Part 2.2 | Numerical Variables by Category

The Economic Question

In Part 0, we saw how Card and Krueger tested a hypothesis about minimum wage. Economic theory predicted one thing: higher minimum wages would reduce employment. But the data told a different story. They compared employment in New Jersey (which raised its minimum wage) to Pennsylvania (which didn't) and found no significant difference.

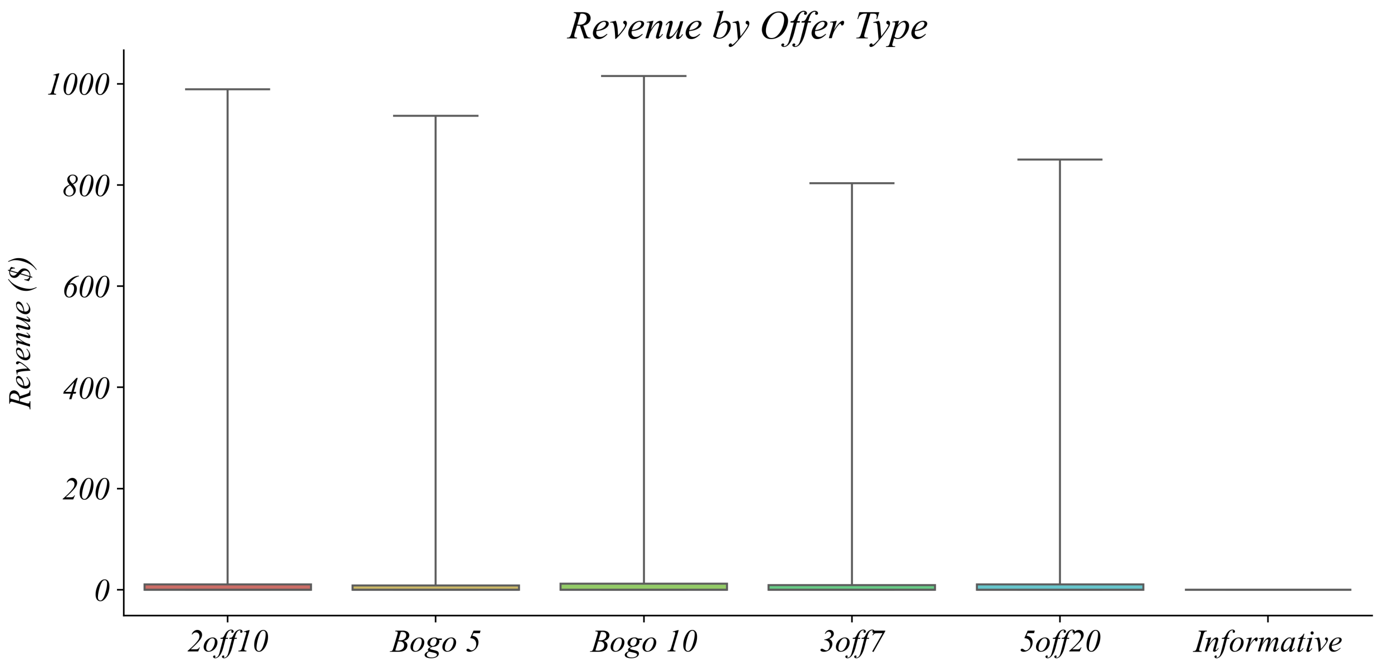
We're going to do something similar here, although not nearly as impactful. One of the basic ideas in economics is that people respond to incentives. Starbucks sends different promotional offers to different customers: BOGO deals, dollars off, percentage discounts. Economic intuition says bigger incentives should lead to bigger responses. But do they?

Our hypothesis: Customers who receive larger offers (like BOGO 10) will spend more than those who receive smaller offers (like BOGO 5).

The data contains records of every time Starbucks sent an offer and every purchase customers made. Our job is to test whether the data supports our hypothesis.

First Attempt: Just Look at the Data

Let's start by visualizing spending by offer type. A boxplot shows the distribution of a numerical variable across categories.



