# Coffee Ratings

Carlos Cardona

### Preliminares

Los siguientes paquetes los usaré para resolver los ejercicios:

```
library(tidyverse)
library(readxl)
library(broom)
library(corrplot)
```

Adicionalmente voy a usar los siguientes paquetes. El paquete `patchwork` que me permite poner las gráficas, una al lado de la otra. Con el paquete `knitr` se pueden formatear las tablas para que sean más estéticas ( La documentación del paquete `patchwork` está en el siguiente link y para `knitr` en este link)

```
library(patchwork)
library(knitr)
```

Con estos paquetes agrego cosas nuevas a los códigos que ustedes ya conocían pero no es necesario que las integren. Simplemente es para que el documento salida se vea mejor. En los preliminares también deberían establecer el directorio de trabajo. En este documento no muestro el código por su longitud. Revisen el archivo `.qmd` para verlo.

## Parte Descriptiva

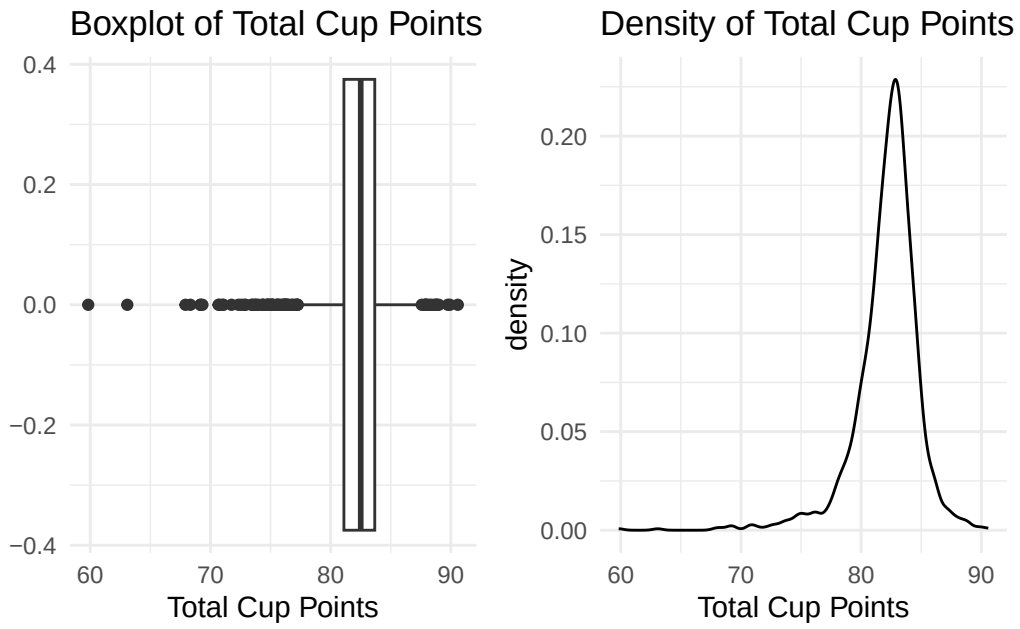
### Punto 1

```
coffee_data <- read.csv("coffee_ratings.csv")
```

