2022 - 2023

ACTIVIDAD GUIADA 1

Máster en Big Data y Data Science

01MBID – Fundamentos de la Tecnología de Big Data

Nombre: **Angelo Ponce Figueroa, Leonidas J. del Rio**

Fecha: **27 de septiembre del 2022**

Curso 2022 – Ed. Abril

viu Universidad Internacional de Valencia

Contenido

1. Clustering

1.1.Descripción de los datos

https://www.kaggle.com/datasets/santhraul/country-data

- **Country:** País
- **Child Mortality Rate:** Tasa de Mortalidad Infantil
- **Exports:** Exportaciones
- **Health:** Salud
- **Imports:** Importaciones
- **Per capita Income:** El ingreso per cápita
- **Inflation:** Inflación
- **Fertility Rate:** Tasa de fertilidad
- **Life Expectancy:** Esperanza de vida
- **The GDP per capita:** El PIB per cápita

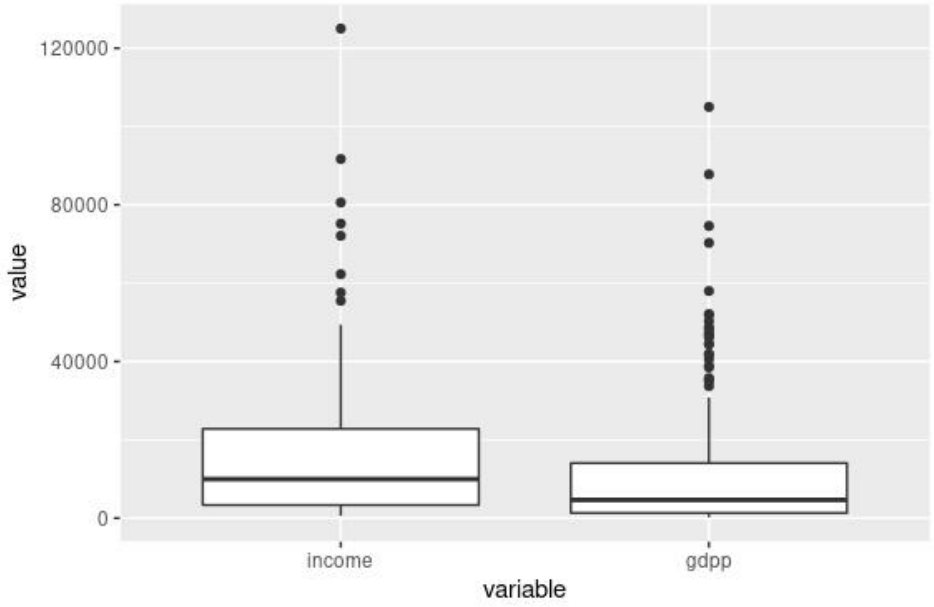
1.2.Motivación del análisis estadístico

Objetivos:

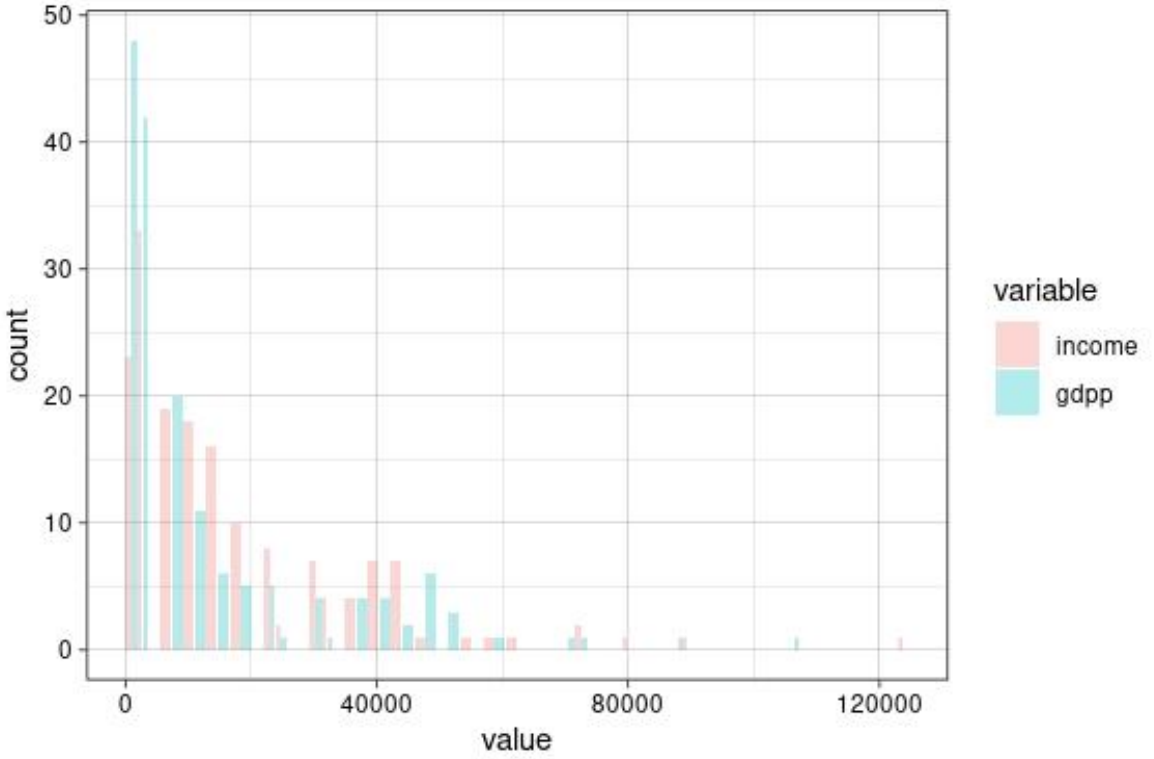
- Categorizar los países utilizando algunos factores socioeconómicos y de salud que determinan el desarrollo general del país.
- Identificar los países que más necesitan ayuda. Globalmente y en Latinoamérica

