Methods of Data Manipulation: Misleading Wholesale Spending Trends

Nour Moghazi (900225966)

Nour Kahky (900221042)

Omar Moustafa (900222400)

Abdelrahman Baioumi (900223218)

Dataset Description

The data set refers to clients of a wholesale distributor. It includes the annual spending in monetary units (m.u.) on diverse product categories

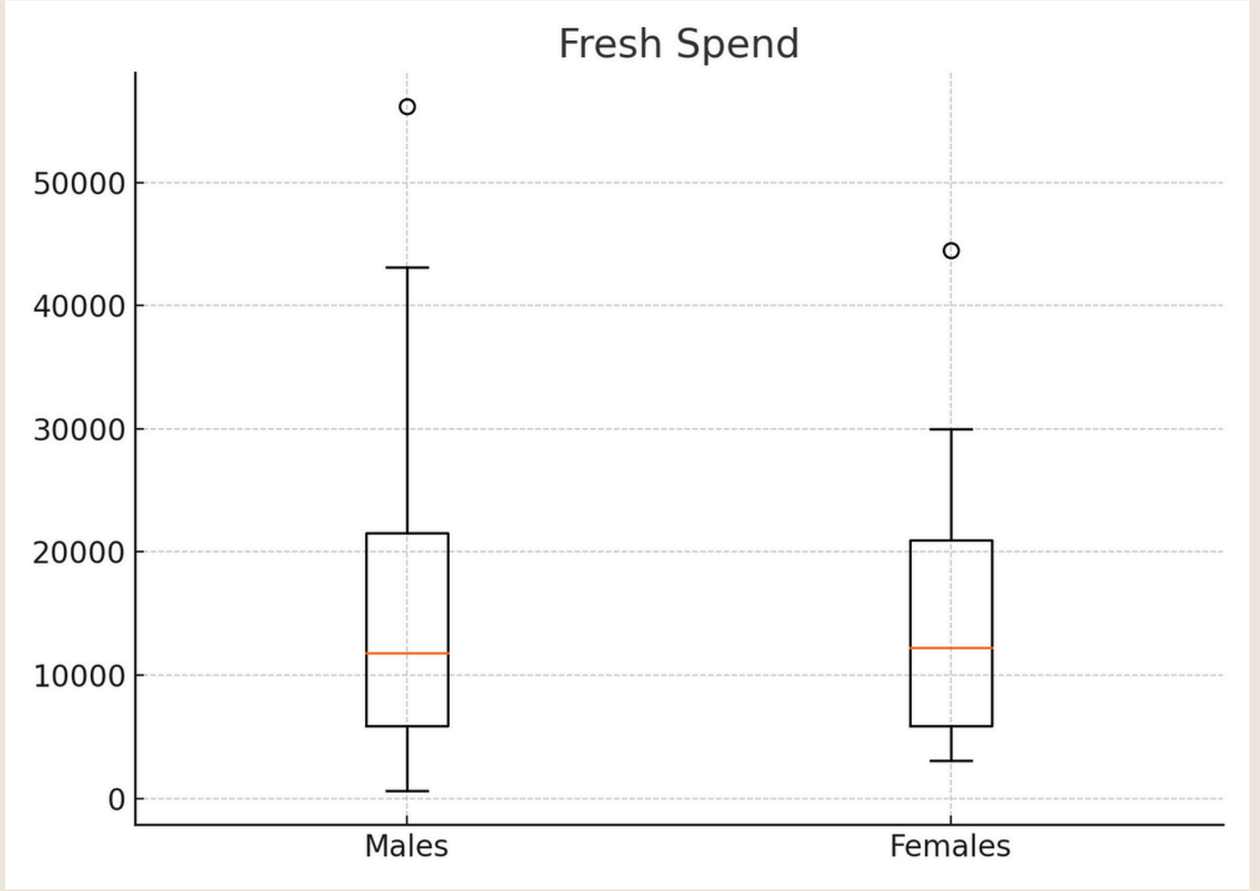
1) FRESH
2) MILK
3) GROCERY
4) FROZEN
5) DETERGENTS PAPER
6) DELICATESSEN
7) CHANNEL
8) REGION
9) Gender