## Punto 2

```
coffee_data <- coffee_data %>%
  filter(!total_cup_points==0)
```

## Punto 3



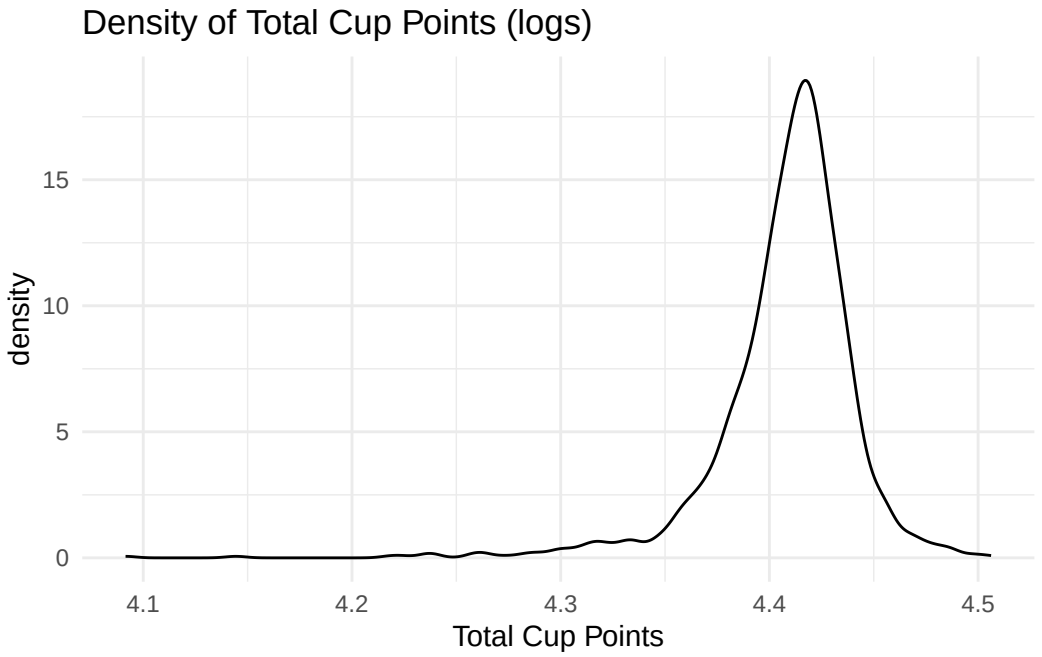

Las estadística descriptivas más desagregadas están a continuación:

```
summary(coffee_data$total_cup_points)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  59.83   81.10   82.50   82.15   83.67   90.58
```

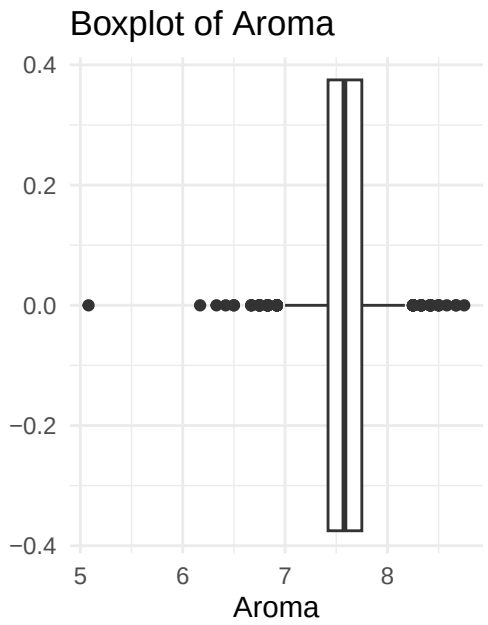
Como se puede observar en el diagrama de caja y la densidad, la distribución de `Total Cup Points` está sesgada hacia la izquierda dada la presencia de varios valores atípicos (outliers). El valor mínimo es 59 y el máximo 90. El valor que divide la distribución en la mitad es 82 puntos. La distribución está concentrada con un rango intercuartílico de aproximadamente 2 (83.67-81.10).

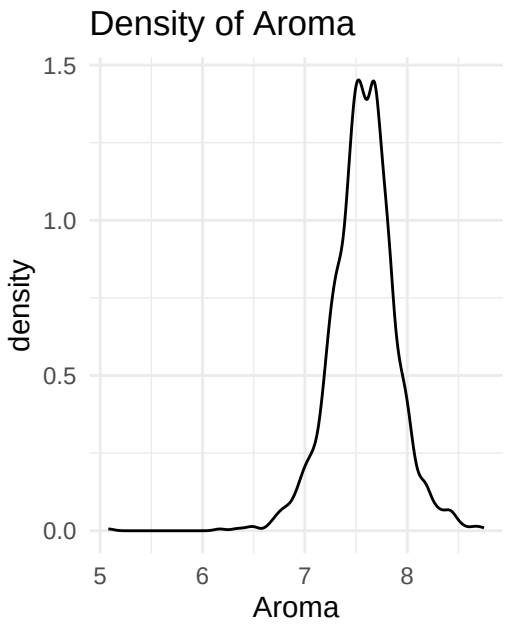
¿Cambiaría la distribución si le aplicamos un logaritmo? Según la gráfica de abajo, no habría necesidad ya que la distribución seguiría siendo similar. Probablemente la presencia de outliers en ambos lados de la distribución impiden que `Total Cup Points` tenga una distribución log-normal.




