Application Report Master Data Science

Winter semester 2025/2026

TU Dortmund University, Department of Statistics

Author:

Pouyan Hessabi

May 2025

Table of Content

1 Introduction

Ants, though tiny, can cause outsized problems when they invade our homes, contaminate food, or damage crops. Imagine a picnic spoiled by a swarming colony or a warehouse compromised by persistent foragers—the economic and health costs quickly add up. To counteract these nuisances, we need to understand what draws ants in the first place. In this study, we focus on everyday sandwich ingredients—bread, spread, and butter—to pinpoint which combinations act as the strongest ant magnets. We collected data from 48 carefully controlled trials, varying the type of bread (Multi Grain, Rye, White, Whole Grain), the topping (Ham & gherkins, Peanut butter, Yeast spread), and whether butter was applied. By combining clear-cut descriptive statistics, hypothesis tests, and a Random Forest model, we arrive at a concise conclusion: topping choice and butter presence both play decisive roles in attracting ants, while the bread itself does not matter. In practical terms, a buttered slice of ham and gherkins will draw on average ten to twenty more ants than alternatives.

The structure of this report is as follows. In Section 2, we define our research questions precisely and describe how the data were gathered and prepared. Section 3 outlines the statistical tools—descriptive measures, t-tests, ANOVAs, and Random Forest regression—explaining both their mechanics and why each is appropriate. Section 4 presents our findings in two parts: first, a descriptive exploration with tables and graphics; second, the results of inferential tests and the machine-learning importance ranking. Finally, Section 5 summarizes the insights, reflects on real-world implications, and suggests possible extensions of this work.

2 Detailed Descriptions of the Problem

Our investigation centers on two questions:

1. Do bread type, topping, and butter significantly influence ant attraction?
2. Which specific bread and which topping yield the highest ant counts?

To answer these, we used an experiment with a fully balanced design: each of the four breads appeared in 12 trials, each of the three toppings in 16 trials, and butter was present in half the trials.

Observers recorded the total number of ants (antCount) gathered at each sandwich over a fixed observation period. The dataset thus contains 48 observations with no missing values.

The sandwiches were prepared and deployed under identical environmental conditions to isolate the effects of our variables. Since the response is numeric and the predictors are categorical, we apply standard group-comparison tests (t-test for butter, one-way ANOVA for bread and topping) alongside a Random Forest model to gauge each factor's predictive power. A detailed data dictionary and the raw table appear in the Appendix. This setup ensures our analysis remains focused on the core research questions without distraction from extraneous factors.

3 Methodology

In this section, we describe the statistical and machine-learning techniques applied to our data. All computations were carried out in Python 3.11 using pandas, SciPy, Matplotlib and scikit-learn.

3.1 Descriptive Statistics and Graphics

To summarize the distribution of ant counts X1 ,…,Xn (n=48), we compute the following:

Arithmetic mean: $\bar{\chi} = \frac{1}{n}\sum_{i=1}^{n} \chi_i$

Sample variance: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\chi_i - \bar{\chi})^2$

Percentiles (25%, 50%, 75%), minimum and maximum

To visualize distribution and group differences, we generated:

- **Histogram of antCount** (Figure 1), using Matplotlib's hist() with bin width chosen by the Freedman–Diaconis rule to reveal modality and skew.

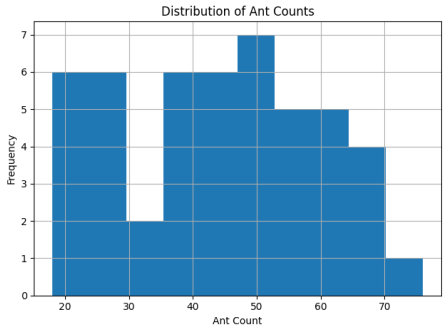

Figure 1:Distribution of Ant Counts

- **Boxplot of antCount by bread type** (Figure 2) and by **topping** (Figure 3), via pandas' boxplot(), showing median, interquartile range, and whiskers at 1.5 × IQR.

