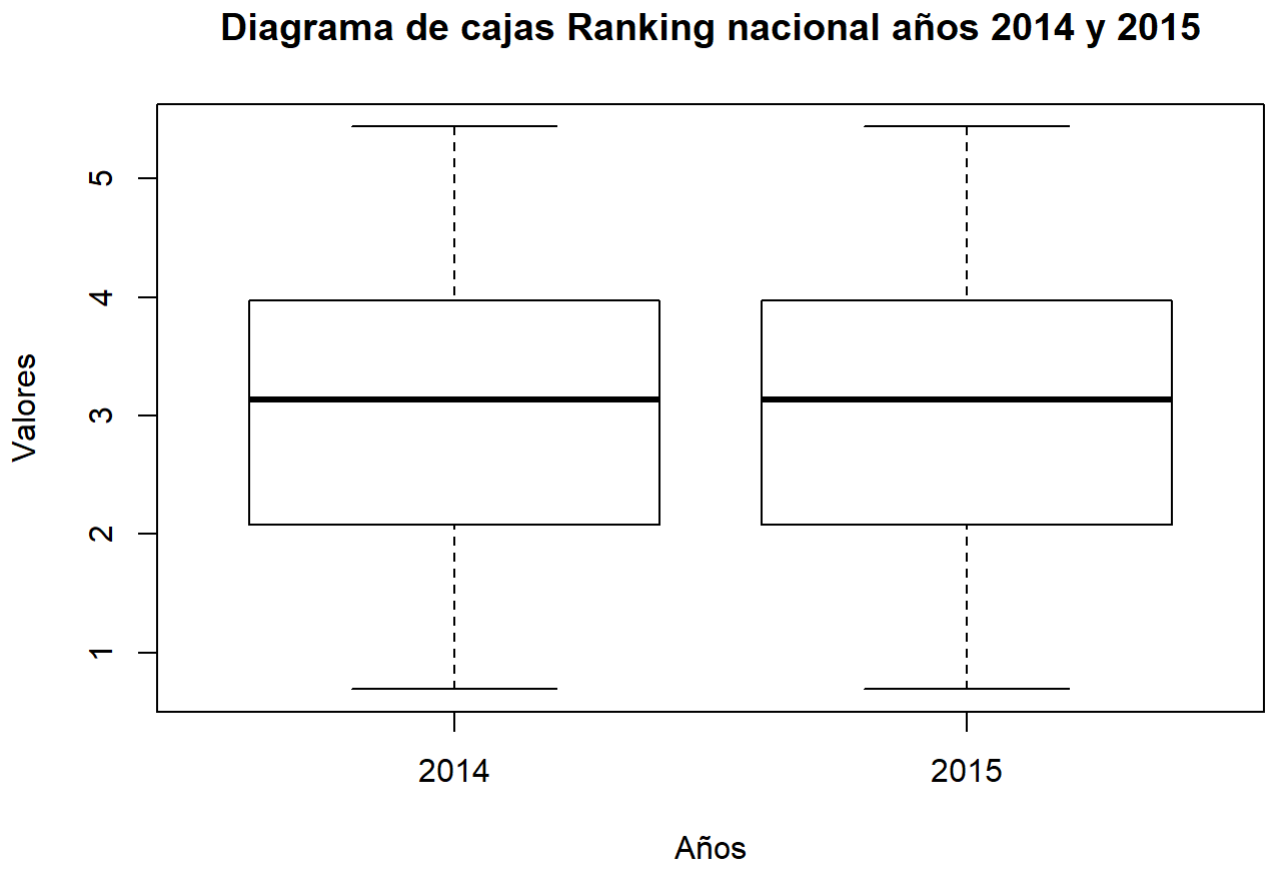


A continuacion vamos a generar los diagramas de caja para el ranking nacional, hemos separado los años 2012 y 2013 de 2014 y 2015, ya que los ultimos dos años se incluyeron mas universidades, con lo cual, al unirlos en un solo grafico nos produciria datos extremos para los años 2014 y 2015.

```
# Años 2012 y 2013
rk_national_all_2012_2013 <- as.data.frame(c(rk_2012[3],rk_2013[3]))
names(rk_national_all_2012_2013) = c("2012","2013")
boxplot(rk_national_all_2012_2013,xlab="Años",ylab="Valores",main="Diagrama de cajas Rank
ing nacional años 2012 y 2013")
```




```
# Años 2014 y 2015
rk_national_all_2014_2015 <- as.data.frame(c(log(rk_2014[3]+1),log(rk_2015[3]+1)))
names(rk_national_all_2014_2015) = c("2014","2015")
boxplot(rk_national_all_2014_2015,xlab="Años",ylab="Valores",main="Diagrama de cajas Rank
ing nacional años 2014 y 2015")
```



Los diagramas evidencian que **no existen datos extremos** en los atributos que hemos seleccionado.