## Punto 4

Con la variable `Aroma` sucede algo similar que en el punto anterior. Aunque el diagrama de caja sugiere la presencia de valores atípicos a ambos lados, la distribución está más sesgada a la izquierda. Nuevamente la distribución está bastante concentrada alrededor de la mediana, con un rango intercuartílico menor a 1 (7.7-7.4).



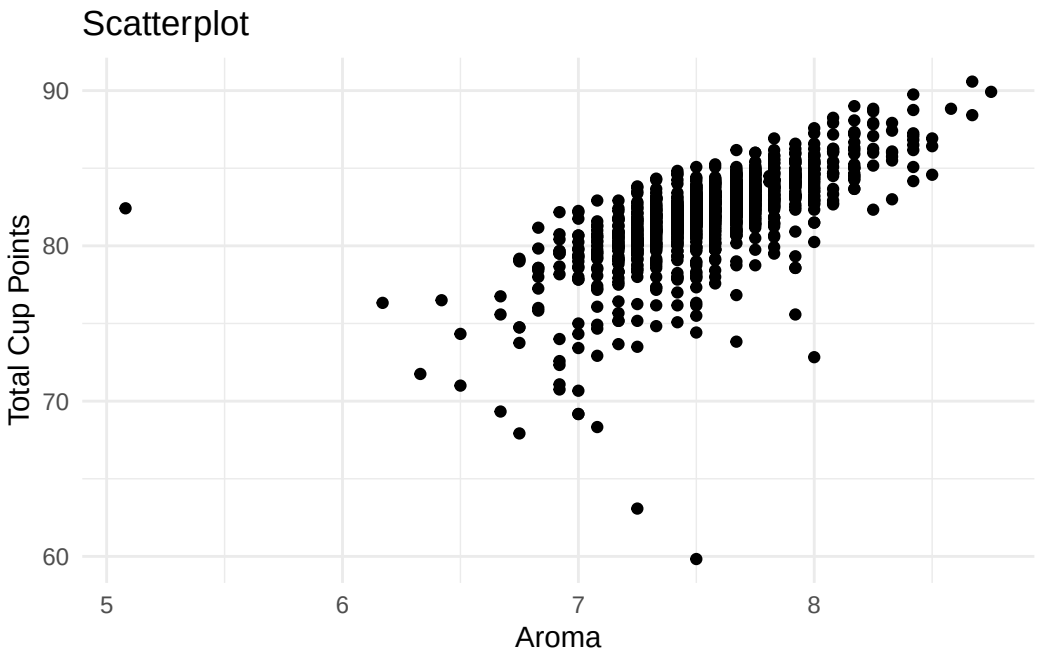




```
summary(coffee_data$aroma)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  5.080   7.420   7.580   7.572   7.750   8.750
```

### Punto 5

Tanto el diagrama de dispersión, como la correlación nos sugieren que hay una fuerte relación lineal entre ambas variables. Sin embargo, es importante resaltar que existen algunos casos que se alejan de la posible línea de regresión. Probablemente esto se conecte con los valores extremos que vimos para ambas variables en los puntos anteriores.



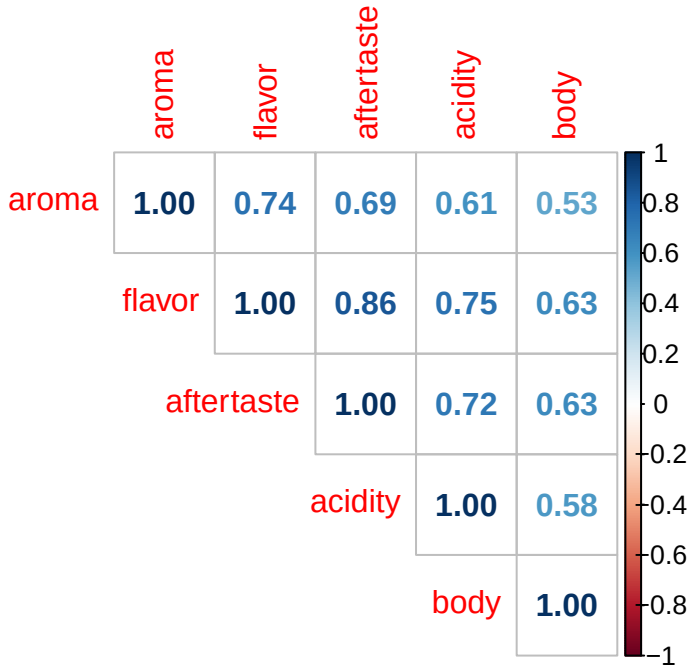

```
cor(coffee_data$total_cup_points, coffee_data$aroma)
```

```
[1] 0.6858045
```

