Universidad del Valle
Data Science
Sección 10

Dana Pérez 17127
Jennifer Barillas
Luisa Arboleda
Anaís Castañeda 17291

RESULTADOS
PROYECTO FINAL

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 los alimentos 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 783,788 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

a. ***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
```

b. 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
```

c. 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)]
```

d. 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.

e. Estandarización de las cantidades de masa contra precio:

Esto se hizo pasando todas las cantidades de masas que el conjunto de datos tenía y se colocó todo en 1 kg, 1 L, docenas y paquetes.

f. Unión de grupos de alimentos

Se trabajó con Guatemala, Perú, Colombia y Bolivia. Estos tres países de Sudamérica fueron los que presentaron los datos más similares a Guatemala.

Para poder trabajar con los datos se hicieron grupos de alimentos. Se puede observar que en los países los bienes muchas veces son bastante específicos como por ejemplo colocan tipo de manzanas; verdes, rojas, amarillas.

Se unieron estos bienes en una misma variable que incluiría todos sus tipos. Luego se procedió a agrupar los datos en grupos alimenticios.

Los grupos que se tomaron en cuenta para clasificar son

1. Frutas
2. Verduras
3. Legumbres
4. Cereales
5. Carnes
6. Grasas y Azúcar
7. Lácteos (Fao, 2010)

A continuación se presentan los datos

Guatemala

```
> table(DatosGuate$cm_name)

                     Bananas                        Beans
                          84                          180
                       Bread                       Cheese
                         180                           67
                      Coffee                         Eggs
                          84                           84
Meat (beef, chops with bones)              Meat (chicken)
                          84                           84
                        Milk                Milk (powder)
                          84                           31
             Oil (vegetable)                      Onions
                          84                           84
                       Pasta                    Plantains
                          84                           84
                    Potatoes Rice (ordinary, second quality)
                          84                          180
                        Salt                        Sugar
                          84                           84
                    Tomatoes             Tortilla (maize)
                          84                          180
>
```

Datos agrupados

```
> DatosGuate[DatosGuate$cm_name %in% c("Beans"),"cm_name"]<-"Legumbres"
> table(DatosGuate$cm_name)

       Carnes       Cereales         Coffee         Frutas
          252            624             84            252
Grasa y azucar       Lacteos      Legumbres           Salt
          168            182            180             84
     Verduras
          168
>
```

Bolivia

```
> table(DatosBolivia$cm_name)

                       Bread                        Eggs
                         426                         258
Meat (beef, chops with bones)       Meat (chicken, whole)
                         745                         745
            Noodles (short)            Potatoes (Dutch)
                         743                         134
     Potatoes (Irish, imilla)            Potatoes (black)
                         321                          29
                        Rice                       Sugar
                         744                         258
>
```

Datos agrupados

```
> table(DatosBolivia$cm_name)

       Carnes       Cereales Grasa y azucar       Verduras
         1748           1913            258            484
>
```

Colombia

```
                Apples                 Bananas
                    60                     162
               Beans                Blackberry
                   616                      87
             Broccoli                 Cabbage
                    87                      64
              Carrots                 Cassava
                   163                     142
          Cauliflower                  Cheese
                    64                      68
   Chickpeas (imported)                  Cocoa
                   343                      77
               Coffee    Cucumbers (greenhouse)
                   124                      57
                 Eggs           Fish (tilapia)
                   442                     163
 Fuel (petrol-gasoline)                 Garlic
                   358                      44
                Guava                 Lentils
                    29                     246
               Maize              Maize flour
                   415                      56
             Mangoes              Meat (beef)
                    87                      71
   Meat (beef, minced)           Meat (chicken)
                   278                     288
          Meat (pork)        Milk (pasteurized)
                    59                     120
       Oil (sunflower)          Oil (vegetable)
                   167                     118
               Onions        Oranges (big size)
                   335                     129
               Papaya                   Pasta
                    58                      23
     Peas (green, dry)                Plantains
                   410                     140
             Potatoes          Potatoes (unica)
                   148                     242
              Pumpkin                    Salt
                    87                      23
              Spinach                   Sugar
                    87                     669
Tamarillos/tree tomatoes               Tomatoes
                    58                     156
          Wheat flour
                   405
```

Datos agrupados