---

4 Análisis de los datos.

4.1 Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

Usaremos los cuatro dataframe que hemos creado en el punto 3.2 del presente proyecto en los cuales agrupamos por año, desde el 2012 hasta el 2015 (rk_2012, rk_2013,rk_2014 y rk_2015), para poder hallar la siguiente información:

- Vamos a averiguar segun los datos, cual es el top 10 de los paises con mayor calidad de educación universitaria segun el numero de universidades incluidas en el ranking
- Compararemos si la tendencia a traves de los años se mantiene o sufrio algún cambio

---

4.2 Comprobación de la normalidad y homogeneidad de la varianza.

Gracias a las bondades de el lenguaje R, podemos comprobar la normalidad y homogenidad de la varianza en nuestros datos de una forma muy sencilla, para ello emplearemos la función **bartlett.test()** para hallar la homogenidad de la varianza, y la función **shapiro.test()** para hallar la normalidad, la aplicaremos sobre el dataframe **datos2** que contiene todos los datos, lo calcularemos con base al año.

```
# Test de homogenidad de la varianza
bartlett.test(score~year,data=datos2)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  score by year
## Bartlett's K-squared = 245.3, df = 3, p-value < 2.2e-16
```

```
# Test de Normalidad
shapiro.test(datos2$year)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  datos2$year
## W = 0.74004, p-value < 2.2e-16
```

Al efectuar estas dos pruebas observamos que p-valor se encuentra muy por debajo se 5% que es la medida convencional, por lo tanto se rechaza la hipotesis nula

4.3 Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc.

Vamos a graficar los paises con mas numero de universidades incluidas entre las mejores del mundo por cada año segun su ranking, agrupado por los 10 primeros paises. Iniciaremos con los años 2012 y 2013

```
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'
```

