Preprocesamiento de datos

# Introducción

Agroindustria Hiui desea mejorar su gestión de inventarios. Para lo cual desea realizar una clasificación ABC multicriterio, cuya base de ponderación serán los pesos de procesos de jerarquía analítica.

No obstante, para poder realizar un proyecto de mejora en los inventarios, es fundamental analizar el comportamiento de los datos, además de validar que los datos sean de calidad para su procesamiento.

El gerente de TI le brinda la base de datos, “data\_inventory.csv”, que presenta en su diseño 7 atributos y contiene 133361 registros de SKUs.

✓ Numero\_SKU: identificador del tipo de producto.  
✓ Tipo\_archivo: indica si el registro es histórico, Historical, o si el producto esta presente en el almacén, Active.  
✓ Tipo\_marketing: la empresa posee dos tipos de promoción de ventas para incrementar la rotación de inventario; tipo D y tipo S.  
✓ Demanda: se registra la demanda estimada del SKU durante el 2022 (unidades de producto).  
✓ Precio: Precio en soles por unidad del del tipo del .  
✓ Lead\_time: tiempo promedio expresados en días que Agroindustrias Hiu demora en abastecerse de los pertenecientes al .  
✓ Estabilidad\_suministro: Dificultad de abastecimiento del del tipo del . durante nuestro escenario postpandemia.

# Carga de datos

data\_inventory <-   
 read\_csv(file = "data\_inventory.csv",  
 show\_col\_types = F)  
  
data\_inventory %>% glimpse()

## Rows: 133,360  
## Columns: 7  
## $ Numero\_SKU <dbl> 1737127, 3255963, 612701, 115883, 863939, 21494…  
## $ Tipo\_archivo <chr> "Historical", "Historical", "Historical", "Hist…  
## $ Tipo\_marketing <chr> "D", "D", "D", "D", "D", "D", "D", "D", "D", "D…  
## $ Demanda <dbl> 682743, 1016014, 340464, 334011, 1287938, 17831…  
## $ Precio <dbl> 167.1, 42.8, 85.2, 80.8, 138.9, 146.4, 188.2, 1…  
## $ Lead\_time <dbl> 12, 11, 10, 10, 12, 12, 7, 9, 8, 11, 6, 4, 10, …  
## $ Estabilidad\_suministro <dbl> 3.9, 2.7, 8.2, 11.7, 6.8, 14.8, 9.1, 7.3, 14.1,…

# Análisis de datos faltantes

Al revisar los datos usted inmediatamente se percata que aun no esta limpia la base de datos, por lo cual se le solicita realizar un informe con los siguientes puntos claves:

## a) Analice la proporción de datos faltantes por tipo de atributo. Realice comentarios acerca de esta problemática.

data\_inventory %>%   
 skim() %>%   
 select(skim\_variable, n\_missing, complete\_rate) %>%   
 mutate(incomplete\_rate = round((1 - complete\_rate) \* 100, 2)) %>%   
 arrange(-incomplete\_rate) %>%   
 select(-complete\_rate)

## # A tibble: 7 × 3  
## skim\_variable n\_missing incomplete\_rate  
## <chr> <int> <dbl>  
## 1 Precio 2765 2.07  
## 2 Estabilidad\_suministro 2709 2.03  
## 3 Lead\_time 2700 2.02  
## 4 Tipo\_archivo 0 0   
## 5 Tipo\_marketing 0 0   
## 6 Numero\_SKU 0 0   
## 7 Demanda 0 0

El atributo precio presenta un 2.07 % de datos faltantes, mientras que la estabilidad de suministro presenta 2.03 % y por ultimo el lead time, presenta un 2.02 %.

## b) Inputar los datos faltantes por métodos univariables y métodos multivariables. Explique las ventajas y desventajas de cada método aplicada a la data de inventarios.

Usualmente cuando el porcentaje de datos faltantes no es alto, como regla menor a 1 %, se puede prescindir de los mismos, a pesar de que se reconoce que esta práctica no es la más apropiada, ya que genera sesgos en los coeficientes de asociación y de correlación (Kalton y Kasprzyk, 1982).

