Solución Taller No. 1 - Análisis de datos con R

Mauricio Mejía Castro

23 de febrero de 2021

1. Medidas de tendencia central

- Operador *pipeline* (`%>%`): Ofrece una sintaxis alternativa a la anidación usual de operaciones `dplyr`. En esencia, permite construir de manera más natural una sucesión de operaciones.
- Función `group_by()`: Operación del paquete `dplyr` que toma una tabla como entrada y retorna una tabla agrupada con respecto a alguna variable.
- Función `summarise_if()`: Este operador ha sido reemplazado por el uso de `across`, según la documentación de R. Su principal uso es el de aplicar una transformación a múltiples columnas de una tabla.
- Función `is.numeric()`: Retorna `TRUE` si el tipo de dato es un entero o entero con doble precisión.

El siguiente código utiliza estas funciones para resumir por medio de la media y la mediana los datos del Apéndice B con respecto a la columna `SamplePosition`.

```
soil_dat_path <- paste(main_path, "appendix-b.csv", sep = "")
soil_dat <- read.csv(soil_dat_path)
soil_summary <- soil_dat %>%
group_by(SamplePosition) %>%
summarise(across(is.numeric, list(Media=mean, Median=median)))
```

2. Medidas de dispersión

El mayor nivel de variación se presenta en los sistemas de producción a campo abierto. Esto lo podemos determinar con el resultado del código adjunto.

Variable	Boyaca	Santander
Longitude.DD_cvar	-0.04479261	-0.06543138
Latitude.DD_cvar	0.8575635	1.7145951
Altitude.masl_cvar	6.672738	15.533085
LotArea.m2_cvar	44.11184	86.60027
CycleDuration.days_cvar	16.30292	10.75022
Density.plants.m2_cvar	16.64172	27.21224
Yield.kg.m2_cvar	45.67017	52.89135
IrrigVolumen.l.m2_cvar	65.45662	141.56878
Rainfall.mm_cvar	NA	41.19937
TotalN.kg.m2_cvar	56.79701	72.86935
P2O5.kg.m2_cvar	60.75209	85.55578
K2O.kg.m2_cvar	63.15631	76.57278

Cuadro 1: Coeficientes de variación agrupados por departamento

```
tomato_summary.1 <- tomato_dat %>%
group_by(Department) %>%
summarise(across(is.numeric, list(cvar=cv)))
```

El Cuadro 1 resume el cálculo del coeficiente de variación para las variables numéricas agrupadas por la variable `Department`:

Para conocer a qué instrumento le corresponde el mayor nivel de variación, agrupamos por la variable `DataSource` a través del código adjunto:

```
tomato_summary.2 <- tomato_dat %>%
group_by(DataSource) %>%
summarise(across(is.numeric, list(cvar=cv)))
```

El resumen obtenido en el Cuadro 2 sugiere que el método de medición con mayor nivel de variación son los seguimientos.

La única variable en la que se evidencia un coeficiente de variación negativo es en la variable `Longitude.DD`.

3. Gráficas para medidas de tendencia central y dispersión

La Figura 1 muestra la gráfica generada por el código dado en el enunciado.

DataSource	FollowUps	Survey
Longitude.DD_cvar	-0.2897137	-0.3014357
Latitude.DD_cvar	6.276944	6.119214
Altitude.masl_cvar	21.15645	18.38336
LotArea.m2_cvar	100.59032	79.97274
CycleDuration.days_cvar	21.48421	25.88937
Density.plants.m2_cvar	42.57766	38.58078
Yield.kg.m2_cvar	45.50939	61.53591
IrrigVolumen.l.m2_cvar	93.33036	83.69995
Rainfall.mm_cvar	144.8402	140.7425
TotalN.kg.m2_cvar	69.18710	73.30321
P2O5.kg.m2_cvar	87.76013	73.37628
K2O.kg.m2_cvar	84.20972	83.40776