Visualization Statistics

Claim: Males Spend More on Fresh Products

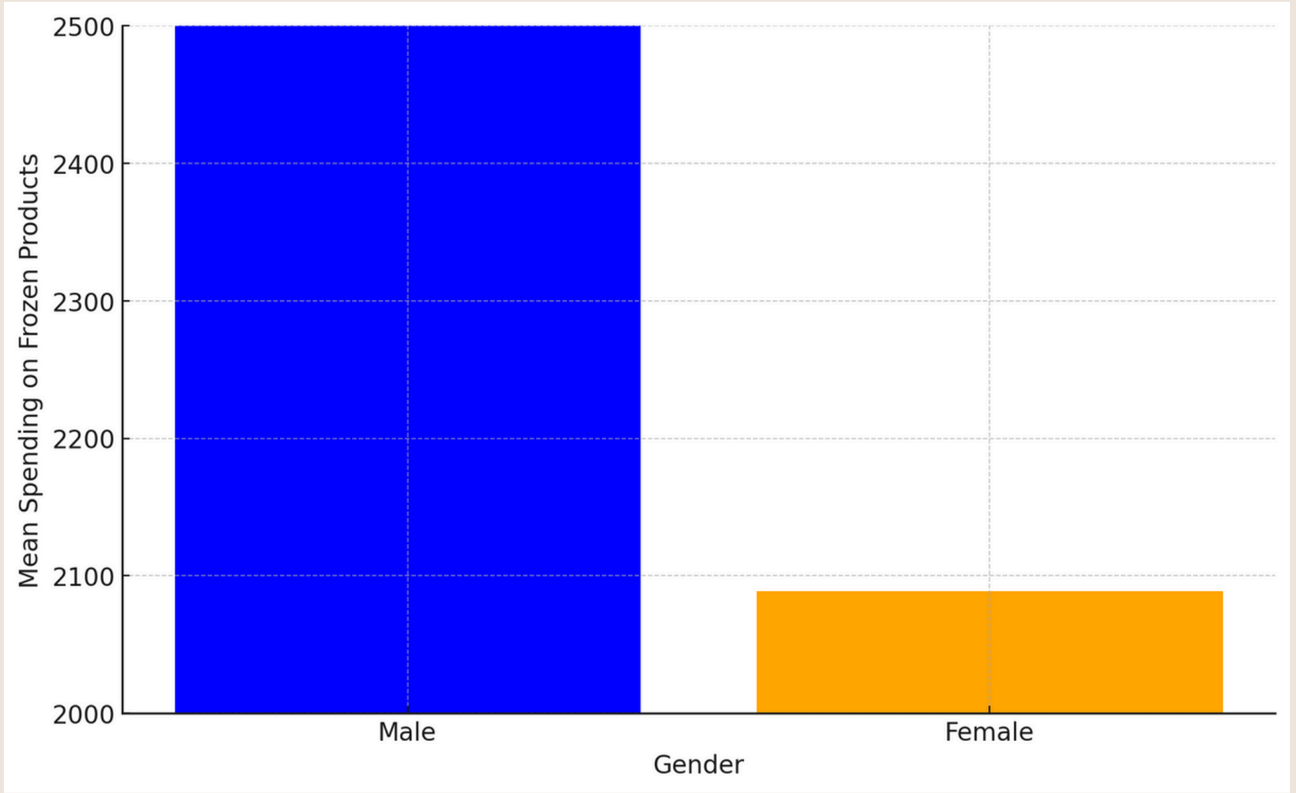
Visualization Statistics

Corrected Visualization

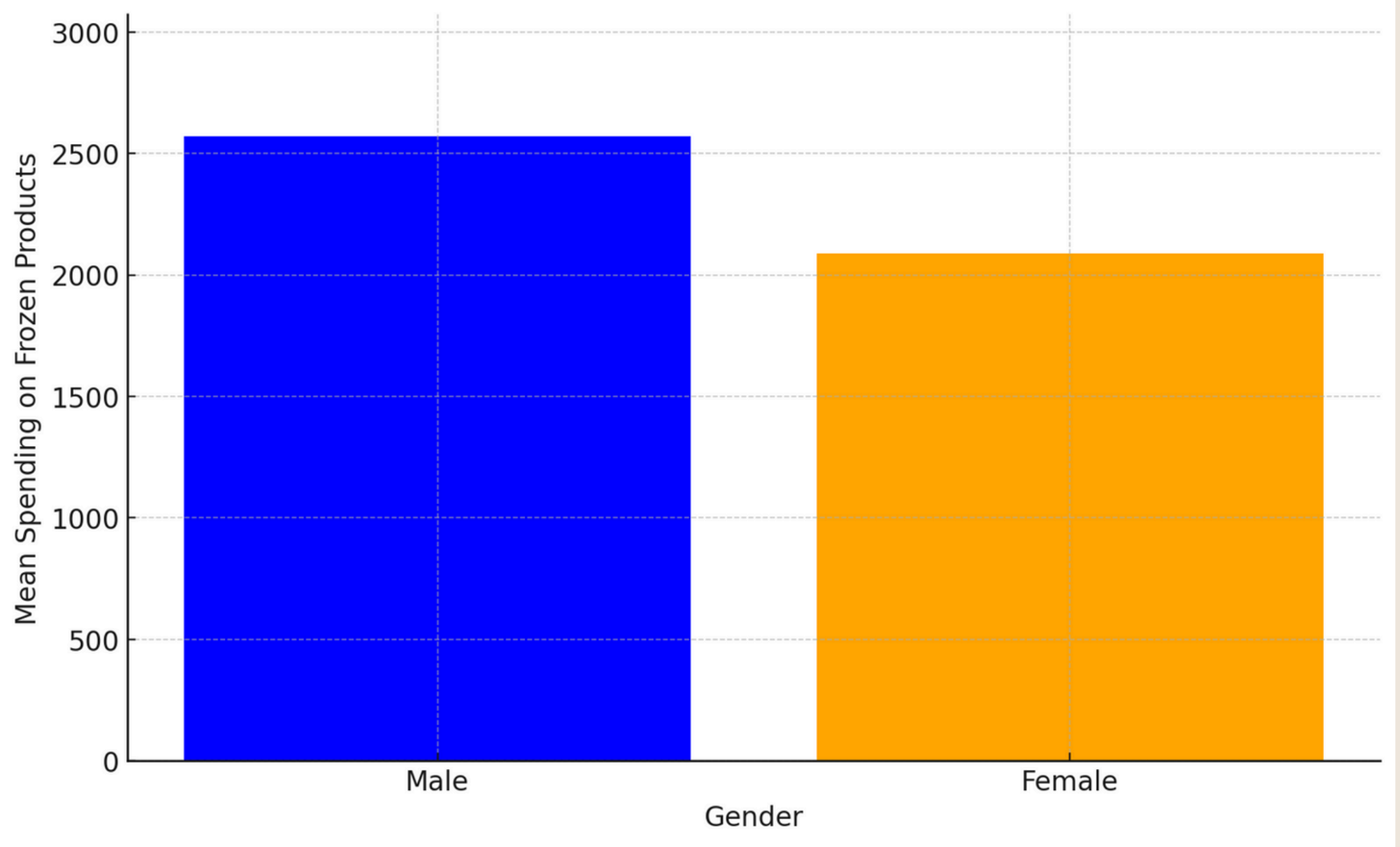


Visualization Statistics

Claim: Males Spend Signifacantly More on Frozen Products



Corrected Visualization



Calculating Confidence Interval For Sample Mean

- An interval estimate at a 99% confidence level for the sample mean of the FRESH variable was calculated
- Based on a significance level, where α = 0.01, we are 99% confident that the population mean lies within the interval [11544.19, 18298.17]
    - This suggests an extremely high degree of confidence in the results

Significance Level Correction

- However, looking at the actual analysis below, it corresponds to a 95% confidence interval meaning that α = 0.05 (not 0.01!)
- x bar ± Z * (s/(n^0.5))
- —> x bar = Sample Mean = 14921.18 —> Z = z-score corresponding to 95% CI = 1.96
- —> s = sample standard deviation = 12183.13 —> n = sample size = 50
- —> CI = 14921.18 ± 1.96 * (12183.13 / (50^0.5)) = [11544.19, 18298.17]
-
- This, therefore, demonstrates how stating a different value for α can lead to a misleading impression about the certainty of the findings
-

Common Methods of Data Manipulation

1. Cherry-picking results
2. Misleading visualizations
3. Falsifying statistical tests

Potential Impacts of Data Manipulation

1. Makes the scientific findings and research untrustworthy

2. Leads to inaccurate conclusions which would affect the decision-making steps that follow the conclusions

Thank You!