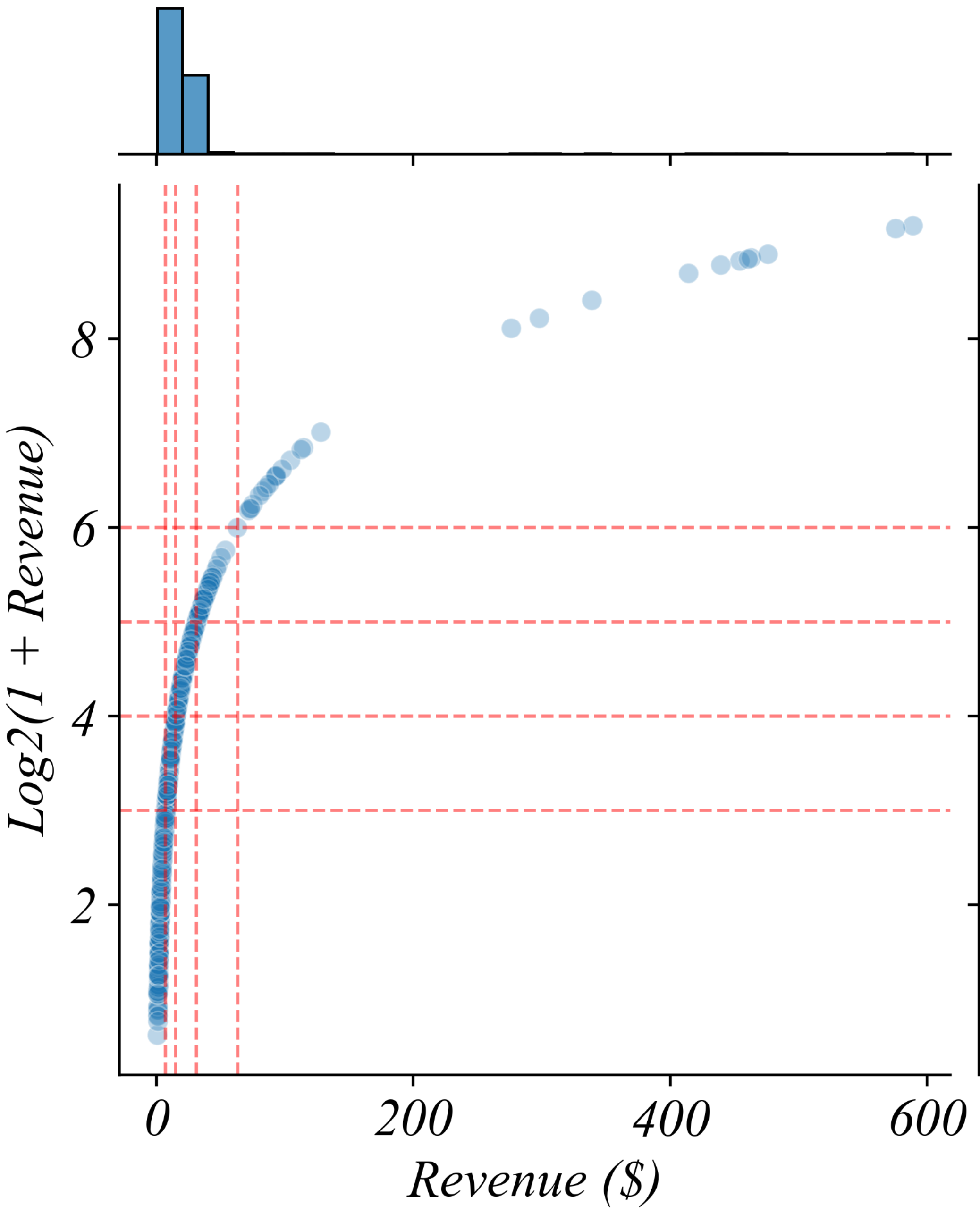
What's wrong with this picture? Everything is compressed at zero. We can't see any meaningful differences between offer types. This is a common problem with spending data — most values are small, but a few large purchases stretch the scale.