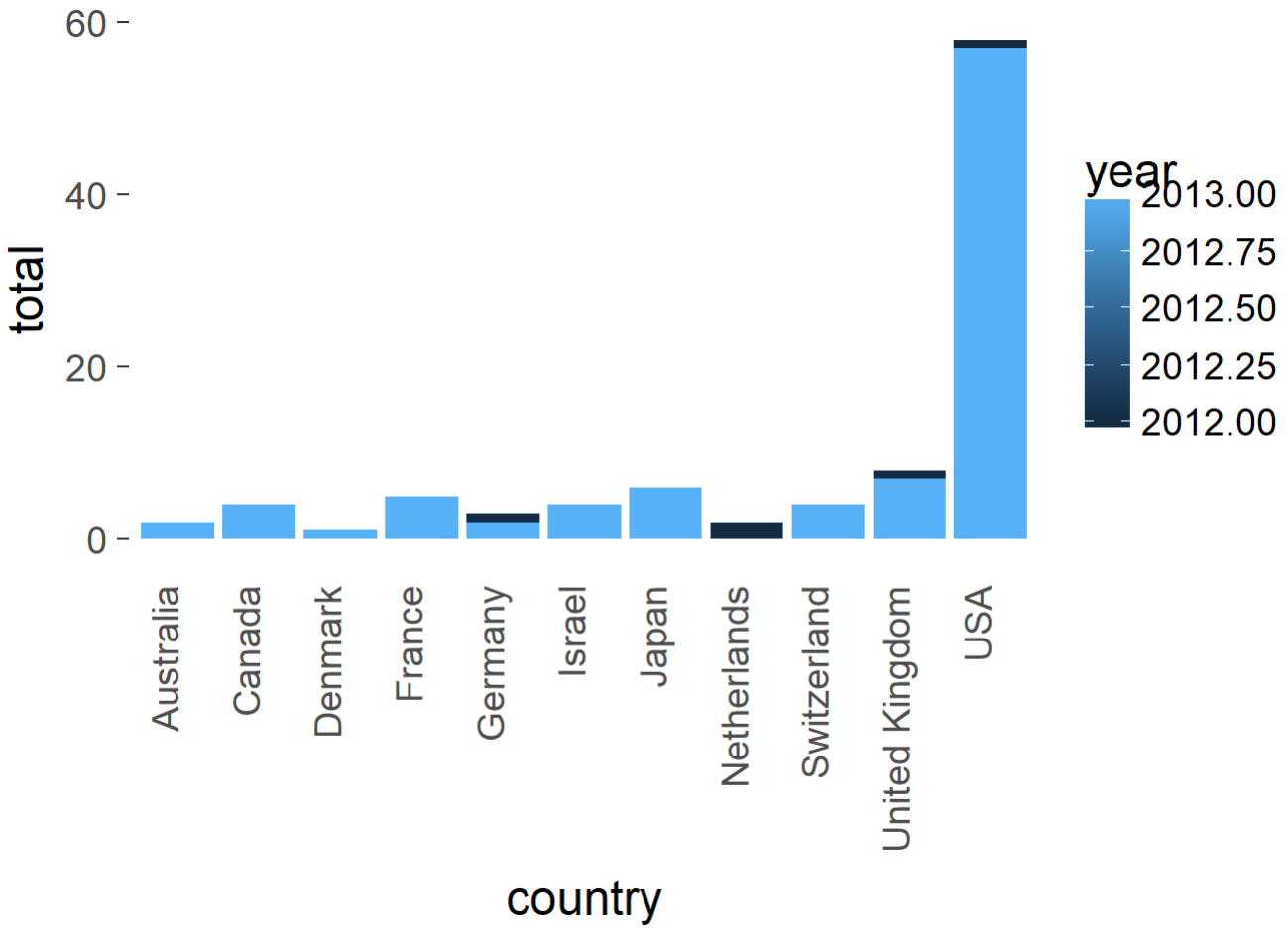
```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```
# Año 2012
gr_rk_2012 <- group_by(rk_2012,country,year)
gr_rk_2012 <- arrange(summarise(gr_rk_2012,total = n()),desc(total))
gr_rk_2012 <- gr_rk_2012[1:10, ]
# Año 2013
gr_rk_2013 <- group_by(rk_2013,country,year)
gr_rk_2013 <- arrange(summarise(gr_rk_2013,total = n()),desc(total))
gr_rk_2013 <- gr_rk_2013[1:10, ]
# Año 2014
gr_rk_2014 <- group_by(rk_2014,country,year)
gr_rk_2014 <- arrange(summarise(gr_rk_2014,total = n()),desc(total))
gr_rk_2014 <- gr_rk_2014[1:10, ]

# Año 2015
gr_rk_2015 <- group_by(rk_2015,country,year)
gr_rk_2015 <- arrange(summarise(gr_rk_2015,total = n()),desc(total))
gr_rk_2015 <- gr_rk_2015[1:10, ]

# agrupando 2012 y 2013
gr_2012_2013 <- rbind(gr_rk_2012,gr_rk_2013)

ggplot(gr_2012_2013) +
  geom_bar(aes(x = country, y = total, fill = year),
           stat = "identity", position = "dodge") +
  theme_classic(base_size = 18) +
  theme(axis.text.x = element_text(angle = 90,
                                   hjust = 1, vjust = 0),
        axis.line = element_blank(),
        axis.ticks.x = element_blank())
```




```
tmp_2012 <- select(gr_rk_2012,country,anio2012 = total)
tmp_2013 <- select(gr_rk_2013,country,anio2013 = total)

mg_2012_2013 <- merge(tmp_2012,tmp_2013)
mg_2012_2013 <- arrange(mg_2012_2013,desc(anio2012))

mg_2012_2013
```