### Imputación por la media

data\_inventory\_media <-   
 data\_inventory %>%   
 impute\_mean\_at(.vars = c("Precio", "Estabilidad\_suministro", "Lead\_time"))

### Imputación por la mediana

data\_inventory\_mediana <-   
 data\_inventory %>%   
 impute\_median\_at(.vars = c("Precio", "Estabilidad\_suministro", "Lead\_time"))

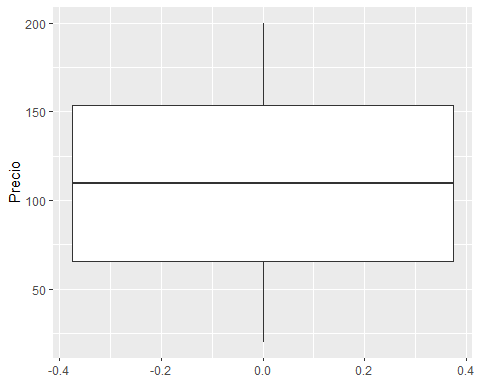
## c) Realizar el análisis de outliers por métodos univariables y métodos multivariables. Explique las ventajas y desventajas de cada método aplicada a la data de inventarios.

### Análisis por el diagrama de cajas

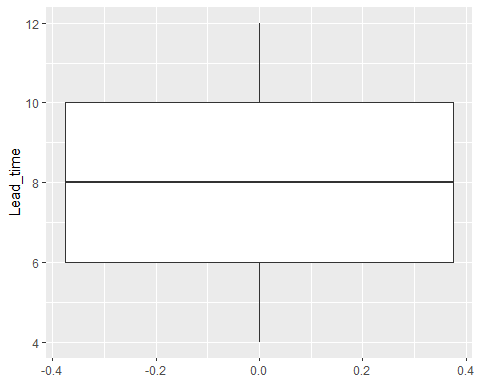
data\_inventory\_media %>%   
 ggplot(mapping = aes(y = Demanda)) +   
 geom\_boxplot()



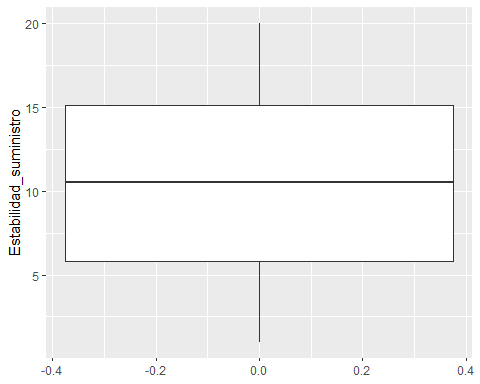
data\_inventory\_media %>%   
 ggplot(mapping = aes(y = Precio)) +   
 geom\_boxplot()



data\_inventory\_media %>%   
 ggplot(mapping = aes(y = Lead\_time)) +   
 geom\_boxplot()



data\_inventory\_media %>%   
 ggplot(mapping = aes(y = Estabilidad\_suministro)) +   
 geom\_boxplot()



## d) Realizar un análisis descriptivo de cada variable de la base preprocesada. Interprete.

### Atributo Lead Time

data\_inventory\_media %>%   
 group\_by(Tipo\_marketing, Tipo\_archivo) %>%   
 summarise(min = min(Lead\_time),  
 p25 = quantile(Lead\_time, probs = 0.25),  
 media = round(mean(Lead\_time), 2),  
 mediana = median(Lead\_time),  
 p75 = quantile(Lead\_time, probs = 0.75),  
 max = max(Lead\_time),  
 sd = round(sd(Lead\_time), 2))

## `summarise()` has grouped output by 'Tipo\_marketing'. You can override using  
## the `.groups` argument.

## # A tibble: 4 × 9  
## # Groups: Tipo\_marketing [2]  
## Tipo\_marketing Tipo\_archivo min p25 media mediana p75 max sd  
## <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 D Active 4 6 7.98 8 10 12 2.54  
## 2 D Historical 4 6 8.01 8 10 12 2.55  
## 3 S Active 4 6 8.01 8 10 12 2.55  
## 4 S Historical 4 6 8.02 8 10 12 2.56