Cuadro 2: Coeficientes de variación agrupados por tipo de medición

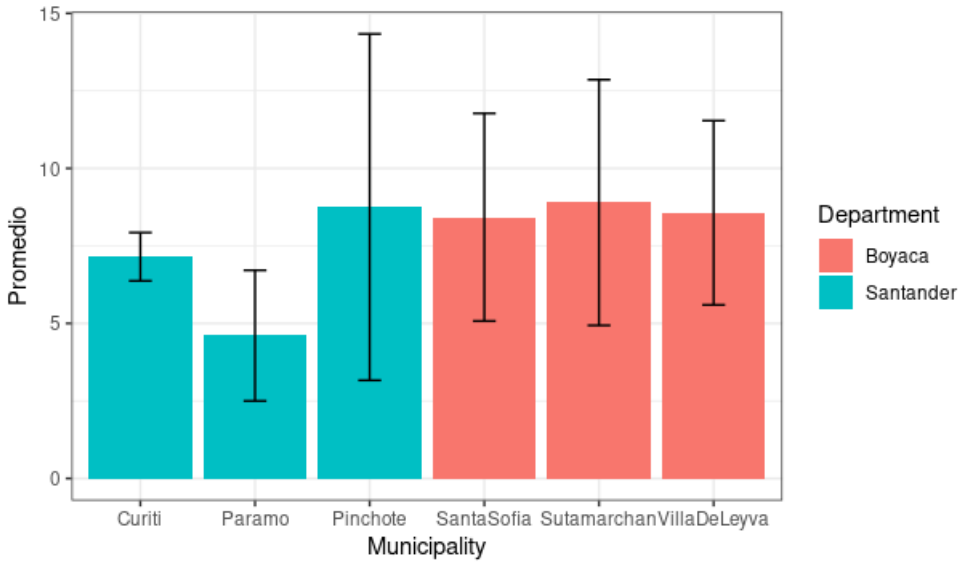

Figura 1: Comparación del promedio de rendimiento por municipio

- Función `ggplot()`: Inicializa un objeto `ggplot`. Este puede ser usado para empezar el marco inicial sobre el cual puede ser graficado algún conjunto de datos.

- Función `aes()`: Proporciona la lista de elementos estéticos que se utilizarán en la gráfica.
- Función `theme_bw()`: Proporciona temas visuales completos para la gráfica.
- Función `geom_bar()`: Indica al objeto que la gráfica debe ser de barras.
- Función `geom_errorbar()`: Agrega un gráfico de intervalo vertical a la gráfica de barras.

4. Medidas de posición relativa

En la Gráfica 2 se muestra el resultado del código dado.

Si nos detenemos en el nivel de acidez (`pH`), es válido mencionar que existe mayor dispersión de esta para los sistemas bajo invernadero (Boyacá) que para los de campo abierto (Santander). Se evidencian más valores atípicos en sistemas de campo abierto que en los de bajo invernadero. Se observa también que en ambos sistemas, el nivel de acidez exhibe asimetría negativa.

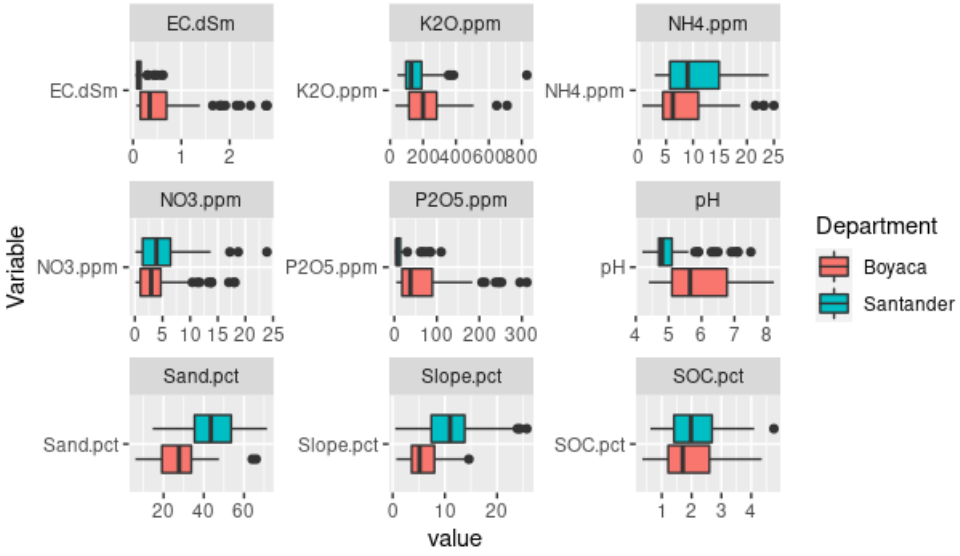

Figura 2: Comportamiento de algunas variables en el conjunto de datos