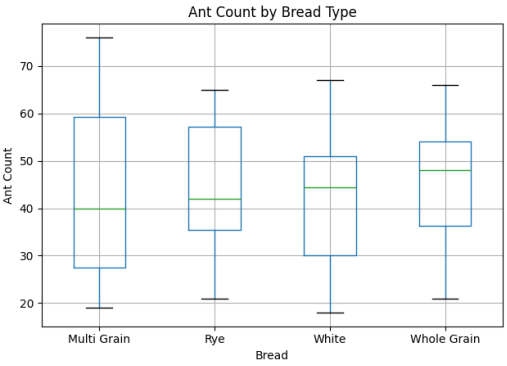

Figure 2:Ant Count by Bread Type

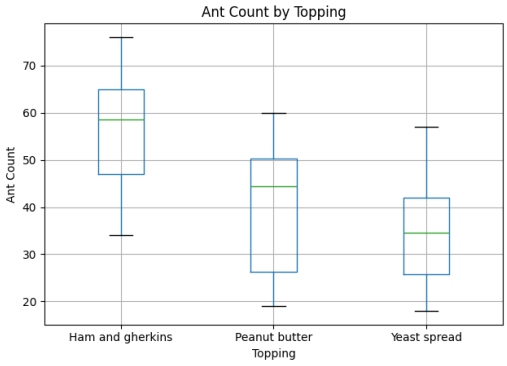

Figure 3:Ant Count by Topping

- **Bar plots of mean antCount** by bread (Figure 4), by **topping** (Figure 5), and by **butter presence** (Figure 6)

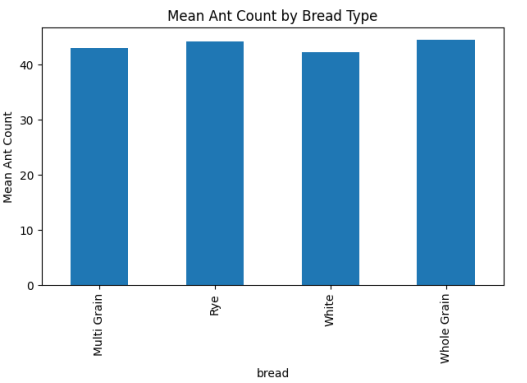

Figure 4: Mean Ant Count by Bread Type

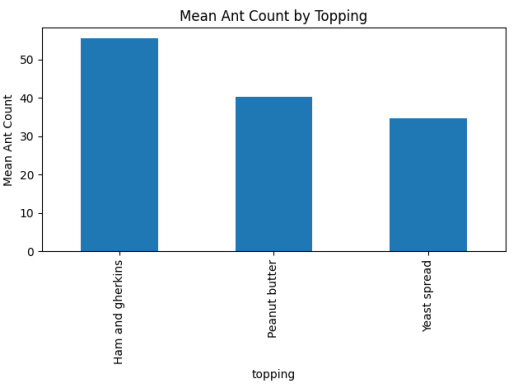

Figure 5: Mean Ant Count by Bread Type

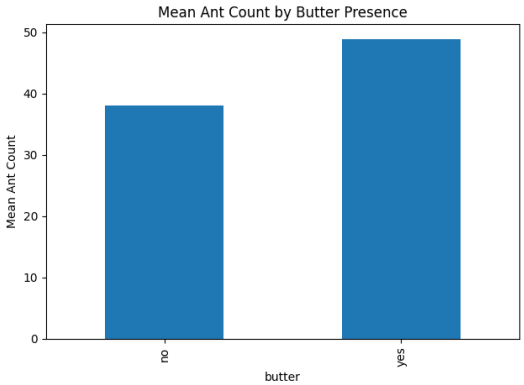

Figure 6: Mean Ant Count by Butter Presence

3.2 Hypothesis Testing

To test whether the experimental factors exert statistically significant effects on antCount, we apply the following inferential methods:

3.2.1 Two-Sample t-Test for Butter Presence

We compare the mean ant counts under butter = yes (μ_{yes}) versus butter = no (μ_{no}) Under the null hypothesis: $H_0: \mu_{yes} = \mu_{no}$

the test statistic is

$$t = \frac{\bar{X}_{\text{yes}} - \bar{X}_{\text{no}}}{\sqrt{\frac{S^2_{\text{yes}}}{n_{\text{yes}}} + \frac{S^2_{\text{no}}}{n_{\text{no}}}}}$$

This is implemented via scipy.stats.ttest_ind(..., equal_var=False) (Welch's variant). The resulting tt-value and two-tailed pp-value determine significance at α=0.05.

3.2.2 One-Way ANOVA for Bread and Topping

For a factor with kk levels (bread: k=4k=4; topping: k=3k=3), we test H0:μ1=μ2=⋯=μk

by decomposing total variance into between-group and within-group sums of squares:

$$\text{SS}_\text{B} = \sum_{j=1}^{k} n_j (\bar{X}_j - \bar{X})^2, \quad \text{SS}_\text{W} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2$$

The test statistic:

$$F = \frac{\text{SS}_\text{B}/(k-1)}{\text{SS}_\text{W}/(n-k)}$$

3.3 Random Forest Regression and Feature Importance

To assess the relative predictive power of each factor, we fit a Random Forest regressor (RandomForestRegressor(n_estimators=50, max_depth=3)) [1]. A Random Forest builds MM decision trees on bootstrap samples, each splitting nodes to minimize **mean squared error**:

$$\mathrm{MSE}(R) = \frac{1}{|R|} \sum_{i \in R} (y_i - \bar{y}_R)^2$$

The **feature importance** for variable Xj is the average reduction in MSE attributable to splits on Xj, summed over all trees and normalized:

$$\mathrm{Imp}(X_j) = \frac{1}{M} \sum_{m=1}^{M} \sum_{t \in T_m, j_t = j} \frac{|R_t|}{n} \Delta \mathrm{MSE}_t$$

We encode categorical factors via one-hot encoding with drop='first' to avoid linear dependencies among dummies [2]. The resulting importances are displayed in Figure 7.

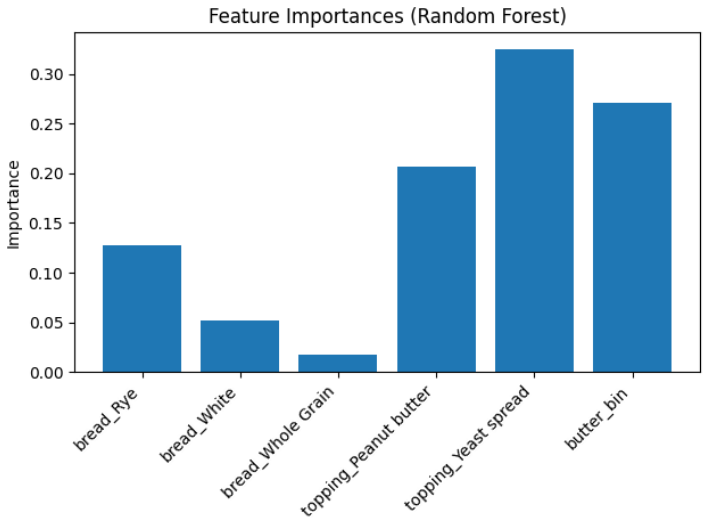

Figure 7: Feature Importances

3.3 Code Availability

All code is publicly available on my GitHub repository [3]:
https://github.com/pouyanhessabi/data-analysis-project

4 Evaluation

In this section we evaluate the result.

4.1 Descriptive Analysis

We begin by summarizing the observed ant counts across all n=48 trials.