1.3.Estadística descriptiva

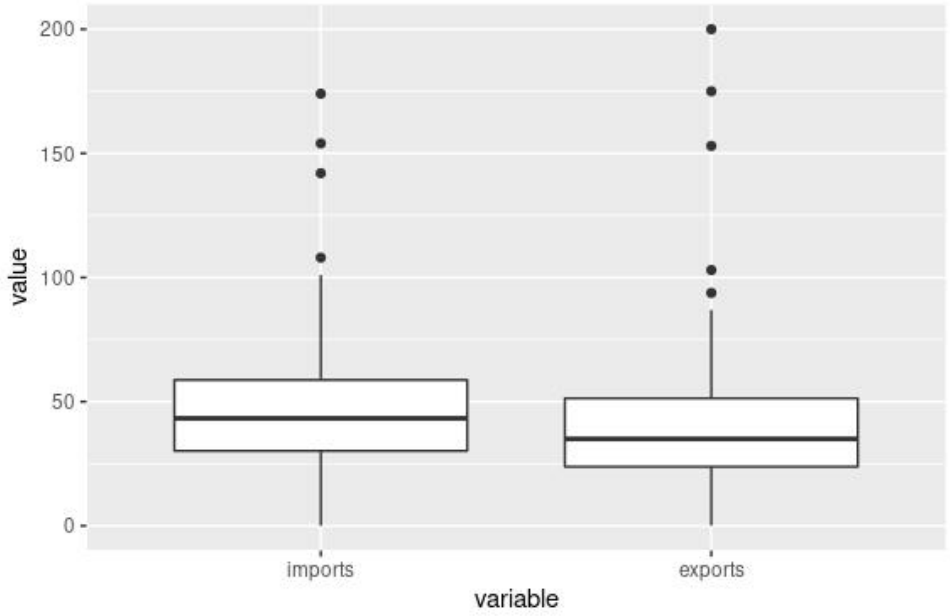
skim_variable <chr>	mean <dbl>	sd <dbl>	p0 <dbl>	p25 <dbl>	p50 <dbl>	p75 <dbl>	p100 <dbl>	hist <chr>
1 child_mort	38.270060	40.328931	2.6000	8.250	19.30	62.10	2.08e+02	
2 health	6.815689	2.746837	1.8100	4.920	6.32	8.60	1.79e+01	
3 life_expec	70.555689	8.893172	32.1000	65.300	73.10	76.80	8.28e+01	
4 exports	41.108976	27.412010	0.1090	23.800	35.00	51.35	2.00e+02	
5 imports	46.890215	24.209589	0.0659	30.200	43.30	58.75	1.74e+02	
6 income	17144.688623	19278.067698	609.0000	3355.000	9960.00	22800.00	1.25e+05	
7 inflation	7.781832	10.570704	-4.2100	1.810	5.39	10.75	1.04e+02	
8 total_fer	2.947964	1.513848	1.1500	1.795	2.41	3.88	7.49e+00	
9 gdpp	12964.155689	18328.704809	231.0000	1330.000	4660.00	14050.00	1.05e+05	



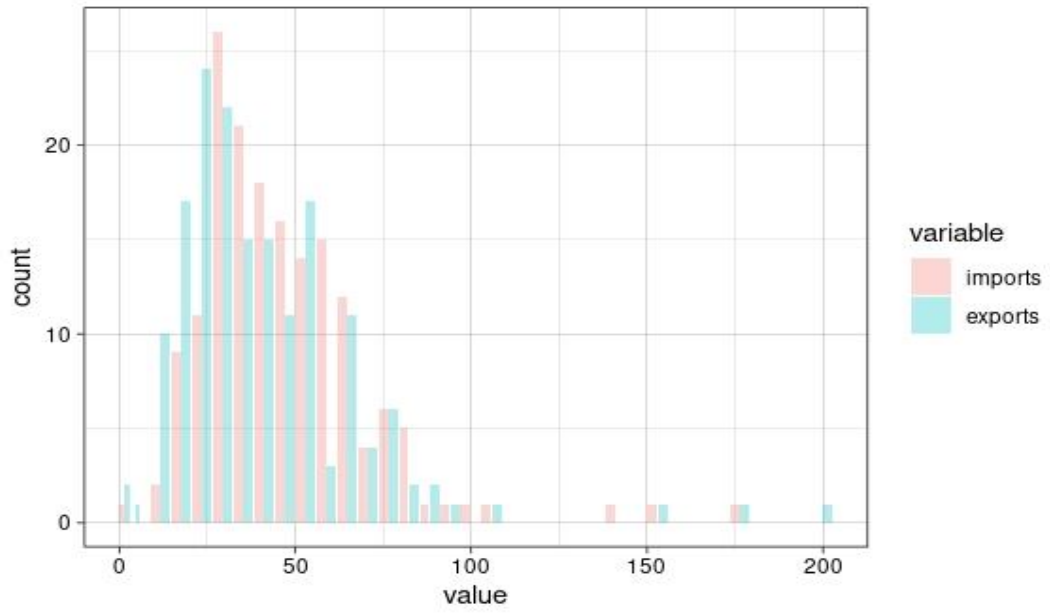
Con respecto a los ingresos y el gdpp notamos una media cercana estas dos variables están relacionadas con el desarrollo. Es de notar los *Outliers* que la verdad no son muchos



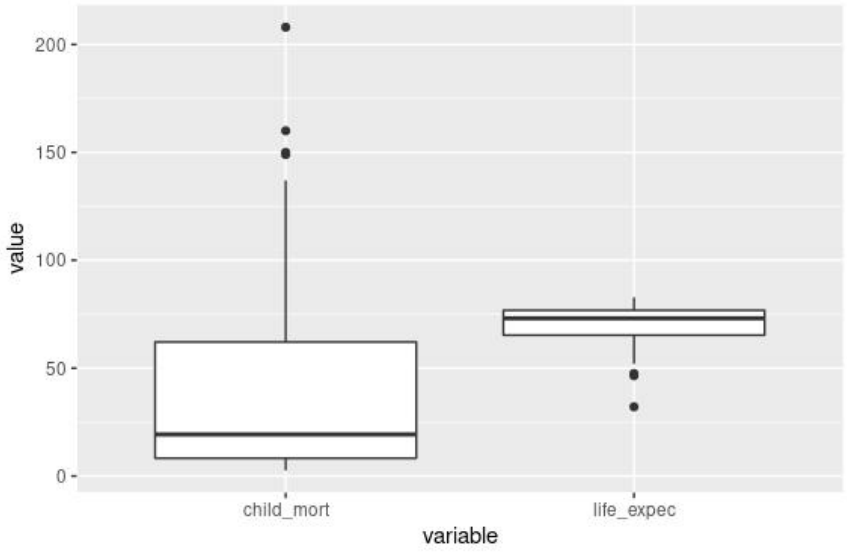
En el histograma de arriba vemos nuevamente que las variables income gdpp están bastante relacionadas un menor gdpp e income se ve en la mayoría de los países a medida que el valor de ambos aumenta los países disminuyen considerablemente y vemos los outliers de baja frecuencia



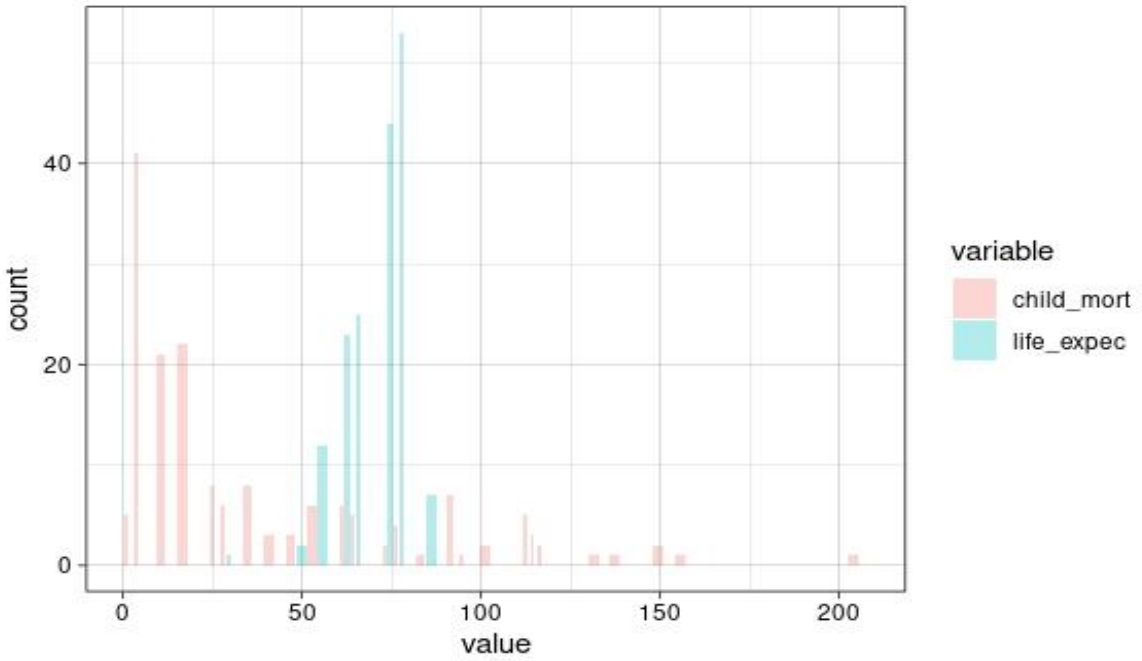
Muy similar a las variables anteriores en cuanto a su relación nuevamente