Log Transformation

When data is heavily skewed, a log transformation helps. We'll use log base 2, where each unit represents a doubling of spending:

- log2(1+$7) = 3
- log2(1+$15) = 4
- log2(1+$31) = 5

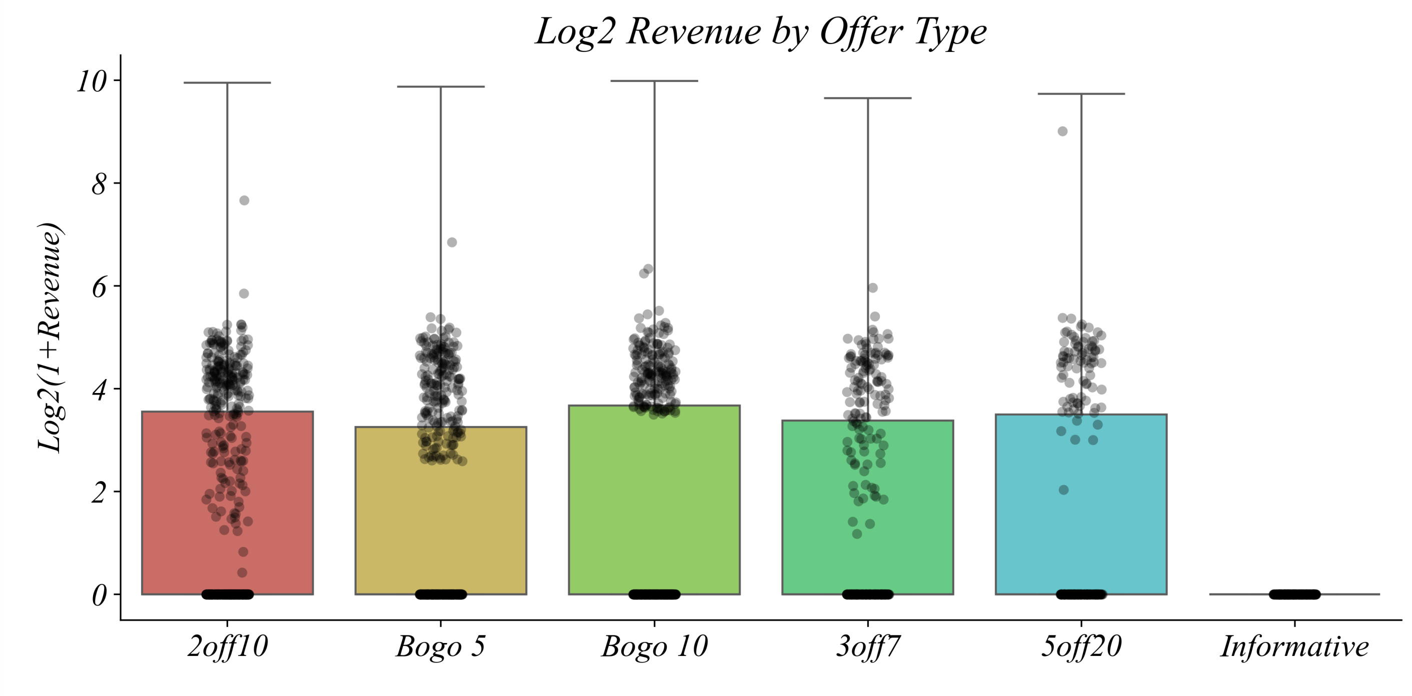
The "+1" handles zero-dollar values (since log of zero is undefined). Now small differences at low spending levels are visible, and large outliers don't dominate the scale.



The x-axis histogram shows the original data which is heavily skewed right. The y-axis histogram shows the transformed data which is much more spread out. This is why we transform.

A Problem in the Data

After transforming, we can see the distributions better. But there's a new problem: lots of zeros.