```
> table(DatosCol$cm_name)

       Carnes       Cereales         Coffee         Frutas
         1301            899            124           1053
Grasa y azucar       Lacteos      Legumbres           Salt
         1031            188           1615             23
     Verduras
         1433
```

PERÚ

```
> table(DatosPeru$cm_name)

                        Maize                Oil (vegetable)
                          128                             19
                     Potatoes                           Rice
                          139                            138
                Sugar Wheat flour (locally processed)
                          140                            140
```

Datos agrupados

```
> table(DatosPeru$cm_name)

     Cereales Grasa y azucar       Verduras
          406            159            139
```

Antecedentes

Time series Data Analytics for stock market prediction using data mining techniques with R:

- En este “paper” se quería desarrollar un modelo de predicción para pronósticas las tendencias del mercado de valores basado en el análisis técnico utilizando datos históricos de series de tiempo y técnicas de extracción de datos. Los resultados del experimento obtenidos demostraron el potencial del modelo ARIMA para predecir los índice s de precios de las acciones obtenidas a corto plazo. Esto da permite guiar a los inversionistas a tomar decisiones para invertir, en el mercado de valores, ya sea para comprar/vender/ mantener participación en algunos proyectos. Con los resultados que se obtuvieron, el modelo ARIMA puede competir bien con las técnicas de pronóstico emergentes en la predicción a corto plazo.(Mahantesh *et all*, 2015)

Predicting House Sale Prices with R:

- En este paper by RStudio nos da un ejemplo de cómo se implementa la regresión lineal para poder predecir precios, en este caso el precio de las casas que se pondrán en venta. Es necesario generar el modelo inicial el cual nos ayudará a poder predecir cualquier situación cuantitativa que sea necesaria.

Algoritmos de predicción

Correlación lineal y regresión lineal

La correlación lineal y la regresión lineal simple son métodos estadísticos que estudian la relación entre dos variables.

- La correlación cuantifica la relación entre las variables mientras que la regresión lineal genera un modelo que permita predecir el valor de una a partir de otra variable.
- Por lo general de primero es necesario establecer si las dos variables están relacionados y luego se procede a generar el modelo de regresión lineal.

Correlación lineal

Para estudiar la relación lineal entre dos variables continuas es necesario disponer de parámetros que permitan cuantificar dicha relación. Como la covarianza la cual indica el grado de variación conjunta de dos variables aleatorias. La covarianza depende de las escalas en que se miden las variables, por lo tanto no es comparable entre distintos pares de variables. Se estandariza la covarianza y se obtiene lo que se conoce como coeficientes de correlación. Podemos obtener diferentes tipos de correlación, entre los que más destacan son coeficiente de Pearson, Rho de Spearman y Tau de Kendall.

La correlación de Pearson funciona con variables cuantitativas que tienen distribución normal como es el caso de los datos a analizar. Varía entre +1 y -1 siendo +1 una correlación positiva perfecta y -1 una correlación negativa perfecta. La correlación puede variar entre 0 a 0.9 siendo 0.5 una asociación moderada.