### Punto 6

La primera conclusión "global" de la matriz de correlación es que todas tienden hacia el color azul oscuro. En efecto, al evaluar los valores numéricos podríamos concluir que todas las variables están fuertemente correlacionadas la una con la otra ya que todos los coeficientes de correlación son mayores a 0,5.

```
subset_coffee <- coffee_data %>%
  select(aroma, flavor, aftertaste, acidity, body)
```




## Parte Inferencial

### Puntos 7 y 8

```
coffee_model <- lm(total_cup_points ~ aroma, coffee_data)
tidy(coffee_model,conf.int = TRUE ) %>%
  knitr::kable(format = "latex", digits= 3)
```

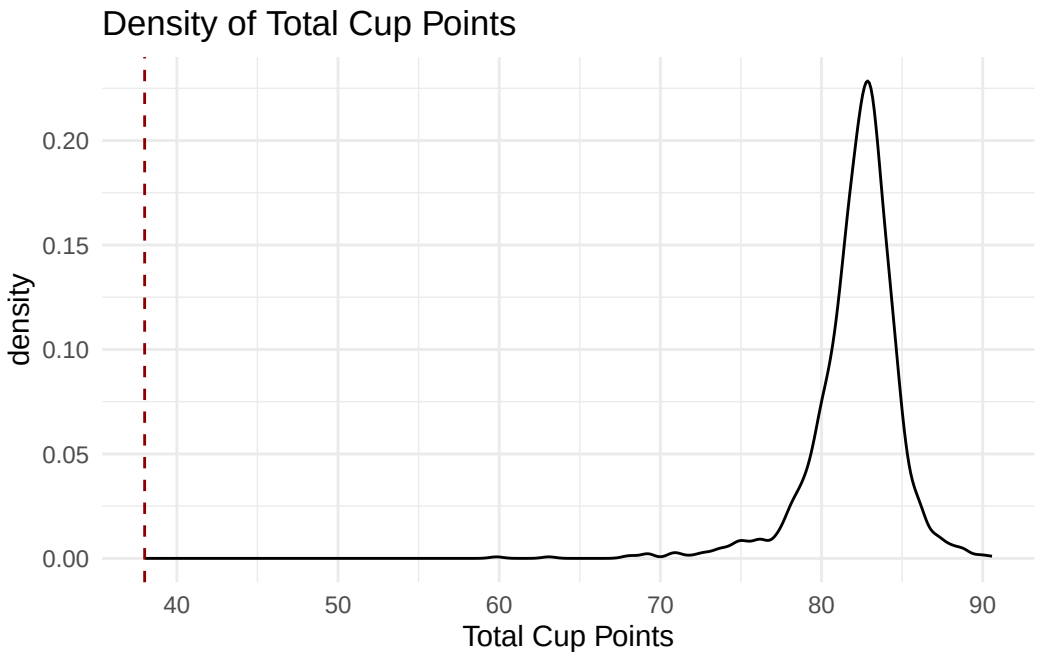
| term | estimate | std.error | statistic | p.value | conf.low | conf.high |
|---|---|---|---|---|---|---|
| (Intercept) | 37.983 | 1.283 | 29.594 | 0 | 35.466 | 40.501 |
| aroma | 5.833 | 0.169 | 34.443 | 0 | 5.501 | 6.165 |

En este caso tenemos un modelo de regresión simple donde ambas variables están en niveles. La interpretación del $\hat{\beta}_1$ en este caso sería que un aumento de 1 punto en la variable `Aroma` está asociado con un aumento de 5.83 puntos en `Total Cup Points`.

No tiene sentido interpretar el intercepto ya que, como vimos en el punto 4, no hay ningún café para el cual `Aroma` sea cero.