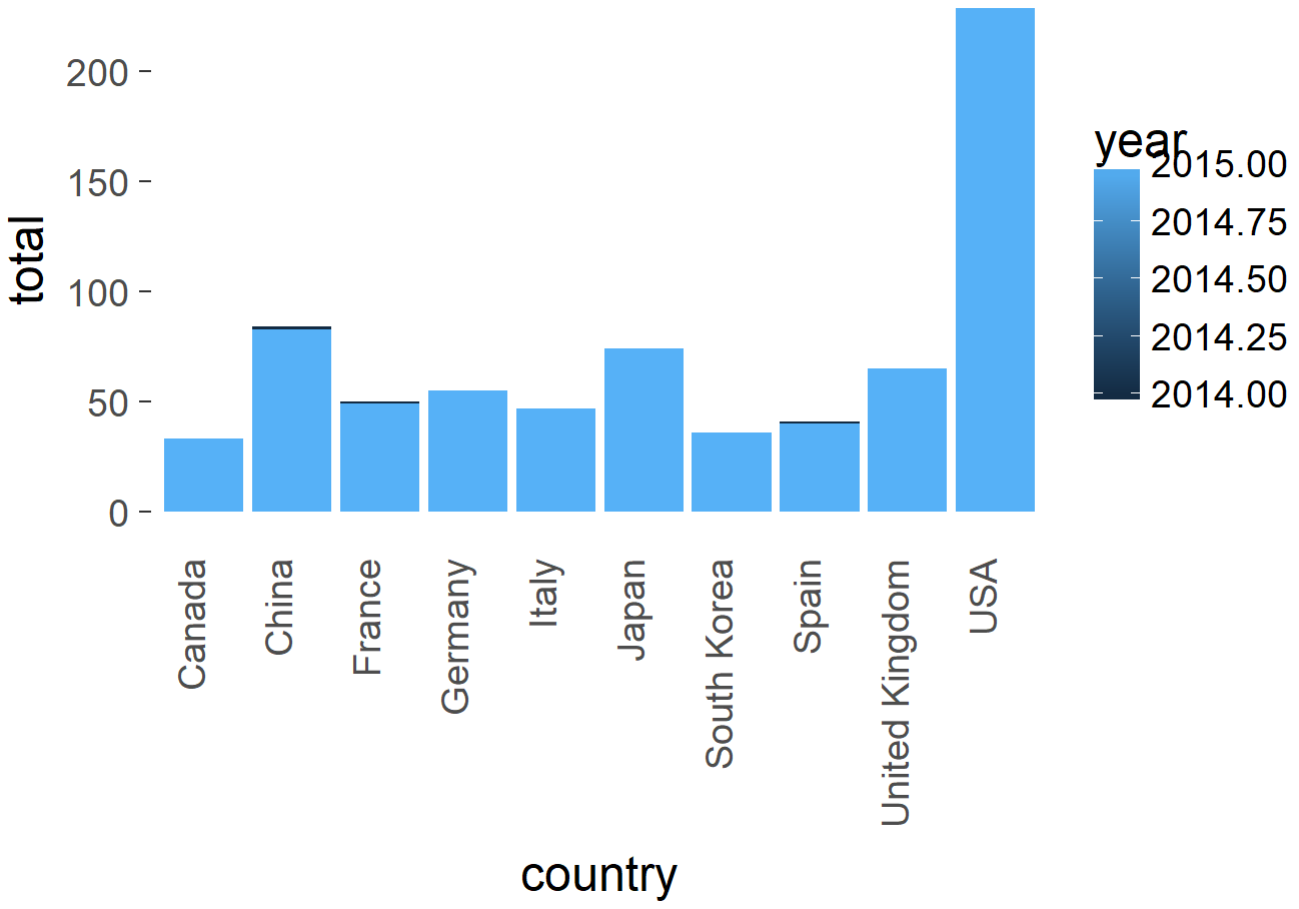
```
##          country anio2012 anio2013
## 1            USA       58       57
## 2 United Kingdom        8        7
## 3         France        5        5
## 4          Japan        5        6
## 5         Israel        4        4
## 6    Switzerland        4        4
## 7         Canada        3        4
## 8        Germany        3        2
## 9      Australia        2        2
```

Podemos observar que USA se mantiene en el primer lugar, con mas universidades en el ranking de las mejores del mundo, podemos comparar en la tabla anterior los años 2012 y 2013

Seguidamente graficaremos los paises con mas numero de universidades, en el estudio realizado en a los años 2014 y 2015

```
# agrupando 2014 y 2015
gr_2014_2015 <- rbind(gr_rk_2014,gr_rk_2015)

ggplot(gr_2014_2015) +
  geom_bar(aes(x = country, y = total, fill = year),
           stat = "identity", position = "dodge") +
  theme_classic(base_size = 18) +
  theme(axis.text.x = element_text(angle = 90,
                                   hjust = 1, vjust = 0),
        axis.line = element_blank(),
        axis.ticks.x = element_blank())
```




```
tmp_2014 <- select(gr_rk_2014,country,anio2014 = total)

tmp_2015 <- select(gr_rk_2015,country,anio2015 = total)

mg_2014_2015 <- merge(tmp_2014,tmp_2015)
mg_2014_2015 <- arrange(mg_2014_2015,desc(anio2014))

mg_2014_2015
```

```
##           country anio2014 anio2015
## 1             USA      229      229
## 2           China       84       83
## 3           Japan       74       74
## 4  United Kingdom       64       65
## 5         Germany       55       55
## 6          France       50       49
## 7           Italy       47       47
## 8           Spain       41       40
## 9     South Korea       34       36
## 10         Canada       32       33
```