* El Lead Time mínimo de productos activos ofertados por el marketing D, presentan un tiempo promedio de abastecimiento mínimo de 4 días, esta estadística se repite para todos los productos activos o históricos según cada tipo de marketing.
* El 25% de los productos tiene un Lead Time de productos activos ofertados por el marketing D, presentan un tiempo promedio de abastecimiento a lo más 6 días, esta estadística se repite para todos los productos activos o históricos según cada tipo de marketing.
* El Lead Time promedio de productos activos ofertados por el marketing D, presentan un tiempo promedio de abastecimiento 7.98 días. Por otro lado, los productos históricos ofertados por el marketing D, presentan un tiempo promedio de abastecimiento 8.01 días. Esta misma estadística se repite para los productos activos ofertados por el marketing S. Por otro lado, los productos históricos ofertados por el marketing S, presentan un tiempo promedio de abastecimiento 8.02 días.
* La mediana de los productos tiene un Lead Time de productos activos ofertados por el marketing D, presentan una mediana de tiempo de abastecimiento de 8 días, esta estadística se repite para todos los productos activos o históricos según cada tipo de marketing.
* El 75% de los productos tiene un Lead Time de productos activos ofertados por el marketing D, presentan un tiempo promedio de abastecimiento a lo más 10 días, esta estadística se repite para todos los productos activos o históricos según cada tipo de marketing.
* El Lead Time máximo de productos activos ofertados por el marketing D, presentan un tiempo promedio de abastecimiento máximo de 12 días, esta estadística se repite para todos los productos activos o históricos según cada tipo de marketing.

### Atributo Demanda

data\_inventory\_media %>%   
 group\_by(Tipo\_marketing, Tipo\_archivo) %>%   
 summarise(min = min(Demanda),  
 p25 = quantile(Demanda, probs = 0.25),  
 media = round(mean(Demanda), 2),  
 mediana = median(Demanda),  
 p75 = quantile(Demanda, probs = 0.75),  
 max = max(Demanda),  
 sd = round(sd(Demanda), 2))

## `summarise()` has grouped output by 'Tipo\_marketing'. You can override using  
## the `.groups` argument.

## # A tibble: 4 × 9  
## # Groups: Tipo\_marketing [2]  
## Tipo\_marketing Tipo\_archivo min p25 media mediana p75 max sd  
## <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 D Active 2517. 9.16e5 2.38e6 1.82e6 3.14e6 1.73e7 2.10e6  
## 2 D Historical 284 2.83e5 1.49e6 8.76e5 2.00e6 1.67e7 1.80e6  
## 3 S Active 6.28 5.29e4 2.30e5 1.14e5 3.07e5 1.74e7 3.45e5  
## 4 S Historical 68 2.18e5 9.96e5 6.19e5 1.27e6 1.53e7 1.23e6