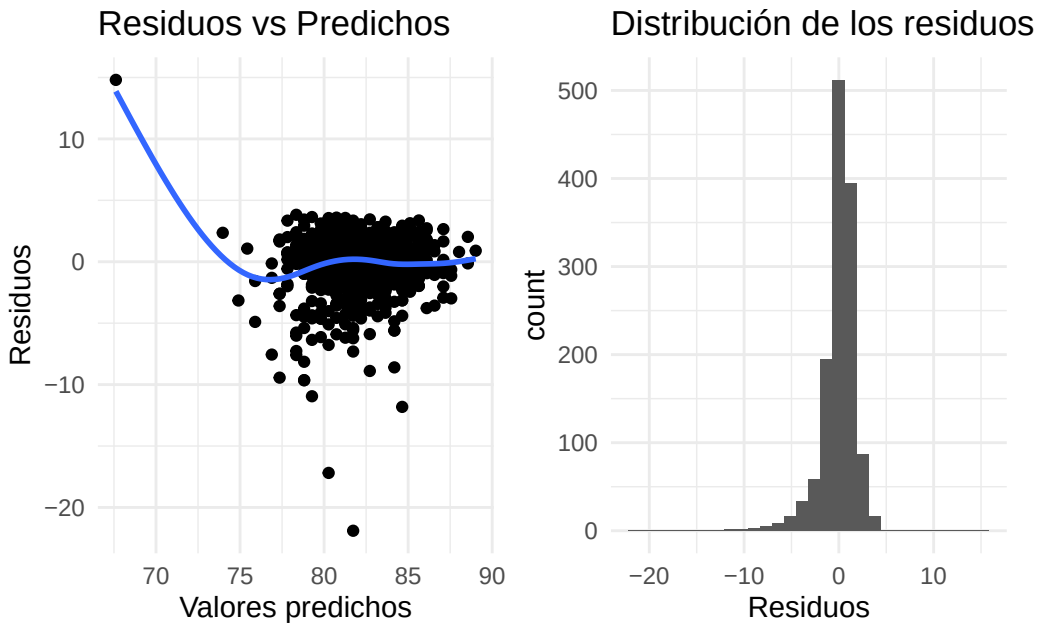
## Punto 9

Al graficar la variable `Total Cup Points` junto con una línea vertical en x=38 (el intercepto de la regresión en el punto 8), podemos ver que un café en esa parte de la distribución no sería la mejor elección.




## Punto 10

Con el gráfico a la izquierda podemos corroborar que no hay un patrón claro en los residuos y que la varianza es (parcialmente) constante. Todo esto es cierto si ignoramos la presencia algunos outliers, lo cual confirma que lo que analizamos en puntos anteriores sobre las distribuciones de `Total Cup Points` y `Aroma`. Todo esto sugeriría que para mejorar nuestro modelo, deberíamos eliminar esos valores atípicos de la estimación.




**Punto 11**

```
tidy(coffee_model,conf.int = TRUE ) %>%
  knitr::kable(format = "latex", digits= 3)
```

| term | estimate | std.error | statistic | p.value | conf.low | conf.high |
|---|---|---|---|---|---|---|
| (Intercept) | 37.983 | 1.283 | 29.594 | 0 | 35.466 | 40.501 |
| aroma | 5.833 | 0.169 | 34.443 | 0 | 5.501 | 6.165 |

La prueba de hipótesis sería:

$$H_0 : \hat{\beta}_1 = 0$$

$$H_1 : \hat{\beta}_1 \neq 0$$

La hipótesis nula es que la pendiente de la línea de regresión es 0. La hipótesis alternativa es que la pendiente es diferente de 0. El estadístico de la prueba es 34 que es mayor al valor crítico 1.96 para un nivel de significancia del 5% y El p-value es 0 incluyendo 3 dígitos. Como conclusión se rechaza la hipótesis nula, $\hat{\beta}_1$ es estadísticamente significativo (diferente de cero).