En la gráfica se nota algo diferente los países menos desarrollados se nota una tendencia de aumento, pero se llega a un límite a partir de eso momento la frecuencia de países disminuye, pero después se nota una nueva tendencia de aumento a partir de 50 notable o un mayor de importaciones o exportaciones por país(es)

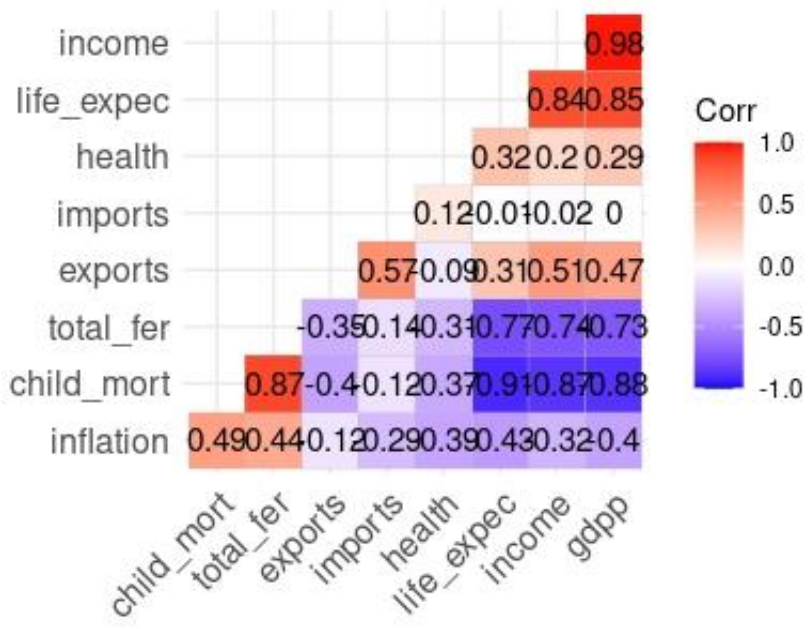


En el boxplot de mortalidad en niños y expectativa de vida vemos un gran porcentaje de datos por encima de la media en la mortalidad de niños esto nos dice que un gran número de países sufren están afectados se notan ciertos *outliers* de números bastante elevados y poca frecuencia. La expectativa de vida es de 50-80 vemos *outliers* también esta vez de muerte temprana en algunos países



En el histograma vemos primero que todo claramente los promedios de expectativa de vida ahora notamos entre 50 años un aumento y después llegamos a 70-80 aproximadamente después caída. En la mortalidad de niños hay bastante frecuencia de mortalidad en varios grupos de países en algunos de ellos con índices muy altos de 40 y 20 se mantiene después 10

que corresponden a un porcentaje grande de países digamos más de un 30%, vemos *outliers* con valores muy altos, pero de poca frecuencia dan prueba de hechos o asesinatos en masa.



En la matriz de correlación los valores se encuentran ordenados podemos ver claramente en la tope correlación entre income y expectativa de vida como vimos en el análisis univariante a mayores ingresos mayor expectativa de vida. En el fondo también vemos la correlación entre mortalidad en niños e inflación vemos también que la variable total fer está correlacionada con mortalidad infantil.

1.4.Clustering

El objetivo de algoritmo K-means es agrupar datos numéricos en grupos K estos grupos se representan por vectores centrados K. Las observaciones en un grupo están cerca del centro de ese grupo que de otros centros.

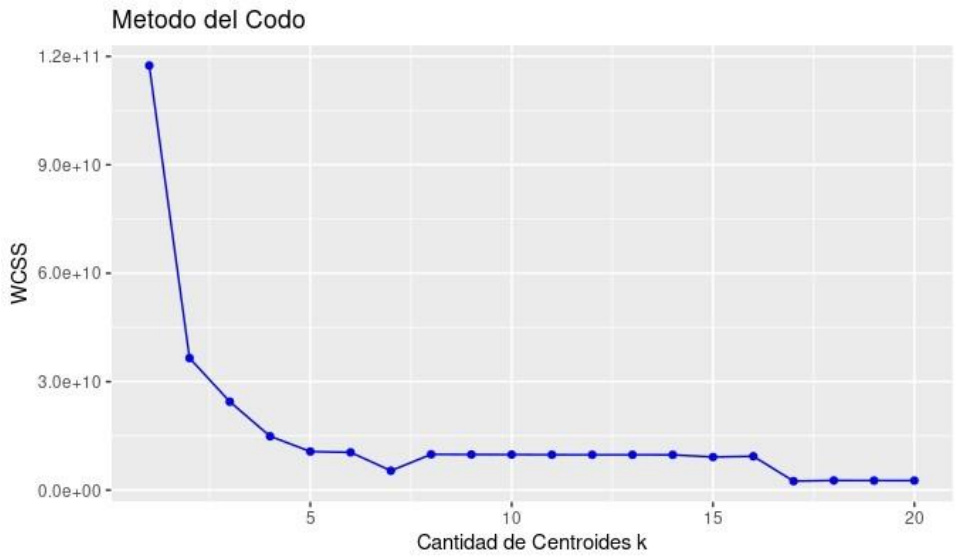
Antes de empezar el proceso de agrupamiento *K-means* hay que tener en cuenta si se van a tratar los *Outliers* que ya se han mencionado en el análisis descriptivo también de acuerdo a la investigación hecha se considera antes de empezar el *clustering* un proceso de escalado de datos y la evaluación de datos para ver si son apropiados para el algoritmo K-means

La librería *factoextra* contiene la estadística de Hopkins para evaluar la tendencia de agrupamiento o clustering, se evaluó y el resultado fue:

[1] 0.933468

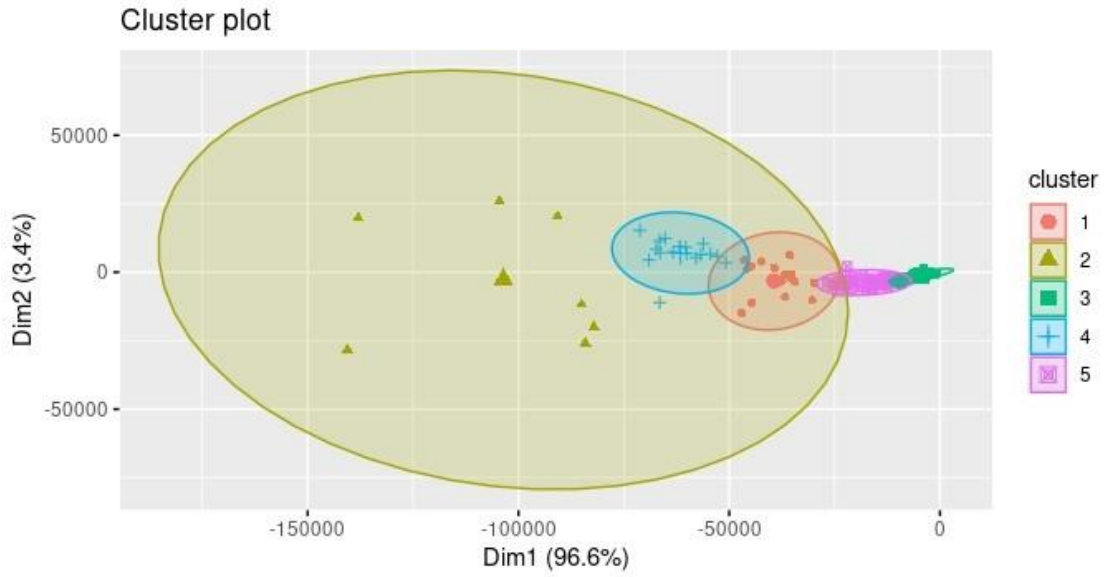
Ya teniendo un grupo de datos en el que podemos hacer un agrupamiento podemos evaluar el valor óptimo de K para el algoritmo *K-means*. EL valor K seria en cuantos grupos dividimos los datos o el valor optimo.

Método del Codo que vimos en clase anexo la librería al final

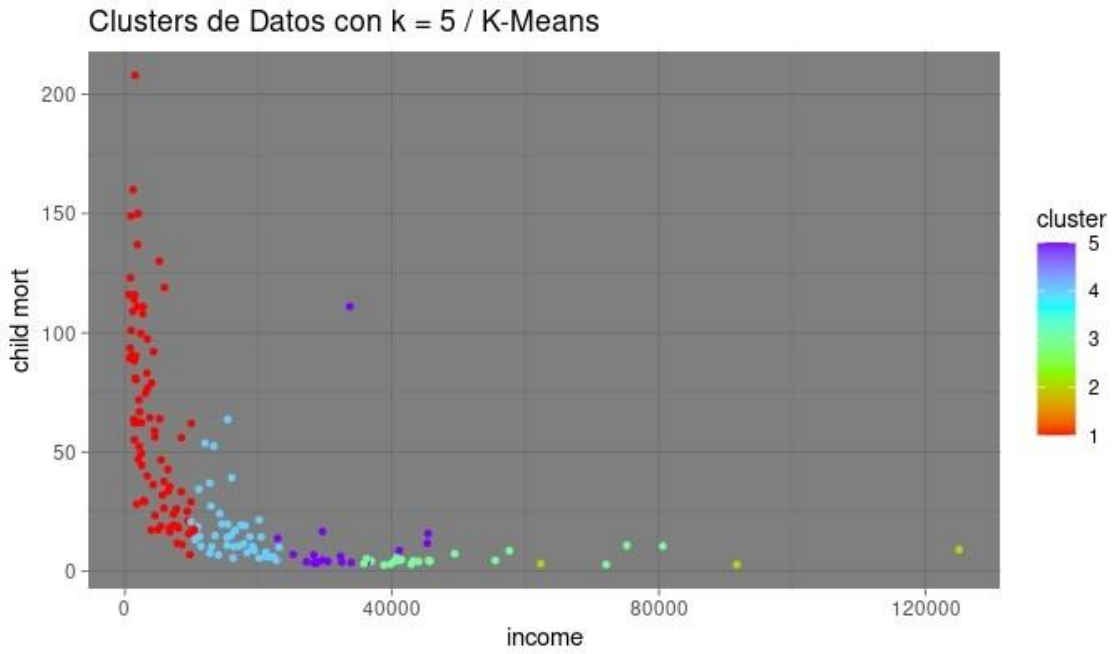


Se opto al final por un K = 5 con este valor tenemos un mejor agrupamiento de los países en este caso

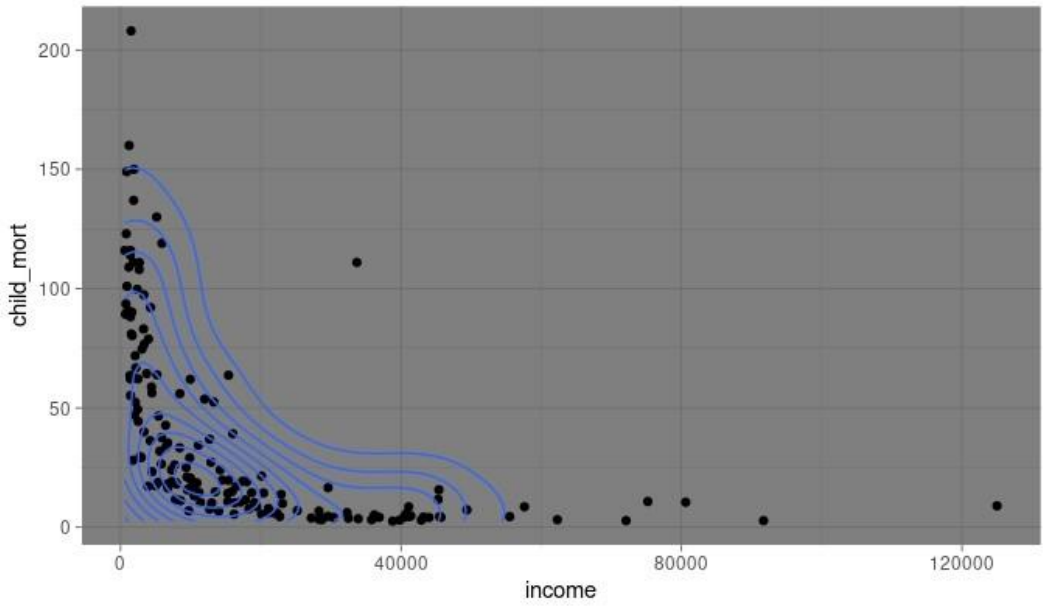
Para este grafica se ve el agrupamiento general de los clústeres.



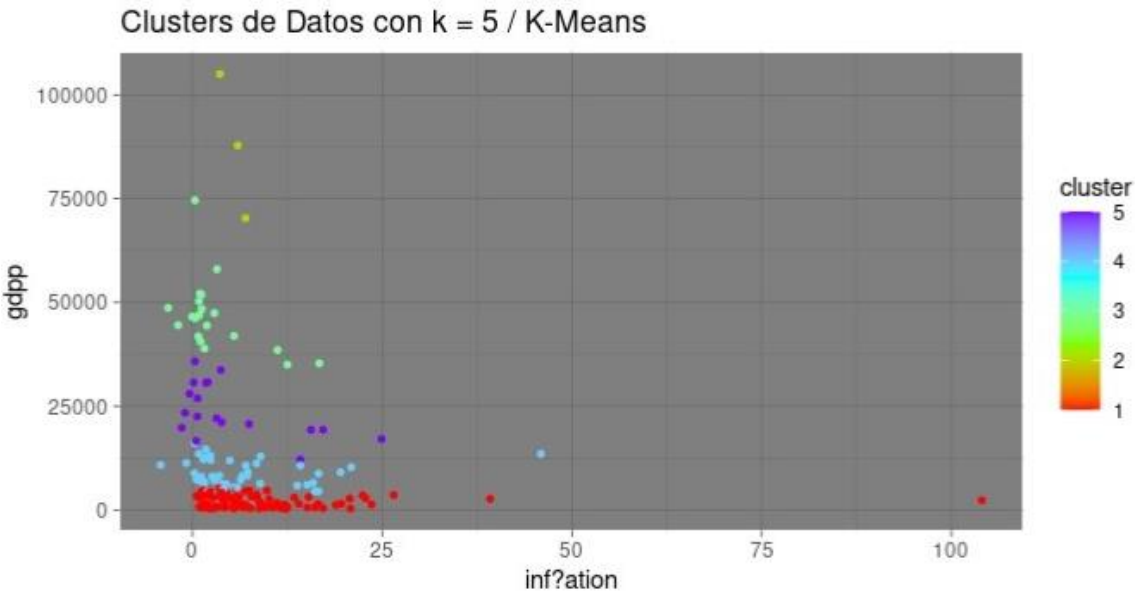
Vamos a detallar diferentes graficas con nuestro nuevo *dataframe* **dfk** el cual ya contiene una columna adicional que es el respectivo valor del clúster de las diferentes naciones en este caso, recordad que en este momento la columna cualitativa del país está excluida para el proceso *Kmeans*.