Además del valor obtenido para el coeficiente de correlación, es necesario calcular su significancia. Solo si el valor p es significativo se puede aceptar que existe correlación entre las variables. El test paramétrico de significancia estadística empleado para el coeficiente de correlación es el t-test. Al igual que ocurre siempre que se trabaja con muestras, por un lado está el parámetro estimado y por otro su significancia a la hora de considerar la población entera.

Se realiza un test de hipótesis, H0 que considera que las variables independientes como el coeficiente de correlación = 0 mientras que la hipótesis Ha considera que existe relación donde no puede ser igual a 0.

El coeficiente de Pearson es la covarianza estandarizada y su ecuación difiere dependiendo de si se aplica a una muestra o si se aplica a una población.

- Las condiciones que este presentan son:
- La relación que se quiere estudiar entre amba variables es lineal. ç
- Deben ser variables cuantitativas
- Las variables se tienen que distribuir de forma normal.
- La varianza de Y debe ser constante a lo largo de la variable X. Esto se puede identificar si en el “scatterplot” los puntos mantienen la misma dispersión en las distintas zonas de la variable X.

El coeficiente de Pearson toma valor entre -1 y +1 como ya mencionado, es una medida independiente de las escalas en las que se midan las variables. No tiene en consideración que las variables sean dependientes o independientes. El coeficiente de correlación de Pearson no equivale a la pendiente de la recta de regresión. Es sensivle a outliers.

Para hacer la correlación lineal se necesitan paquetes como **Mass y Ggplot2.**
Además R ya contiene funciones que permiten calcular los diferentes tipos de correlaciones y sus niveles de significancia: cor() y cor.test(). La segunda función nos ayuda a calcular el coeficiente de correlación e indica su significancia (p-value) e intervalo de confianza.

Con un diagrama de dispersión podremos identificar si existe relación lineal. La pendiente que tome la gráfica nos indicará si existe o no regresión. Para elegir el coeficiente de correlación adecuado es necesario analizar el tipo de variables y la distribución que presentan.

Luego se realiza un análisis de normalidad representado en una gráfica con la función par() e hist().

Se obtiene la normalidad con qq(norm) y se traza la línea con qq(line). El test de hipótesis se puede obtener mediante la función shapiro.test(). El análisis gráfico y el contraste de normalidad nos muestran si se puede o no asumir normalidad.

Para la significancia de la correlación usaremos la función cor.test() que como antes mencionamos nos indica un número de relación entre las variables. Luego obtenemos el coeficiente de determinación en R2 con cor().

Matriz de correlaciones

Cuando se disponen de múltiples variables y se quiere estudiar la relación entre todas ellas se recurre al cálculo de matrices con el coeficiente de correlación de cada par de variables (pairwise correlation). También se generan gráficos de

dispersión dos a dos. En R existen diferentes funciones que permiten realizar este estudio, las diferencias entre ellas son el modo en que se representan gráficamente los resultados. Con (pairs) podemos analizar de dos en dos la correlación entre las variables.

Para poder realizar un histograma de las variables y poder analizar su dispersión una respecto a la otra utilizamos el paquete **psych** y la función (multi.hist) esto nos dará una vista gráfica y nos ayudará a entender de mejor forma la relación entre las variables.

