Pre-Parcial

Analítica de Datos
Pontificia Universidad Javeriana

El conjunto de datos para este ejercicio proviene de la base de datos Coffee Quality y fue obtenido del repositorio de GitHub de TidyTuesday. Contiene datos detallados de más de 1000 variedades de café, incluyendo su procedencia, productor, características específicas y evaluación de calidad.

Realicen los siguientes ejercicios en orden, siguiendo cada paso cuidadosamente. Incluyan nombres en los ejes y un título informativo para todos los gráficos. Escriban todas las interpretaciones en el contexto de los datos.

1. Parte Descriptiva

1. Carguen los datos 'coffee_ratings.csv' a R.
2. Eliminen las observaciones para las cuales total_cup_points==0
3. Grafiquen la distribución de la variable dependiente *total_cup_points* y calculen las estadísticas descriptivas apropiadas. Usen la gráfica y las estadísticas descriptivas para describir la distribución. Incluyan un título informativo y etiquetas en los ejes del gráfico.
4. Hagan lo mismo del literal anterior para la variable *aroma*.
5. Elaboren un diagrama de dispersión que represente la relación entre las variables *total_cup_points* y *aroma*. Además, calculen el coeficiente de correlación entre ambas variables. Basándose en estos resultados, analice e interprete la relación entre estas dos características del café.
6. Usando la función **corrplot** (diapositivas 6 y 7 que están en este link), calculen la correlación entre las variables *aroma*, *flavor*, *aftertaste*, *acidity* y *body*. Sin analizar 1 por 1 las caracterízticas, ¿qué puede decir de la relación general de estas variables?

2. Parte Inferencial

7. Estimen la regresión lineal:

   $\widehat{\text{Total Cup Points}} = \hat{\beta}_0 + \hat{\beta}_1 \times Aroma$

   Interpreten la pendiente en el contexto de los datos.
8. ¿Tiene sentido interpretar el intercepto? Si es así, escriban la interpretación en el contexto de los datos.
9. ¿Tomarían una taza de café representada por el intercepto? Grafiquen la densidad de la variable dependiente y una línea punteada señalando la ubicación del intercepto en la gráfica.

10. Ahora evaluemos las condiciones del modelo. Verifiquen las condiciones de linealidad, varianza constante y normalidad. Para cada condición, indiquen si se cumple junto con una breve explicación de su conclusión. Incluyan los gráficos y/o estadísticas descriptivas que utilizaron para justificar su respuesta.

11. Ahora realicen la prueba de hipótesis para la pendiente. En su respuesta, enuncien las hipótesis nula y alternativa en palabras, y expongan la conclusión en el contexto de los datos.

12. Interpreten el R^2

13. Creen una variable dummy llamada *colombia* que sea igual a 1 si el país de origen es Colombia. ¿Qué porcentaje de los cafés en la muestra provienen de Colombia?

14. Estimen la regresión lineal:

    $\widehat{\text{Total Cup Points}} = \hat{\beta}_0 + \hat{\beta}_1 \times Aroma + \hat{\beta}_2 \times Colombia$

    Interpreten $\hat{\beta}_2$ en el contexto de los datos.

15. ¿Cómo cambia el R^2 con respecto al de la regresión en el punto 6?¿Qué nos dice esto sobre la variable *colombia*?

16. Estimen la regresión lineal:

    $\widehat{\text{Total Cup Points}} = \hat{\beta}_0 + \hat{\beta}_1 \times Flavor$

    Interpreten el R^2. Al comparar con el punto 11, ¿qué pueden decir sobre las variables *aroma* y $flavour$ y su relación con total_cup_points?

17. Imaginen que tienen que hablar con un caficultor que está en el dilema entre mejorar el aroma o el sabor del café que produce. Basado en estos datos y en la regresión:

    $\widehat{\text{Total Cup Points}} = \hat{\beta}_0 + \hat{\beta}_1 \times Aroma + \hat{\beta}_2 \times Flavor$

    ¿Qué le recomendarían?

18. Elijan una de las características del punto 6 y estimen la regresión:

    $\log(\widehat{\text{Total Cup Points}}) = \hat{\beta}_0 + \hat{\beta}_1 \times \log(\text{Característica})$

    Interpreten la pendiente en el contexto de los datos.