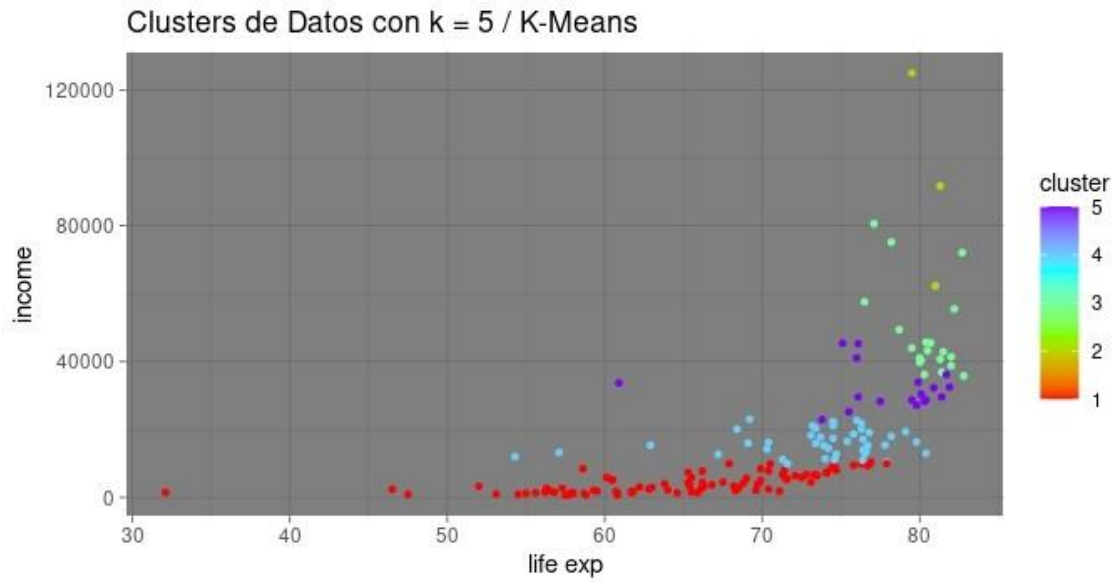
El índice de mortalidad que debería estar en rojo por el *seed* que se usó muestra en el eje y un alto índice de mortalidad infantil en los países con bajo *income* o ingresos....



...



Aquí se muestra que la inflación es mayor en los países con bajo *gdpp* o dicho de otra forma las naciones con mayor *gdpp* tienden a tener un menor índice de inflación



En la última grafica se compara la expectativa de vida comparado con los ingresos. Se nota que si el país tiene menos ingresos la tendencia de la edad que mueren es menor.

1.5.Análisis de resultados

Anexo el código en R de todo el proceso. Al final creamos un df nuevo se agregó la información de países del df original con comandos con el siguiente se pueden mostrar los clústeres en pantalla siendo 2 uno de los países necesitados de ayuda y con bajo ingresos

```
fviz_dend(hclust(dist(dfk)), k = 5,  cex = 0.5)


dfk2<-df
# Al final de Kmeans asignamos el valor cluster a una copia del
dfk2$cluster<-dfk$cluster

# Con este comando vemos las entradas en en cluster asignado si
dfk2[dfk2$cluster=="2",]
```

Países Globalmente que necesitan ayuda

```
Session restored from your saved work on 2022-Sep-27 15:54:14 UTC (2 minutes ago)
> head(dfk2[order(dfk2$income, dfk2$child_mor,  dfk2$gdpp),])
                    country child_mort exports health imports income inflation life_expec total_fer gdpp cluster
38         Congo, Dem. Rep.      116.0   41.10   7.91    49.6    609     20.80       57.5      6.54  334       2
89                  Liberia       89.3   19.10  11.80    92.6    700      5.47       60.8      5.02  327       2
27                  Burundi       93.6    8.92  11.60    39.2    764     12.30       57.7      6.26  231       2
113                   Niger      123.0   22.20   5.16    49.1    814      2.55       58.8      7.49  348       2
32  Central African Republic      149.0   11.80   3.98    26.5    888      2.01       47.5      5.21  446       2
107              Mozambique      101.0   31.50   5.21    46.2    918      7.64       54.5      5.56  419       2
>
```

Ahora creamos un subset para Latinoamérica

```
dfl <-subset(dfk2 ,  country=="Antigua and Barbuda" | country=="Argentina"  |
              country=="Barbados" | country=="Belize" | country=="Bolivia"  | country=="Brazil"  |
              country=="Colombia" | country=="Costa Rica" | country=="Chile"  | country=="Dominican Republic"
              country=="Ecuador" | country=="El Salvador" | country=="Grenada"  | country=="Guatemala"  |
              country=="Guyana" | country=="Haiti" | country=="Jamaica"  | country=="Panama"  |
              country=="Paraguay" | country=="Peru" | country=="St. Vincent and the Grenadines"  |
              country=="Suriname"  |  country=="Uruguay" | country=="Venezuela" | country=="Guyana" )


head(dfl[order(dfl$income, dfl$child_mor, dfl$gdpp),])
```

(Top Level)

```
 Central African Republic      149.0   11.80   3.98    26.5    888      2.01       47.5      5.21  446
97               Mozambique      101.0   31.50   5.21    46.2    918      7.64       54.5      5.56  419
head(dfl[order(dfl$income, dfl$child_mor, dfl$gdpp),])
      country child_mort exports health imports income inflation life_expec total_fer gdpp cluster
7       Haiti      208.0    15.3   6.91    64.7   1500      5.45       32.1      3.33  662       2
9     Bolivia       46.6    41.2   4.84    34.3   5410      8.78       71.6      3.20 1980       2
5      Guyana       37.6    51.4   5.38    79.1   5840      5.73       65.5      2.65 3040       2
3   Guatemala       35.4    25.8   6.85    36.3   6710      5.14       71.3      3.38 2830       2
19   Paraguay       24.1    55.1   5.87    51.5   7290      6.10       74.1      2.73 3230       2
9  El Salvador       19.2    26.9   6.91    46.6   7300      2.65       74.1      2.27 2990       2
```

Quiero mencionar que se trataron varios procesos de escalado y eliminación de outliers pero se presentaron comportamientos extraños en las gráficas valores negativos en un futuro se puede profundizar un poco para realizar estos procesos en caso la agrupación lo necesite.

2. Serie Temporal

2.1.Descripción de los datos

El conjunto de datos con el que se va a trabajar contiene el número de incendios forestales en Brasil desde 1998 hasta el 2017.

