Situación Problemática:

Conforme a los años que han pasado podemos notar que el precio de la canasta básica ha cambiado de precio. Gracias a las fuentes de información: WFP, RRD, Program Unit SO Tindouf Algeria,VAM Unit SO Tindouf, ONASA, SIMA, el gobierno no Benin, INE, entre otras pudimos encontrar la información respecto al precio de diferentes años de los alimentos básicos en países alrededor del mundo. Ya que como seres humanos es vital ingerir alimento, el cual es comprado; surge la inquietud de saber si este precio subirá, se mantendrá o bajará en los próximos años.

Problema científico:

Se desea saber si el precio de la canasta básica va a cambiar en los próximos años. La moneda alrededor del mundo, el valor que tiene se va deteriorando haciendo que todo lo que tiene que ver con el movimiento consumista aumente en su valor. Con esto se encontraran personas sin la capacidad de cubrir como mínimo los alimentos básicos para una buena alimentación con esto se aumentará la tasa de desnutrición alrededor del mundo.

Objetivos:

Crear un análisis predictivo en el precio de la canasta básica por si esta subirá o bajará porcentualmente de forma global en los próximos años. Para ser específicos se buscará hacer una predicción de los precios de los alimentos básicos del año actual a 5 años adelante.

Descripción de datos

Estos datos cuentan con variables bastante independientes entre sÍ. solamente se encontró que el coeficiente de correlación más elevado es entre la ciudad y el tipo de cambio el cual es de 0.558. Además de este todo es independiente entre sÍ. Este dataset tiene un acantidad de 18 variables, con un total de 783788 observaciones. Se cuentan tanto con variables cuantitativas como cualitativas.

Los datos encontrados son datos recaudados alrededor del mundo acerca de los precios de algunos alimentos en diferentes países con su respectivo peso, moneda, ciudad, etc. Para comenzar se realizó un análisis de datos con una tabla de correlación la cual se muestra a lo largo del documento.

Información de los datos:

Las variables de las cuales consta el documento son las siguientes junto con 783,788 filas de datos, a continuación se encuentra el significado de cada una de ellas.

Nombre de la variable	Descripción	Tipo de variable
"adm0_id"	Id del país	Numérica
"adm0_name"	Nombre del país	Categórica
"adm1_id"	Departamento, estado o provincia	Numérica
"adm1_name"	Nombre del departamento, estado o provincia	Categórica
"mkt_id"	Id de la ciudad	Numérica
"mkt_name"	Nombre de la ciudad	Categórica
"cm_id"	Id de la comida	Numérica
"cm_name"	Nombre de la comida	Categórica
"cur_id"	Id del tipo de cambio	Numérica
"cur_name"	Nombre del tipo de cambio	Categórica
"pt_id"	Id del nombre del sector económico	Numérica
"pt_name"	Nombre del sector económico	Categórica
"um_id"	Id de la cantidad de masa	Numérica
"um_name"	Cantidad de masa	Categórica
"mp_month"	Mes en el que se midió el producto	Numérica
"mp_year"	Año de medición	Numérica
"mp_price"	Precio del producto	Numérica
"mp_commoditysource"	Nombre de la fuente de datos	Categórica

Limpieza de datos

Precios en dólares estadounidenses:

Como se lee en la descripción de datos se encuentra información de varios países por ende consta de varias monedas, por lo que llevamos a cabo un conversión con el tipo de cambio del día 1 de septiembre del año 2018. Con este cambio se puede eliminar la columna de moneda ya que todas serán la misma, lo que hace no necesario el tenerla. En el código se usó el mismo para todos los tipos de moneda.

```
#ZMW - 0.096
rows <- datos[which(datos$cur_name == "ZMW"),]
rows$mp_price <- rows$mp_price*0.096
datos[which(datos$cur_name == "ZMW"),] <- rows
```

Eliminación de las columnas del tipo de moneda:

Las mayoría de las variables son conforme el nombre de las variables y el id de las mismas, por lo que al eliminar la variable de la moneda también sus id´s.

```
datos$cur_id <- NULL
datos$cur_name <- NULL
```

Eliminación de filas con precio 0:

Si nuestro objetivo es predictivo, se debe eliminar la mayor cantidad de desviación que se pueda de la base de datos. En algunos datos se encuentra en 0 el precio de los productos, sin embargo, ningún producto es gratis según una investigación realizada, por lo que se procede a eliminar dichos datos y así reducimos el ruido en el análisis. Con esto también se eliminan las filas vacías las cuales constaban solo del id de algunos países correspondientes.

```
datos <- datos[-which(datos$mp_price == 0)]
```

Eliminación de algunos artículos alimenticios irrelevantes:

- *Livestock:*
  Podemos observar al analizar los precios, habían algunos datos atípicos por lo que al evaluar existían artículos irrelevantes como animales para consumo humano como pollos, cerdos, reses, etc. los precios de dicho artículos se colocaban tomando en cuenta la edad del animal y estos aun con vida. Este tipo de artículos no pueden ser medidos para un estudio de alimentos básicos para una persona, lo cual hace factible la eliminación de los mismos.

- *Unión de elementos*
  En artículos como frutas, verduras, legumbres y lacteos se encuentran repetidos pero con nombres o cantidades diferentes, por lo que se procede a juntarlos.

Análisis exploratorio

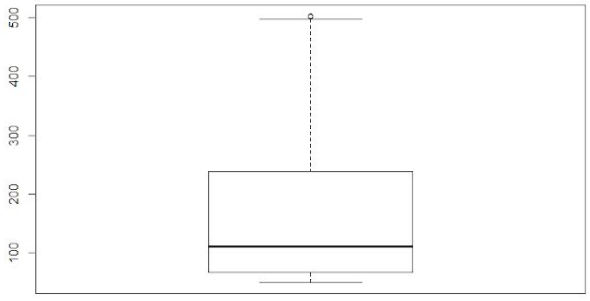
Las variables cuantitativas constan la mayoría de id's de las variables categóricas, se estudiaron mediante estadísticas descriptivas las variables de importante en el conjunto, entre ellas se encuentra el país, la comida, tipo de cambio, año y precio de dicha comida. En este caso se escogieron esas debido a que son las más generales para discutir las predicciones, son las variables que más importan a nivel global.

```
> #nombre de la comida
> summary(datos$cm_id)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     50      67     111     162     239     503
> #pais
> summary(datos$adm0_id)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    1.0   115.0   170.0   959.9   205.0 70001.0
> #tipo de cambio
> summary(datos$cur_id)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  22.00   31.00   63.00   57.42   77.00   95.00
> #precio
> summary(datos$mp_price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0      30     175    4235     504 5833333
> #año
> summary(datos$mp_year)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1992    2011    2014    2013    2015    2017
> |
```

Para hacer una revisión de los datos atípicos encontrados en los datos se van a realizar gráficos de caja y bigotes. Donde la línea en medio de la caja es la mediana la cual divide en 50 50 los datos.

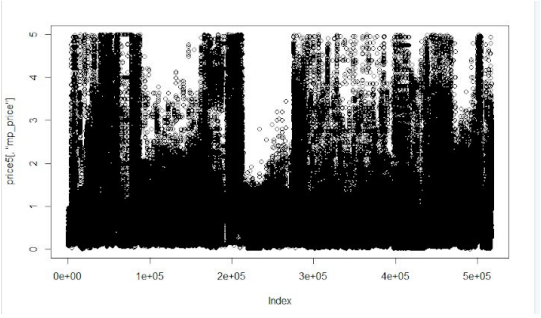
Comida

En esta variable se encuentra que en una mitad están menos dispersos que la otra y en esa misma se encuentra un dato atípico en la parte entre 400 - 500.

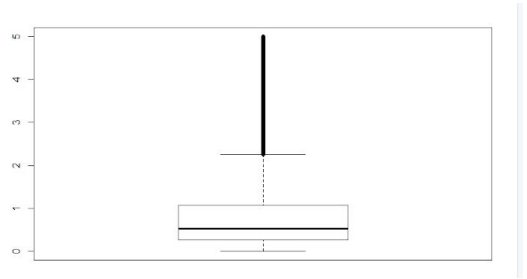


Precio

Los precios están muy dispersos sin embargo la mayoría se encuentran entre el precio de 0.2 a 5, ya que en la limpieza se quitaron los datos que tenían valor 0 y al poner todos en un precio estándar se puede observar que la distribución de los datos es bastante variada.