En los estudios de los años 2014 y 2015 se incluyeron 1000 universidades mas, por ello los valores se amplian considerablemente, esto lo podemos observar en la tabla anterior.

```
# Las 15 mejores universidades de cada pais Año 2012
gr_rk_2012 <- filter(rk_2012,national_rank==1)
gr_rk_2012 <- select(gr_rk_2012,institution,country,score)
gr_rk_2012 <- gr_rk_2012[1:15, ]
gr_rk_2012
```

```
##                                       institution        country  score
## 1                             Harvard University            USA 100.00
## 2                          University of Cambridge United Kingdom  86.17
## 3                              University of Tokyo          Japan  69.49
## 4  Swiss Federal Institute of Technology in Zurich    Switzerland  66.69
## 5                    Weizmann Institute of Science         Israel  65.09
## 6                            University of Toronto         Canada  53.43
## 7                          University of Paris-Sud         France  50.44
## 8                          University of Edinburgh United Kingdom  48.43
## 9                             Karolinska Institute         Sweden  47.61
## 10                       Seoul National University    South Korea  46.74
## 11                     Sapienza University of Rome          Italy  46.34
## 12          Ruprecht Karl University of Heidelberg        Germany  45.33
## 13                               Leiden University    Netherlands  45.13
## 14                          University of Helsinki        Finland  44.44
## 15                              University of Oslo         Norway  44.26
```

```
# Las 15 mejores universidades de cada pais Año 2013
gr_rk_2013 <- filter(rk_2013,national_rank==1)
gr_rk_2013 <- select(gr_rk_2013,institution,country,score)
gr_rk_2013 <- gr_rk_2013[1:15, ]
gr_rk_2013
```

```
##                                       institution        country  score
## 1                              Harvard University            USA 100.00
## 2                            University of Oxford United Kingdom  92.54
## 3                             University of Tokyo          Japan  76.23
## 4  Swiss Federal Institute of Technology in Zurich    Switzerland  64.99
## 5                    Hebrew University of Jerusalem         Israel  59.98
## 6                              University of Toronto         Canada  56.11
## 7                            University of Paris-Sud         France  51.72
## 8                         Seoul National University    South Korea  51.31
## 9                              Karolinska Institute         Sweden  47.98
## 10                      Sapienza University of Rome          Italy  47.75
## 11          Ludwig Maximilian University of Munich        Germany  47.25
## 12                           University of Copenhagen        Denmark  47.12
## 13                                 University of Oslo         Norway  46.10
## 14                                 Utrecht University    Netherlands  45.73
## 15                 National University of Singapore      Singapore  45.20
```

```
# Las 15 mejores universidades de cada pais Año 2014
gr_rk_2014 <- filter(rk_2014,national_rank==1)
gr_rk_2014 <- select(gr_rk_2014,institution,country,score)
gr_rk_2014 <- gr_rk_2014[1:15, ]
gr_rk_2014
```

```
##                                       institution        country  score
## 1                              Harvard University            USA 100.00
## 2                          University of Cambridge United Kingdom  97.64
## 3                              University of Tokyo          Japan  80.64
## 4  Swiss Federal Institute of Technology in Zurich    Switzerland  72.18
## 5                    Hebrew University of Jerusalem         Israel  66.76
## 6                         Seoul National University    South Korea  66.06
## 7                              University of Toronto         Canada  60.87
## 8                 Ã‰cole normale supÃ©rieure - Paris         France  59.72
## 9                   Lomonosov Moscow State University         Russia  56.42
## 10                               Peking University          China  55.30
## 11                        National Taiwan University         Taiwan  54.19
## 12                             Karolinska Institute         Sweden  53.64
## 13                  National University of Singapore      Singapore  53.52
## 14                           University of Copenhagen        Denmark  52.94
## 15          Ludwig Maximilian University of Munich        Germany  52.75
```

```
# Las 15 mejores universidades de cada pais Año 2015
gr_rk_2015 <- filter(rk_2015,national_rank==1)
gr_rk_2015 <- select(gr_rk_2015,institution,country,score)
gr_rk_2015 <- gr_rk_2015[1:15, ]
gr_rk_2015
```

```
##                                       institution        country  score
## 1                             Harvard University            USA 100.00
## 2                         University of Cambridge United Kingdom  96.81
## 3                             University of Tokyo          Japan  78.23
## 4  Swiss Federal Institute of Technology in Zurich    Switzerland  66.93
## 5                    Hebrew University of Jerusalem         Israel  65.71
## 6                         Seoul National University    South Korea  64.82
## 7                             University of Toronto         Canada  60.04
## 8                              Ã‰cole Polytechnique         France  59.20
## 9                                 Peking University          China  54.26
## 10                       National Taiwan University         Taiwan  54.23
## 11               Lomonosov Moscow State University         Russia  54.19
## 12                National University of Singapore      Singapore  53.44
## 13                             Karolinska Institute         Sweden  52.79
## 14                          University of Copenhagen        Denmark  52.51
## 15          Ruprecht Karl University of Heidelberg        Germany  52.32
```