* La demanda mínima de productos activos ofertados por el marketing D, presentan un tiempo promedio de abastecimiento de 2,517.457 unidades, siendo la mas alta registrada. Por otro lado, la demanda mínima de productos activos ofertados por el marketing S, presentan un tiempo promedio de abastecimiento mínimo de 6.275 unidades, siendo esta la mas baja.
* El 25% de los productos tienen una demanda de productos activos ofertados por el marketing D, presentan una demanda 916,390.10 unidades, siendo esta la mas alta registrada. Por otro lado, el 25% de los productos tienen una demanda de productos activos ofertados por el marketing S, presentan un tiempo promedio de abastecimiento mínimo de 52,927.65 unidades, siendo esta la mas baja.
* La demanda media de productos activos ofertados por el marketing D, presentan un tiempo promedio de abastecimiento mínimo de 2,376,512.4 unidades, siendo la mas alta registrada. Por otro lado, la demanda media de productos activos ofertados por el marketing S, presentan un tiempo promedio de abastecimiento mínimo de 229,926.8 unidades, siendo esta la mas baja.
* La mediana de la demanda de productos activos ofertados por el marketing D, presentan un tiempo promedio de abastecimiento mínimo de 1,818,957.0 unidades, siendo la mas alta registrada. Por otro lado, la mediana de la demanda de productos activos ofertados por el marketing S, presentan un tiempo promedio de abastecimiento mínimo de 113,754.1 unidades, siendo esta la mas baja.
* El 75% de los productos tienen una demanda de productos activos ofertados por el marketing D, presentan una demanda 3,141,345.0 unidades, siendo esta la mas alta registrada. Por otro lado, el 75% de los productos tienen una demanda de productos activos ofertados por el marketing S, presentan un tiempo promedio de abastecimiento mínimo de 307,478.5 unidades, siendo esta la mas baja.
* La demanda máxima de productos activos ofertados por el marketing S, presentan un tiempo promedio de abastecimiento de 17,384,454 unidades, siendo la mas alta registrada. Por otro lado, la demanda máxima de productos históricos ofertados por el marketing S, presentan un tiempo promedio de abastecimiento mínimo de 15,318,509 unidades, siendo esta la mas baja.

### Atributo Precio

data\_inventory\_media %>%   
 group\_by(Tipo\_marketing, Tipo\_archivo) %>%   
 summarise(min = min(Precio),  
 p25 = quantile(Precio, probs = 0.25),  
 media = round(mean(Precio), 2),  
 mediana = median(Precio),  
 p75 = quantile(Precio, probs = 0.75),  
 max = max(Precio),  
 sd = round(sd(Precio), 2))

## `summarise()` has grouped output by 'Tipo\_marketing'. You can override using  
## the `.groups` argument.

## # A tibble: 4 × 9  
## # Groups: Tipo\_marketing [2]  
## Tipo\_marketing Tipo\_archivo min p25 media mediana p75 max sd  
## <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 D Active 20 65.4 110. 110. 154. 200 51.5  
## 2 D Historical 20 66 110. 110. 154. 200 51.4  
## 3 S Active 20 66.1 110. 110. 154. 200 51.3  
## 4 S Historical 20 65.7 110. 110. 154. 200 51.5

* El precio mínimo de productos activos ofertados por el marketing D, presentan un tiempo promedio de abastecimiento de 20 unidades, siendo esta la misma para el resto de productos.
* El 25% de los productos tienen un precio de productos activos ofertados por el marketing S, presentan una demanda 66.1 unidades, siendo esta la mas alta registrada. Por otro lado, el 25% de los productos tienen una demanda de productos activos ofertados por el marketing D, presentan un tiempo promedio de abastecimiento mínimo de 65.4 unidades, siendo esta la mas baja.
* El precio medio de productos histórico ofertados por el marketing S, presentan un tiempo promedio de abastecimiento de 110.04 unidades, siendo la mas alta registrada. Por otro lado, el precio medio de productos activos ofertados por el marketing D, presentan un precio promedio de abastecimiento de 109.66 unidades, siendo esta la mas baja.

## e) Después de una semana que usted realiza el análisis de datos faltantes, le brindan la data que se recuperó del sistema, data\_completa. ¿su imputación fue de calidad? Sustente.

data\_completa <-   
 read\_excel(path = "data\_completa.xlsx")  
  
precio\_na <-   
 data\_inventory %>%   
 filter(is.na(Precio)) %>%   
 select(Numero\_SKU) %>%   
 as.vector()  
  
data\_completa\_precio <-   
 data\_completa %>%   
 select(Numero\_SKU, Precio) %>%   
 filter(Numero\_SKU %in% precio\_na$Numero\_SKU)  
  
data\_error <-   
 data\_completa\_precio %>%   
 left\_join(data\_inventory\_media, by = c("Numero\_SKU" = "Numero\_SKU"))  
  
data\_error <-   
 data\_error %>%   
 mutate(dif = Precio.x - Precio.y)  
  
summary(data\_error$dif)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -89.8039 -45.5039 0.8961 0.2816 43.6961 90.1961