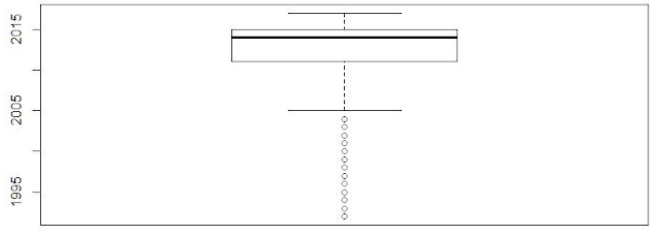


En esta gráfica se muestra que aunque en plot anterior se ven muy dispersos alrededor del cuadro en realidad los datos siguen estando entre 0-1 dólares y si hacemos la analogía hacia quetzales la mayoría de datos se encuentran en un precio entre Q 0 - Q 7.5.



Año

En la variable de años según el diagrama de caja y bigotes podemos observar que los años que nos ayudarían más es del 2005 hacia el 2017 debido a que por debajo de estos se tiene muy poca información



En la siguiente imagen se demuestra que entre la variables numéricas se tiene independencia, siendo la mayoría de ellas id. El coeficiente de correlación más elevado es entre la ciudad y el tipo de cambio el cual es de 0.558.

```
>
> numdata <- datos[c(1,3,5,7,9,11,13,15,17)]
> cor(numdata)
             adm0_id      adm1_id       mkt_id       cm_id      cur_id       pt_id       um_id     mp_month     mp_price
adm0_id   1.000000000 -0.006097482  0.064510623 -0.031140056 0.099794940  0.031012434 0.180174332  0.002049329 -0.009393203
adm1_id  -0.006097482  1.000000000  0.164457824  0.064479933 0.047966388  0.014457845 0.085320771 -0.004492119 -0.004293638
mkt_id    0.064510623  0.164457824  1.000000000  0.254048721 0.558547455  0.066791527 0.088780281  0.002726445 -0.011469169
cm_id    -0.031140056  0.064479933  0.254048721  1.000000000 0.324316512  0.083560408 0.205512057  0.003153072  0.011317339
cur_id    0.099794940  0.047966388  0.558547455  0.324316512 1.000000000  0.070499033 0.085650431  0.015601708  0.003464396
pt_id     0.031012434  0.014457845  0.066791527  0.083560408 0.070499033  1.000000000 -0.044784517  0.003153959 -0.131667531
um_id     0.180174332  0.085320771  0.088780281  0.205512057 0.085650431 -0.044784517  1.000000000 -0.003464338  0.035885140
mp_month  0.002049329 -0.004492119  0.002726445  0.003153072 0.015601708  0.003153959 -0.003464338  1.000000000  0.001505465
mp_price -0.009393203 -0.004293638 -0.011469169  0.011317339 0.003464396 -0.131667531  0.035885140  0.001505465  1.000000000
>
```

Variables categóricas

Para evaluar las variables categóricas se hicieron tablas de frecuencia de las variables. A Continuación se mostraran las importantes:

```r
#--------VARIABLES CATEGORICAS-----

#Tablas de frecuencias
#Nombre de comida
table(datos$cm_name)
#nombre del pais
table(datos$adm0_name)
#Nombre del tipo de cambio
table(datos$cur_name)
#cantidad de masa
table(datos$um_name)
#fuente de datos
table(datos$mp_commoditysource)
```

CLUSTER

Para la parte de cluster se hizo mediante rangos de años en los datos, debido a que nuestros datos llegaban a casi un millón. El rango se contempló para cada 5 años comenzando desde 1992 hasta el 2017, la razón por la cual se decidió hacer cada 5 fue porque según nuestro análisis exploratorio los datos entre 1992 y 1996 eran casi atípicos por lo cual hacía ventajosa su manipulación si se lograba analizar cómo un grupo separado, de esta manera evitaremos la distorsión o posibles errores en a siguiente parte la cual es la predicción de precios.