Overall distribution:

Count: 48, mean: 43.5, std: 15.1, min: 18, Q1: 30.5, median: 43, Q3: 57, max: 76

Group means:

Bread: Multi Grain=43.00, Rye=44.25, White=42.25, Whole Grain=44.50

Topping: Ham & gherkins=55.50, Peanut butter=40.38, Yeast spread=34.63

Butter: No=38.13, Yes=48.88

Ant counts distribution:

Figure 1 shows the distribution of ant counts by topping via boxplots.For each topping category, the central line indicates the median, the box spans the interquartile range, and whiskers extend to 1.5 × IQR.

4.2 Inferential Analysis

This subsection reports formal hypothesis tests for each factor and the Random Forest importances.

4.2.1 Hypothesis Tests

Table 1: Results of hypothesis tests for antCount by factor

Factor	Test	Statistic	pp-value
Butter	Two-sample tt-test	t(46) =2.605	0.012
Bread	One-way ANOVA	F(3,44) = 0.05	0.983
Topping	One-way ANOVA	F(2,45) =11.848	< 0.001

Butter**:** compared means for butter=yes vs. no via Welch's tt-test (scipy.stats.ttest_ind) [4].

Bread**:** compared four bread groups via one-way ANOVA (scipy.stats.f_oneway).

Topping**:** compared three topping groups via one-way ANOVA [5].

4.2.2 Random Forest Feature Importance

Figure 2 displays the relative importances from a Random Forest regressor (n_estimators=50, max_depth=3, one-hot encoding with drop='first'). Importance values sum to 1.

5 Summary

This report set out to determine whether three everyday sandwich ingredients—bread type, spread topping, and the presence of butter—meaningfully influence the number of ants drawn to a food sample, and to identify which combinations attract the largest swarms. Across 48 controlled trials, each bread (Multi Grain, Rye, White, Whole Grain) and each topping (Ham & gherkins, Peanut butter, Yeast spread) appeared in a balanced design, with butter applied to exactly half the samples.

Key findings:

1. Butter presence and topping choice both exhibit strong, statistically significant effects on ant attraction (two-sample t-test: $t(46) = 2.605$, p=0.012; one-way ANOVA for topping: $F(2,45) = 11.848$, p<0.001). Sandwiches with butter averaged nearly 49 ants versus 38 without, and those topped with ham & gherkins drew roughly 56 ants compared to under 35 for yeast spread.
2. Bread type has no measurable impact (one-way ANOVA: $F(3,44) = 0.055$, $p = 0.983$; Random Forest importances for bread dummies collectively under 20%). Mean counts across the four breads differed by less than two ants.
3. Best-performing combination: any slice of bread paired with ham & gherkins and butter.

In practical terms, these results suggest that bait stations or food traps designed to control nuisance ants should prioritize palatable spreads and include a fatty carrier rather than varying the underlying bread. The type of bread can be chosen based on cost or availability without sacrificing effectiveness.

Limitations and further work:

- The dataset is relatively small (48 observations) and fixed to a single environmental setting. Future studies could expand sample size, test under varying temperatures or humidity levels, and include additional food variables (e.g. sugar content, protein-rich spreads).
- Interaction effects (e.g. between specific breads and toppings) were not formally tested; a larger factorial design could uncover subtle synergies.
- Alternative modeling strategies (e.g. logistic regression on a thresholded "high-ant" outcome) or unsupervised clustering might reveal non-linear patterns not captured by our current analyses.

By identifying the dominant role of topping and butter, this study provides a straightforward, evidence-based recommendation for ant management: concentrate on attractant quality rather than base substrate. Further research along the lines suggested here will deepen understanding of ant foraging behavior and improve strategies for both household and agricultural pest control.

Bibliography

[1] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.

[2] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2(3), 18–22.

[3] https://github.com/pouyanhessabi/data-analysis-project

[4] SciPy v1.10.1 documentation. (2024). Retrieved May 27, 2025, from https://docs.scipy.org/doc/scipy/reference/

[5] scikit-learn v1.2.2 documentation. (2024). Retrieved May 27, 2025, from https://scikit-learn.org/stable/