Los datos están divididos por año, meses y estados de Brasil.

A continuación, se describe el dataset con el que se va a trabajar.

- Year: Año en que ocurren los incendios forestales.
- State: Estados brasileños.
- Month: Mes en que ocurren los incendios forestales.
- Number: Número de Incendios Forestales reportados.
- Date: Fecha en que se reportaron los Incendios Forestales.

2.2.Motivación del análisis estadístico

Los incendios forestales son un grave problema para la conservación de los bosques, por tal motivo es recomendable identificar la frecuencia con la que pasan estos eventos.

Para comprender la frecuencia de los incendios forestales en Brasil con una serie temporal, aportaría a tomar medidas para prevenirlos.

Con los datos es posible evaluar los incendios forestales a lo largo de los años, así como en las regiones que suceden.

2.3.Estadística descriptiva

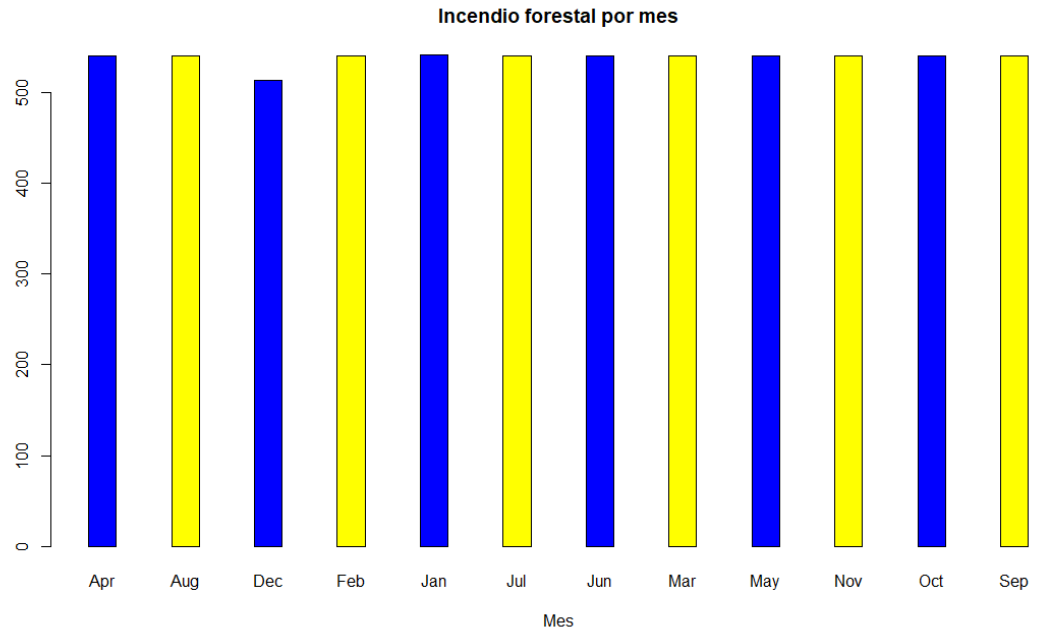
Para este análisis se cuenta con un dataset de 5 columnas y 6454 registros.

	year	state	month	number	date
1	1998	Acre	Janeiro	0	1998-01-01
2	1999	Acre	Janeiro	0	1999-01-01
3	2000	Acre	Janeiro	0	2000-01-01
4	2001	Acre	Janeiro	0	2001-01-01
5	2002	Acre	Janeiro	0	2002-01-01
6	2003	Acre	Janeiro	10	2003-01-01
7	2004	Acre	Janeiro	0	2004-01-01
8	2005	Acre	Janeiro	12	2005-01-01
9	2006	Acre	Janeiro	4	2006-01-01
10	2007	Acre	Janeiro	0	2007-01-01

La columna *month* la cambiaremos por las abreviaturas de los meses en el idioma inglés.

```
> unique(dataForest$month)
 [1] "Janeiro"   "Fevereiro" "Mar\xe7o"  "Abril"     "Maio"      "Junho"
 [7] "Julho"     "Agosto"    "Setembro"  "Outubro"   "Novembro"  "Dezembro"
> unique(dataForest2$month)
 [1] "Jan" "Feb" "Mar" "Apr" "May" "Jun" "Jul" "Aug" "Sep" "Oct" "Nov" "Dec"
```

En la siguiente gráfica se visualiza la cantidad de registros que se tiene por mes.




```
> table(dataForest2$month)

Apr Aug Dec Feb Jan Jul Jun Mar May Nov Oct Sep
540 540 513 540 541 540 540 540 540 540 540 540
```

Con estos datos, se puede observar que los datos del mes de *Dec* y *Jan* son diferentes al resto de meses.

Análisis Mes

En la siguiente consulta, se identificará esa diferencia por año.

```
> table(dataForest2$year)

1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013
 324  324  324  324  324  324  324  324  324  324  324  324  324  324  324  324
2014 2015 2016 2017
 324  324  324  298
```

Se puede observar que el año 2017 tiene menos datos en comparación al resto de años.

```
> table(year_2017$month)

Apr Aug Feb Jan Jul Jun Mar May Nov Oct Sep
 27  27  27  28  27  27  27  27  27  27  27
```

Con este análisis se puede concluir que existe la misma cantidad de datos por cada año excepto para el 2017 que no tiene datos totalmente actualizados, falta el mes de diciembre. También se puede observar que el mes de *Jan* tiene un dato adicional.

```
> year_2017 %>% filter(month == "Jan") %>% head(10)
   year            state month number       date
1  2017             Acre   Jan      0 2017-01-01
2  2017          Alagoas   Jan     38 2017-01-01
3  2017          Alagoas   Jan     38 2017-01-01
4  2017            Amapa   Jan      2 2017-01-01
5  2017         Amazonas   Jan     65 2017-01-01
6  2017            Bahia   Jan    154 2017-01-01
7  2017            Ceara   Jan     91 2017-01-01
8  2017 Distrito Federal   Jan      1 2017-01-01
9  2017   Espirito Santo   Jan     13 2017-01-01
10 2017            Goias   Jan     53 2017-01-01
```

En la consulta anterior se puede observar que existe el estado Alagoas se repite por dos ocasiones con los mismos datos.

Análisis de State

En la siguiente consulta se puede observar los datos de *State*.

- La mayoría de los estados tienen el mismo número 239.
- El state de Alagoas tiene un registro adicional en comparación al resto, se puede suponer que se debe a un error de entrada en el dataset debido a tener el registro duplicado. Al tener este análisis se puede eliminar dicho registro.
- Los estados de Mato Grosso y Paraiba contienen el doble de registros en comparación al resto.
- El estado de Rio contiene el triple de registros.

```
> table(dataForest2$state)

       Par\xe1           Acre        Alagoas          Amapa
           239            239            240            239
      Amazonas          Bahia         Ceara Distrito Federal
           239            239            239            239
 Espirito Santo          Goias       Maranhao   Mato Grosso
           239            239            239            478
  Minas Gerais         Paraiba      Pernambuco          Piau
           239            478            239            239
           Rio       Rondonia        Roraima Santa Catarina
           717            239            239            239
     Sao Paulo        Sergipe      Tocantins
           239            239            239
```

Ilustración 1 Mapa político con la división entre los estados de Brasil
Fuente: https://mapamundi.online/america/del-sur/brasil/

Con los datos obtenidos por estado y el mapa político de Brasil se puede concluir que existen 3 estados que comienzan con Rio, 2 estados que comienzan con Mato Grosso y solo hay un estado de Paraiba.

```
> dataForest2 %>%
+   filter(state %in% c("Piau","Bahia","Rio","Mato Grosso"), month == "Apr", year == 20
06)
  year       state month number       date
1 2006       Bahia   Apr     60 2006-01-01
2 2006 Mato Grosso   Apr    161 2006-01-01
3 2006 Mato Grosso   Apr     51 2006-01-01
4 2006        Piau   Apr      1 2006-01-01
5 2006         Rio   Apr      8 2006-01-01
6 2006         Rio   Apr      2 2006-01-01
7 2006         Rio   Apr     59 2006-01-01
```

Con la consulta anterior se puede confirmar la distribución de los estados.

