

Logit difference heatmap: $\text{logit}(\text{'Smith'}) - \text{logit}(\text{'Jones'})$