```
>
> 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
>
```

Regresión lineal simple

La regresión lineal consiste en generar un modelo de regresión que permita explicar la relación lineal entre las variables. La variable dependiente se le identifica como X y la **variable predictora** se identifica como X.

De esta forma podremos identificar si existe alguna relación entre dos variables a partir de la relación que se observa en la muestra y por lo tanto está sujeta a variaciones. Para cada uno de los parámetros de la ecuación de regresión lineal se puede calcular su significancia.

Se realiza un test de hipótesis entre ambas variables siendo Ho la variable que indica que no hay relación lineal entre las variables por lo que la pendiente del modelo es 0 y la Ha que indica que si hay relación lineal entre ambas variables por lo que la pendiente del modelo lineal es distinta de 0.

Calculamos nuestro RSE que puede entenderse como la diferencia promedio que se desvía de la variable respuesta de la verdadera regresión lineal.

También obtenemos nuestros intervalos de confianza la cual nos indica la estimación de nuestros parámetros y si son significativamente distintos de 0. Si se desea conocer los intervalos de confianza para cada uno de los parámetros se puede calcular con la función coint().

Los residuos del modelo son una estimación que se define como la diferente entre el valor observado y el valor esperado acorde al modelo. Una vez que se ha ajustado el modelo es necesario verificar su eficiencia, ya que aun siendo la línea que mejor se ajusta a las observaciones de entre las posibilidades, el modelo puede ser malo. Las medidas más utilizadas para medir la calidad del ajuste son: error estándar de

los residuos, el test F y el coeficiente de determinación en R2. Este mide la desviación promedio de cualquier punto estimado por el modelo respecto de la verdadera recta de regresión poblacional. Luego tenemos el test F el cual es un test de hipótesis que considera la hipótesis nula que todos los coeficientes de correlación estimados son cero, frente a la hipótesis alternativa de que al menos uno de ellos no lo es.

Las condiciones para la regresión lineal son:
Linealidad: La relación entre ambas variables debe ser lineal. Para comprobarlo se puede recurrir a scatterplots o diagramas de dispersión. Calcular los residuos de cada observación acorde al modelo generado. Distribución normal de los residuos: Los residuos se tiene que distribuir de forma normal, con media igual a 0. Esto se puede comprobar con un histograma, con la distribución de cuantiles qqnorm y qqline o con un test de hipótesis de normalidad.

La varianza de residuos constante (homocedasticidad), la varianza de los residuos debe ser aproximadamente constante a lo largo del eje X. Los valores atípicos y de alta influencia deben ser estudiados ya que pueden generar falsa correlación que realmente no existe, u ocultar una existente. La independencia de autocorrelación indica que las variables deben ser dependientes una de otra.

Una vez generado un modelo que se pueda considerar válido es posible predecir el valor de la variable dependiente Y para nuestros valores de la variable predictora X. Es importante tener en cuenta que las predicciones deben limitarse al rango de valores dentro del que se encuentran las observaciones con las que se ha generado el modelo.

Dado que el modelo generado se ha obtenido a partir de una muestra, por lo tanto las estimaciones de los coeficientes de regresión tienen un error asociado.

Selección de algoritmo

Regresión lineal

Para estudiar la relación lineal existente entre dos variables continuas es necesario disponer de parámetros que permitan cuantificar dicha relación. Uno de estos parámetros es la *covarianza*, que indica el grado de variación conjunta de dos variables aleatorias.

$$Covarianza\ muestral = Cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{N-1}$$

Las comparaciones que se pueden realizar por medio de este algoritmo se mencionan las de pearson, spearman y kendall. Con el número de covarianza muestral se identifica la correlación haciendo que sea + o -, y se mantiene en el rango entre 0 y 1.

Resultados

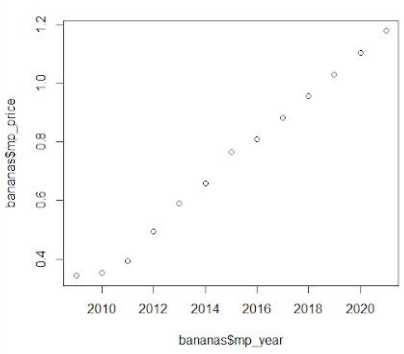
En la investigación el objetivo es usar algoritmos para predecir el precio de los alimentos, lo cual como anteriormente se menciona se hará por medio de la regresión lineal simple. Pero se hará tomando 4 países en donde se incluye Guatemala por ser el país objetivo, los otros serán de america latina entre ellos Bolivia, Colombia y Perú. Sin embargo se tiene le problema que no todos los países tienen de todo tipo de comida por cual se escogen varios para poder referir cada item, como ejemplo se tienen las legumbres de Perú donde tienen verduras, frutas, etc. pero no legumbre por lo cual se usa a colombia para comparar los precios de Guatemala.

