Modelos probabilísticos y análisis estadístico / Métodos estadísticos para data analytics

Primner taller de análisis de datos en R

- **Propósito:** Emplear el lenguaje de programación R para describir por medio de medidas de tendencia central, dispersión y distribucionalidad para variables de un conjunto de datos.
- **Datos:** Se emplearán los datos disponibles en el artículo: Gil, R., Bojacá, C. R., Schrevens, E. (2019). Datasets of the environmental factors and management practices of the smallholder tomato production systems in the Colombian Andes. Data in brief, 24, 103844.
- **Herramientas:** Se emplearán paquetes avanzados de R, específicamente el conjunto de paquetes contenidos dentro de la colección tidyverse, el cual está diseñado específicamente para científicos de datos. Aunque este tipo de herramientas no fueron presentadas en la clase, justamente este taller tiene como propósito introducir estas herramientas, de manera que la próxima clase se desarrolle a un ritmo más rápido.

Puntos a desarrollar

1. **Medidas de tendencia central**. Los datos que se emplearán serán los contenidos en el archivo del apéndice C, disponibles en el artículo relacionado anteriormente. En este archivo se presentan algunas características relacionadas con la producción de tomate bajo dos sistemas: campo abierto (Santander) y bajo invernadero (Boyacá), esto está especificado en la columna Department. También se discrimina de acuerdo con dos estrategias de captura de la información: encuestas a productores (Surveys) y seguimientos detallados a ciclos de producción (Follow-ups), en este caso especificado en la columna DataSource.

   Realizaremos un resumen de las variables numéricas por medio del cálculo de la media y la mediana para cada sistema de producción y método de recolección de la información. Ver código en R.

   Indague acerca de las siguientes funciones o instrucciones en R:

   - % > %
   - group_by
   - summarise_if
   - is.numeric

   Aplique este mismo enfoque para describir los datos del apéndice B, en el cual se describen las características de suelos dentro y fuera de invernaderos, es decir la variable SamplePosition

2. **Medidas de dispersión**. Ahora usaremos los datos del apéndice C para describir el nivel de variación al interior del conjunto de datos, para poder establecer una comparación entre las diferentes variables lo haremos por medio del coeficiente de variación. Ver código en R.

   Responda las siguientes preguntas

   - En términos generales, donde es mayor el nivel de variación: ¿En sistemas de producción a campo abierto o bajo invernadero?
   - Con qué instrumento de medición se obtuvo el mayor nivel de variación: ¿Con las encuestas o con los seguimientos detallados?
   - ¿En qué casos el coeficiente de variación es negativo?

3. **Gráfica uniendo las medidas de tendencia central y de dispersión**. Elaboremos una gráfica en la cual podamos comparar el promedio del rendimiento en cada uno de los municipios, pero teniendo en cuenta las siguientes consideraciones:

   - Sólo trabajaremos con los datos de rendimiento que provienen de los seguimientos.
   - Excluiremos los datos del municipio de Valle de San José por tener muy pocas observaciones.
   - La medida de tendencia central que usaremos será la media y la de dispersión será la desviación estándar.

   Ver código en R. Indague acerca de las siguientes funciones o instrucciones en R:

   - ggplot
   - aes
   - theme_bw
   - geom_bar
   - geom_errorbar

4. **Medidas de posición relativa**. Una representación típica de la distribución o agrupación de los datos es por medio de los cuartiles. El diagrama de cajas y bigotes es una representación gráfica de los cuartiles útil que se emplea para describir la distribución de los datos y para establecer comparaciones entre categorías. A continuación, un conjunto de criterios para la interpretación básica de un diagrama de cajas y bigotes.

   - Mientras más larga la caja y los bigotes, mayor dispersión en el conjunto de datos.
   - La distancia entre las cinco medidas descritas en el boxplot (sin incluir la media aritmética) puede variar; sin embargo, recuerde que la cantidad de elementos entre una y otra es aproximadamente la misma. Se considera aproximado porque pudiera haber valores atípicos, en cuyo caso la cantidad de elementos se ve levemente modificada.
   - La línea que divide la caja representa la mediana e indica la simetría. Si está relativamente en el centro de la caja la distribución es simétrica. Si, por el contrario, se acerca al primer o tercer cuartil, la distribución pudiera ser sesgada a la derecha (asimétrica positiva) o sesgada a la izquierda (asimétrica negativa) respectivamente.

- Los puntos por encima o por debajo del límite de los bigotes representan datos con valores atípicos. Un valor atípico o inusual (outlier, en inglés) corresponde a una observación que presenta un valor distante con respecto al resto de los datos.

Hagamos un diagrama de cajas y bigotes para describir la manera como se distribuye o agrupan los datos. En esta ocasión empleemos los datos del anexo A. Estos datos corresponden a los resultados de análisis de suelos realizados en las dos áreas productoras de tomate: Santander y Boyacá. En este caso estamos interesados en describir el comportamiento de las variables: acidez (pH), conductividad eléctrica (EC.dSm) contenido de nitrógeno amoniacal (NH4.ppm) y nítrico (NO3.ppm), potasio (K2O.ppm), fósforo (P2O5.ppm), carbono orgánico (SOC.pct) y contenido de arenas (Sand.pct); por departamento o zona de producción. Ver código en R.

Eliga una de las variables representadas en la gráfica y haga un análisis comparativo