Se debe etiquetar y diferenciar los estados repetidos, pero debido a la falta de información que proporciona el dataset, se puede intentar solucionar asumiendo que los registros están ordenados.

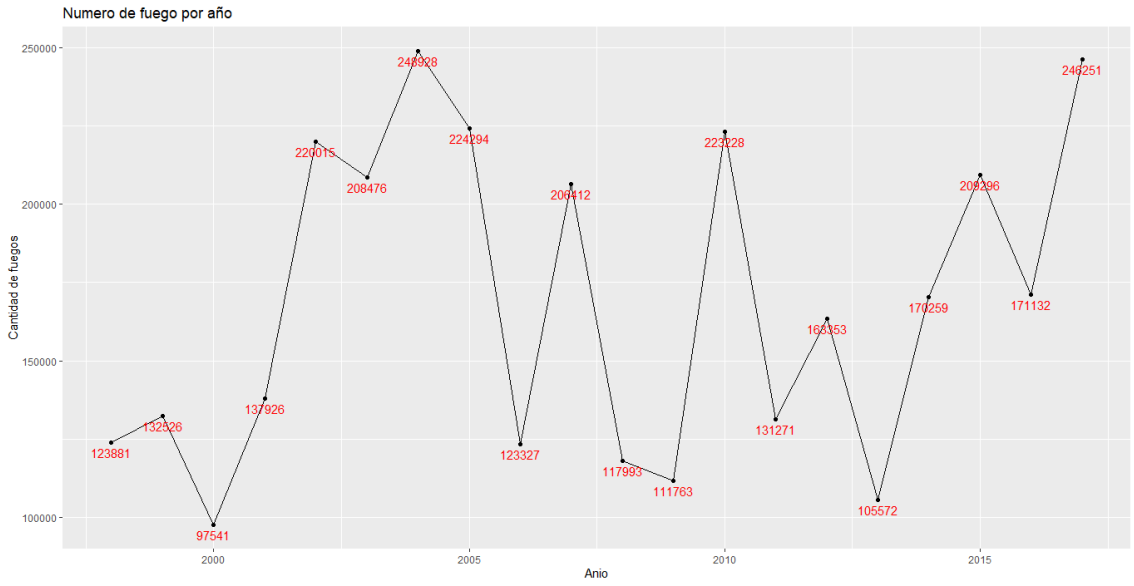
A continuación, se presentarán los datos moldeados a los cambios mencionados.

```
> head(newDataForest %>% arrange(year, month), 10)
   year          state month number
1  1998 Mato Grosso 1   Apr      0
2  1998 Mato Grosso 2   Apr      0
3  1998     Paraiba 1   Apr      0
4  1998     Paraiba 2   Apr      0
5  1998         Rio 1   Apr      0
6  1998         Rio 2   Apr      0
7  1998         Rio 3   Apr      0
8  1998          Acre   Apr      0
9  1998       Alagoas   Apr      0
10 1998         Amapa   Apr      0
> table(newDataForest$month)

Apr Aug Dec Feb Jan Jul Jun Mar May Nov Oct Sep
540 540 513 540 540 540 540 540 540 540 540 540
```

```
> table(newDataForest$state)

          Acre          Par\xe1        Alagoas          Amapa
           239             239             239             239
      Amazonas          Bahia          Ceara Distrito Federal
           239             239             239             239
 Espirito Santo          Goias        Maranhao    Mato Grosso 1
           239             239             239             239
  Mato Grosso 2   Minas Gerais       Paraiba 1       Paraiba 2
           239             239             239             239
     Pernambuco           Piau           Rio 1           Rio 2
           239             239             239             239
         Rio 3       Rondonia         Roraima  Santa Catarina
           239             239             239             239
      Sao Paulo        Sergipe       Tocantins
           239             239             239
```

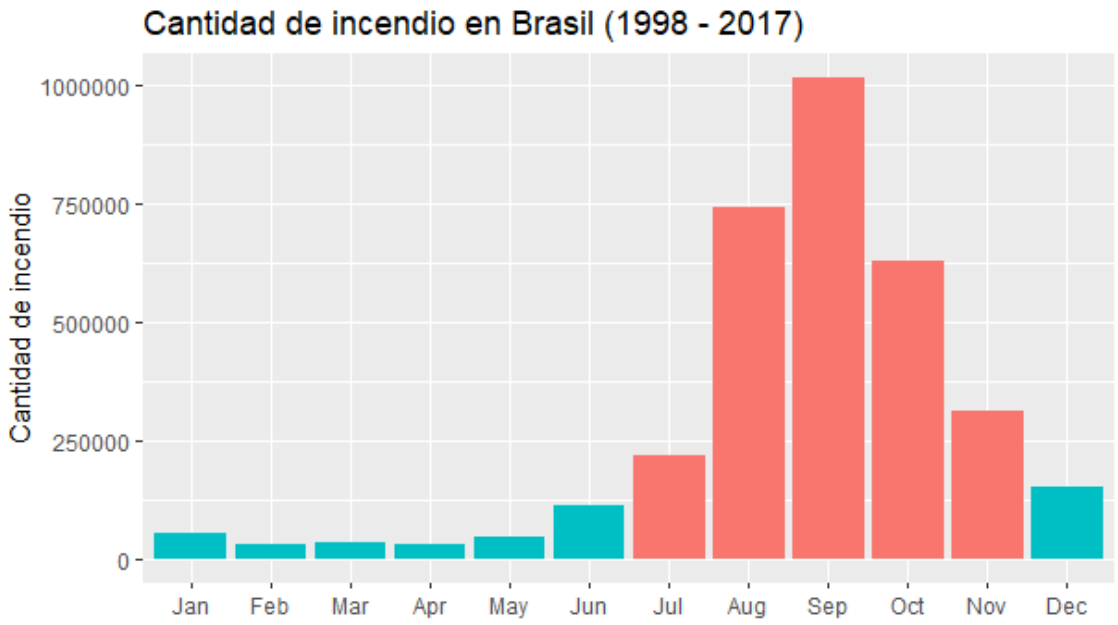
2.4.Descomposición de la serie temporal

La siguiente imagen representa una serie temporal por año.