En los resultados previos se daban las predicciones con el precio en dólar, sin embargo se ve que el comportamiento es mayormente de la actividad del dólar que el precio de los al alimentos en sí. Esto produce incertidumbre y menos confiabilidad.

Gráficas con precio

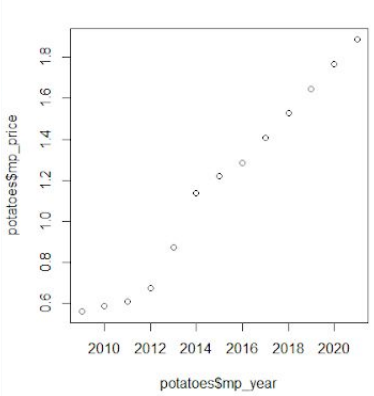
Gráficos de Guatemala

Precio de bananas



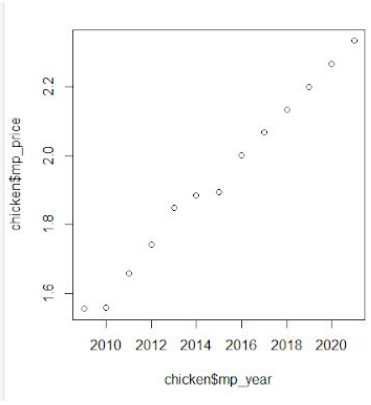
Podemos observar como el precio de los bananos ha ido incrementando desde el 2010 y seguirá creciendo.

Precio de papas



El precio de las papas incrementó muy poco del 2010 al 2012; del 2014 al 2015 hubo un gran incremento y se predice que desde esa fecha hasta el 2020 el precio irá ganando un precio similar cada año.

Precio de pollo

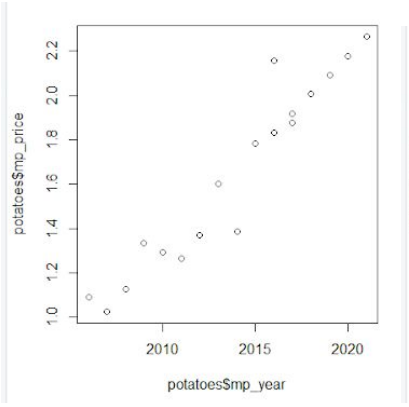


El pollo mantuvo su precio del 2009 al 2010, luego tiene un gran incremento de aproximadamente $0.4 llegando aproximadamente a los $1.9. Luego el l precio crece poco a poco hasta llegar al 2020

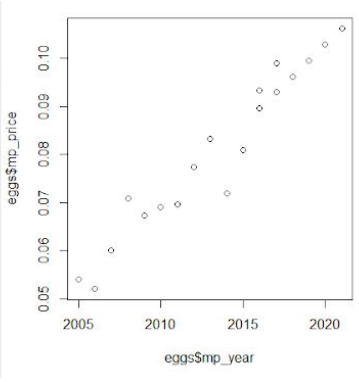
Gráficos de América del Sur

En estas gráficas se incluyeron 3 países de américa del sur debido a que no todos los países tenían datos de todos los tipo de comida, entre estos están Bolivia, Perú y Colombia.

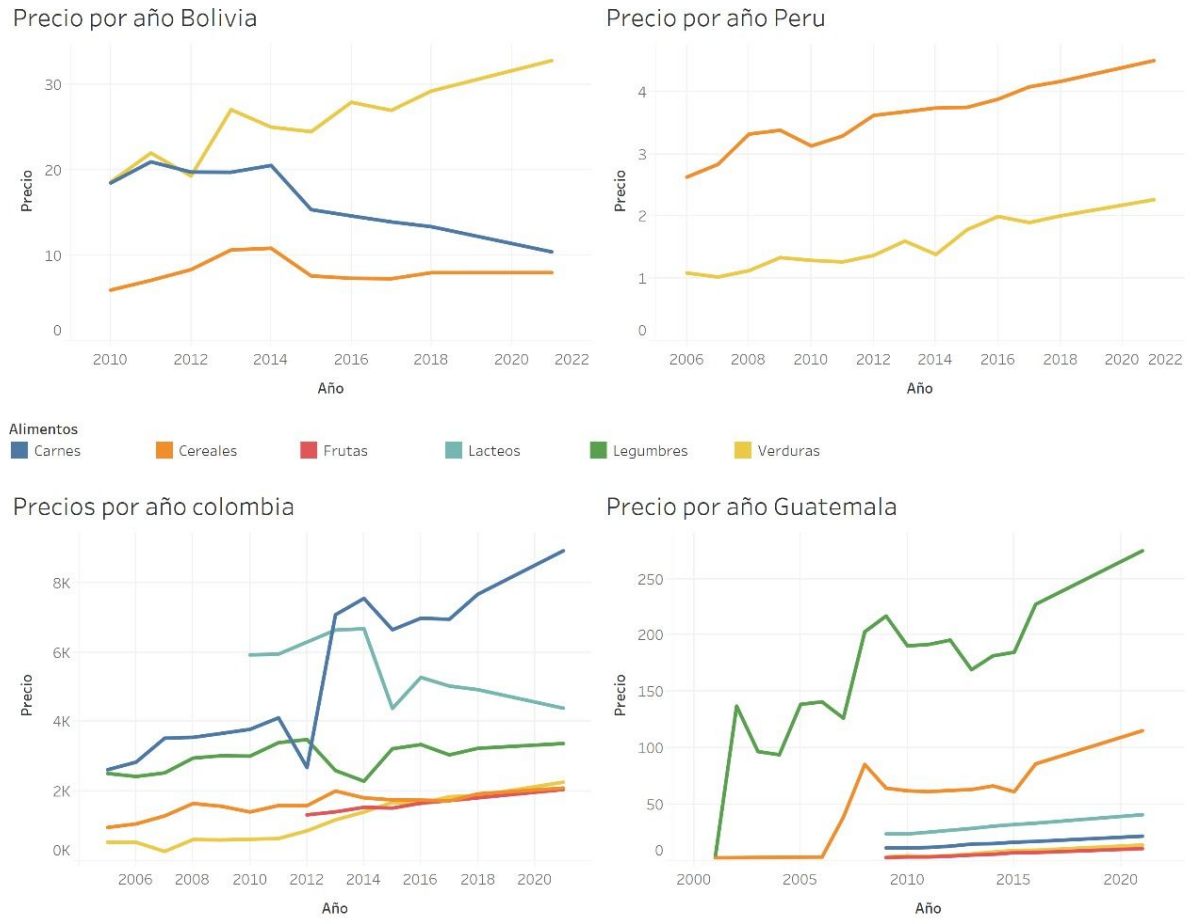
Papas



Huevos



Al comparar los precios que se predijo fueron alineados en grupos alimenticios y extendidos a cada país seleccionado haciendo de esto una predicción ascendente.



Conclusión

Se aprueba la hipótesis que el precio aumenta mediante el paso del tiempo sin importar el país del cual venga la información. Lo que pone en riesgo la nutrición de las personas Bolivia, Colombia, Perú y Guatemala.

Bibliografía:

Mahantesh, A, *et al.* (2015). Time series Data Analytics for stock market prediction using data mining techniques with R. Extraído de International Journal of advanced research in computer science https://www.researchgate.net/publication/286890497_Time_Series_Data_Analysis_for_Stock_Market_Prediction_using_Data_Mining_Techniques_with_R

Amat, J, et al. (2016) *Correlación lineal y regresión lineal simple*. RPubs. Extraído el 10 de octubre de 2018 de https://rpubs.com/Joaquin_AR/223351

Romero, D. (2016) *Predicting sale prices in R.* RPubs by RStudio. Extraído el 10 de octubre de 2018 de https://rpubs.com/darioromero/housesaleprice

FAO(2010) Los http://www.fao.org/docrep/013/am044s/am044s00.pdf