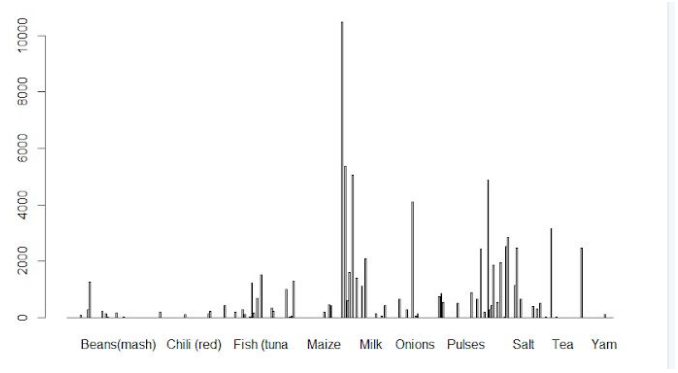
En cada grupo de se utilizó el método de eclidean y usamos el método de ward para poder obtener la mejor cantidad de cluster. El número asignado para K el cual es 5 se escogió por el hecho que las variables relevantes o más influyentes en el precio del producto son 5 en las cuales entra el precio mismo junto al país de origen, cantidad, nombre y año.

```r
#se dividieron los datos por year (5 por subset) ya que habian demasiados registros
#por cada grupo de 5 se realizo la agrupacion de clusters.
datos1<-subset(datos,mp_year >= 1992 & mp_year <= 1996)
#distancia
d1 <- dist(datos1, method = "euclidean")
fit1 <- hclust(d1, method="ward.D")
plot(fit1)#dendograma
groups <- cutree(fit1, k=5)
rect.hclust(fit1, k=5, border="red") #dendograma dividido por clusters

#datos de 1997 a 2001-----------------------------------------------------------
datos2<-subset(datos,mp_year >= 1997 & mp_year <= 2001)
d2 <- dist(datos2, method = "euclidean")
fit2 <- hclust(d2, method="ward.D")
plot(fit2)
groups2 <- cutree(fit2, k=5)
rect.hclust(fit2, k=5, border="red")

#datos de 2002 a 2006-----------------------------------------------------------
datos3<-subset(datos,mp_year >= 2002 & mp_year <= 2006)
d3 <- dist(datos1, method = "euclidean")
fit3 <- hclust(d3, method="ward.D")
plot(fit3)
groups3 <- cutree(fit3, k=5)
rect.hclust(fit3, k=5, border="red")

#datos de 2007 a 2011-----------------------------------------------------------
datos4<-subset(datos,mp_year >= 2007 & mp_year <= 2011)
d4 <- dist(datos1, method = "euclidean")
fit4 <- hclust(d4, method="ward.D")
plot(fit4)
groups4 <- cutree(fit4, k=5)
rect.hclust(fit4, k=5, border="red")

#datos de 2012 a 2017-----------------------------------------------------------
datos5<-subset(datos,mp_year >= 2012 & mp_year <= 2017)
d5 <- dist(datos1, method = "euclidean")
fit5 <- hclust(d5, method="ward.D")
plot(fit5)
groups5 <- cutree(fit5, k=5)
rect.hclust(fit5, k=5, border="red")
```

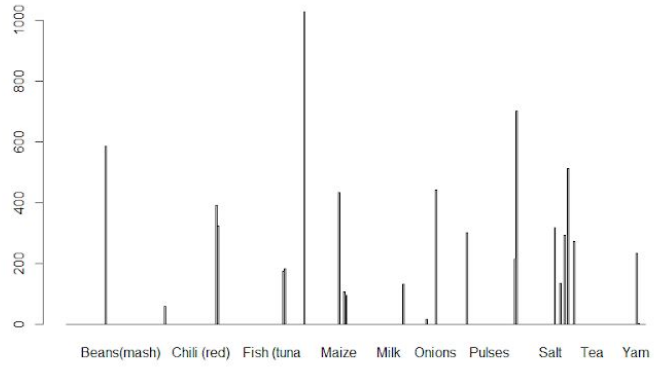
Se usa el método para sacar las k medias para cada grupo y con este procedimiento se obtienen los grupos por cluster.

```
#--------------------km--------------
#se obtienen los datos numericos de la matriz
nums <- unlist(lapply(datos, is.numeric))
#se usan esos datos para sacar el km y crear los grupos
numDatos1<-datos[ ,nums]
km1<-kmeans(numDatos1,4)
datos$grupo<-km1$cluster

#obtenemos el grupo de datos por cluster
g1<- datos[datos$grupo==1,]
prop.table(table(g1$cm_name))*100
plot(g1$cm_name)
g2<- datos[datos$grupo==2,]
prop.table(table(g2$cm_name))*100
plot(g2$cm_name)
g3<- datos[datos$grupo==3,]
prop.table(table(g3$cm_name))*100
plot(g3$cm_name)
g4<- datos[datos$grupo==4,]
prop.table(table(g4$cm_name))*100
plot(g4$cm_name)

#ploteamos los datos, al tener muchos rgistros se demora bastante en plotear
plotcluster(numDatos1,km1$cluster)
```