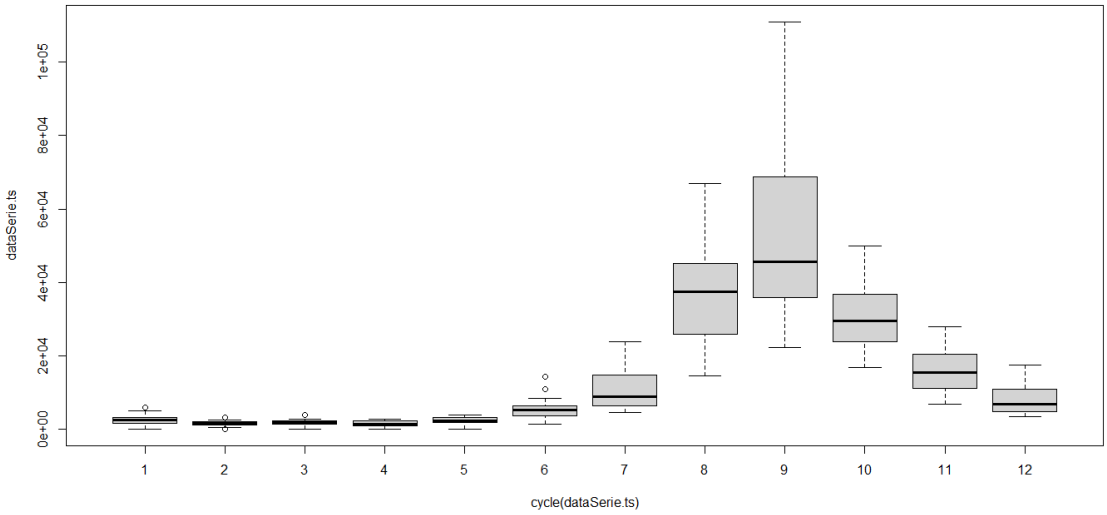
2003 fue el peor año en el recuento de incendios forestales en los estados de Brasil.

Para el año 2016 existe de la misma manera una cantidad muy elevada de incendios forestales



Con la gráfica anterior se puede observar que la temporada de mayor cantidad de incendio es desde Julio hasta noviembre.

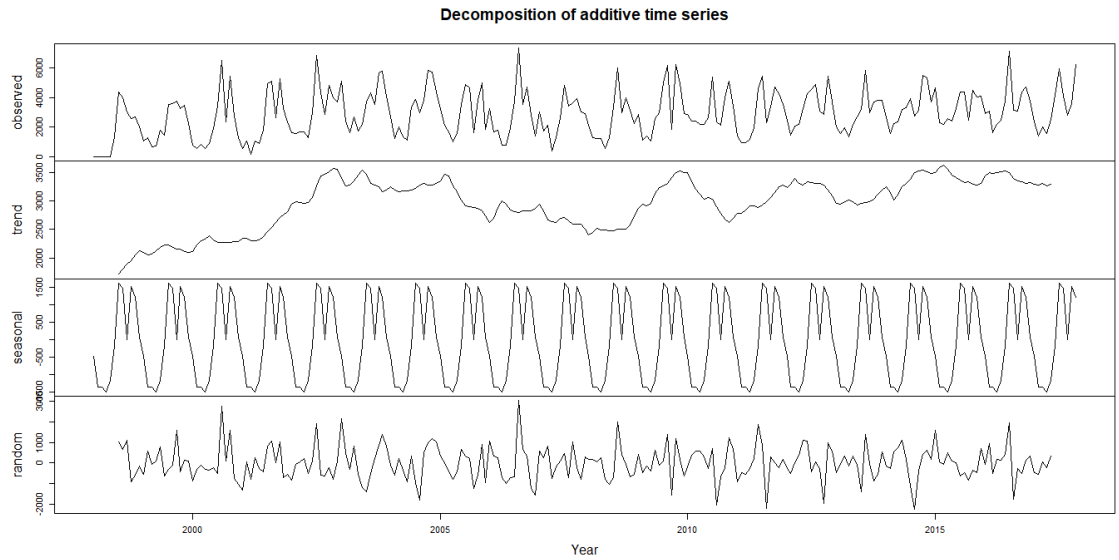
Con esta gráfica se puede observar que el mes de diciembre también hay incendios forestales, pero no se vio reflejada en la serie temporal por mes, debido que el año 2017 no se tienen datos para el mes de diciembre.



Con las gráficas anteriores y la gráfica de Boxplot se puede observar que el mes de septiembre, es donde tiene más incendios forestales a diferencias de los otros meses.

De la misma manera la temporada desde Julio hasta noviembre es donde ocurren más incendios forestales.

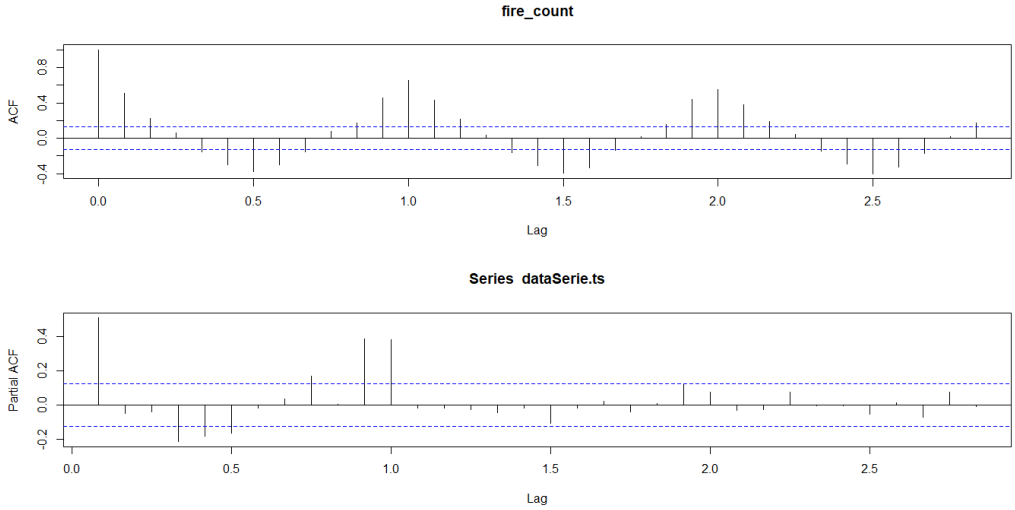
Con la siguiente gráfica se puede observar con mejor detalle el análisis de la serie temporal.



La tendencia de la serie se comporta como estacionaria.

2.5.Moving average

A continuación, se va a explorar el modelo que se usara para la serie temporal.



Se usa el auto.arima para que nos brinde los parámetros adecuados del modelo.

Se evaluará por cada dos años y a su vez por 2 medias móviles por cada 12 meses.

El valor de AIC sería el más bajo en comparación a otros modelos.

```
> ArimaModel <- auto.arima(dataSerie.ts)
> ArimaModel
Series: dataSerie.ts
ARIMA(2,0,2)(0,1,2)[12] with drift

Coefficients:
         ar1     ar2     ma1      ma2     sma1     sma2   drift
      0.0637  0.8542  0.0597  -0.8573  -0.7115  -0.1546  6.0630
s.e.  0.0814  0.0713  0.0837   0.0753   0.0775   0.0738  2.7563

sigma^2 = 870864:  log likelihood = -1878.76
AIC=3773.51   AICc=3774.17   BIC=3800.91
```

Modelo con parámetros diferentes.

AIC incrementa.

```
> (fitARIMA <- arima(dataSerie.ts,
+                    order=c(1,0,0),method="ML"))

Call:
arima(x = dataSerie.ts, order = c(1, 0, 0), method = "ML")

Coefficients:
         ar1  intercept
      0.5276  2925.0112
s.e.  0.0560   178.4221

sigma^2 estimated as 1713618:  log likelihood = -2054.61,  aic = 4115.21
>
```