## Punto 12

El 47% de la varianza de `Total Cup Points` es explicado por la variación de `Aroma`.

```
glance(coffee_model) %>%
  knitr::kable(format = "latex", digits = 3) %>%
  kableExtra::kable_styling(latex_options = c("scale_down", "hold_position"))
```

| r.squared | adj.r.squared | sigma | statistic | p.value | df | logLik | AIC | BIC | deviance | df.residual | nobs |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.47 | 0.47 | 1.956 | 1186.315 | 0 | 1 | -2795.337 | 5596.673 | 5612.27 | 5112.454 | 1336 | 1338 |

## Punto 13

13% de los cafés en nuestra muestra provienen de Colombia.

```
coffee_data <- coffee_data %>%
  mutate(colombia = ifelse(country_of_origin == "Colombia", 1, 0))
mean(coffee_data$colombia, na.rm = TRUE)
```

```
[1] 0.1368736
```

## Punto 14

Manteniendo todo constante, los cafes colombianos tienen en promedio 0.5 puntos más que los cafés que vienen de otros lugares.

```
coffee_model_colombia <- lm(total_cup_points ~ aroma + colombia, coffee_data)
tidy(coffee_model_colombia,conf.int = TRUE ) %>%
  knitr::kable(format = "latex", digits= 3)
```

| term | estimate | std.error | statistic | p.value | conf.low | conf.high |
|---|---|---|---|---|---|---|
| (Intercept) | 38.298 | 1.286 | 29.788 | 0 | 35.776 | 40.820 |
| aroma | 5.781 | 0.170 | 34.027 | 0 | 5.448 | 6.114 |
| colombia | 0.568 | 0.156 | 3.646 | 0 | 0.262 | 0.873 |

## Punto 15

El $R^2$ aumento en 0.5%. La variable `colombia` no añade mucho poder explicativo al modelo.

```r
glance(coffee_model_colombia) %>%
  knitr::kable(format = "latex", digits = 3) %>%
  kableExtra::kable_styling(latex_options = c("scale_down", "hold_position"))
```

| r.squared | adj.r.squared | sigma | statistic | p.value | df | logLik | AIC | BIC | deviance | df.residual | nobs |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.475 | 0.475 | 1.947 | 604.318 | 0 | 2 | -2786.728 | 5581.455 | 5602.248 | 5059.052 | 1334 | 1337 |

## Punto 16

En este caso, `flavour` es un mejor predictor ya que su varianza explica el 68% de la varianza de `Total Cup Points` en comparación con el 47% del punto 12.

```r
model_flavor <- lm(total_cup_points ~ flavor, coffee_data)
glance(model_flavor) %>%
  knitr::kable(format = "latex", digits = 3) %>%
  kableExtra::kable_styling(latex_options = c("scale_down", "hold_position"))
```

| r.squared | adj.r.squared | sigma | statistic | p.value | df | logLik | AIC | BIC | deviance | df.residual | nobs |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.683 | 0.683 | 1.512 | 2883.298 | 0 | 1 | -2451.142 | 4908.284 | 4923.88 | 3056.248 | 1336 | 1338 |

## Punto 17

La recomendación sería que invirtiera más en `flavour` ya que el retorno en la puntuación de la calidad es mayor por cada unidad más en el sabor del café.

```r
model_aroma_flavor <- lm(total_cup_points ~ aroma + flavor, coffee_data)
tidy(model_aroma_flavor,conf.int = TRUE ) %>%
  knitr::kable(format = "latex", digits= 3)
```

| term | estimate | std.error | statistic | p.value | conf.low | conf.high |
|---|---|---|---|---|---|---|
| (Intercept) | 29.828 | 1.008 | 29.602 | 0 | 27.851 | 31.805 |
| aroma | 1.389 | 0.191 | 7.270 | 0 | 1.014 | 1.764 |
| flavor | 5.554 | 0.177 | 31.410 | 0 | 5.208 | 5.901 |

## Punto 18

En este caso elegí la variable `body`.

```r
coffee_data <- coffee_data %>%
  mutate(log_cup_points=log(total_cup_points),
         log_body=log(body))
model_body <- lm(log_cup_points ~ log_body, coffee_data)
tidy(model_body,conf.int = TRUE ) %>%
  knitr::kable(format = "latex", digits= 3)
```

| term | estimate | std.error | statistic | p.value | conf.low | conf.high |
|---|---|---|---|---|---|---|
| (Intercept) | 3.428 | 0.035 | 96.907 | 0 | 3.359 | 3.498 |
| log_body | 0.486 | 0.018 | 27.704 | 0 | 0.451 | 0.520 |

Un aumento de 1% en el cuerpo del café está asociado a un aumento de 0.48% en `Total Cup Points`.