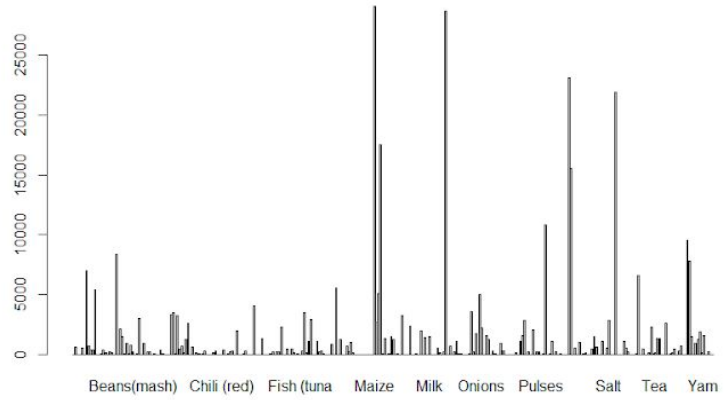
grupo 1



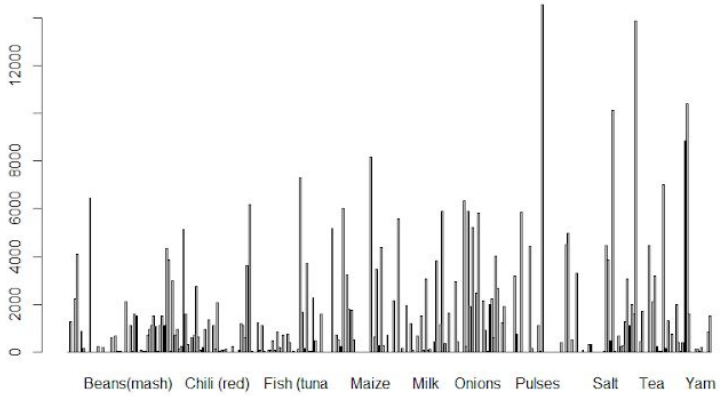
grupo 2



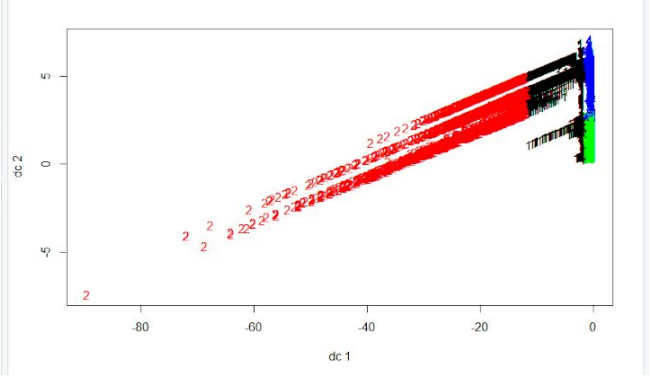
grupo 3



grupo 4



Plot del cluster



Al analizar la distribución en los grupos de la derecha es evidente su separación y las características del mismo, sin embargo el grupo más distante es el grupo 2 por debajo de 0 y en inclinación hacia el negativo por ambas partes.

Conclusiones:

- El precio de la comida si varía conforme al país y el año de producción.
- Los alimentos básicos se mantienen por debajo de 1 dólar estadounidense con la comida medida en libra.

- La diferencia de precios entre países se mantiene precisa, es decir que giran alrededor del mismo precio estándar.