Why so many zeros? Let's investigate. The data has three types of events:

- **offer** — Starbucks sent an offer to a customer
- **transaction** — a customer made a purchase
- **offer completed** — a customer redeemed an offer

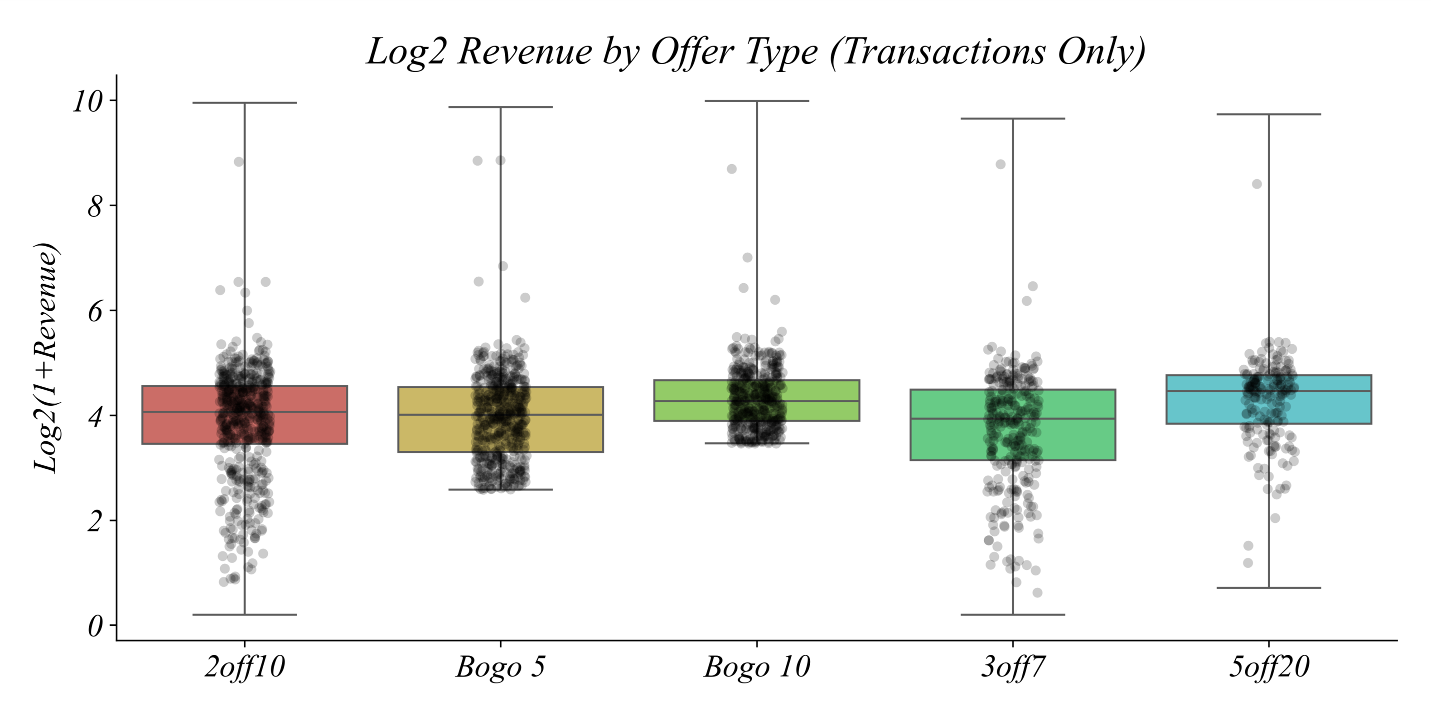
Only transactions have revenue. Offers and completions have zero revenue because they're not purchases. We were accidentally including non-purchases in our analysis.

Filter First, Then Analyze

This is a crucial lesson: *know your data before you analyze it.* We need to filter for transactions only.

```
transactions = data[data['Event'] == 'transaction']
```

Now every row is a real purchase.



The stripplot shows individual transactions. Now we can see actual variation in spending across offer types.

Grouped Statistics

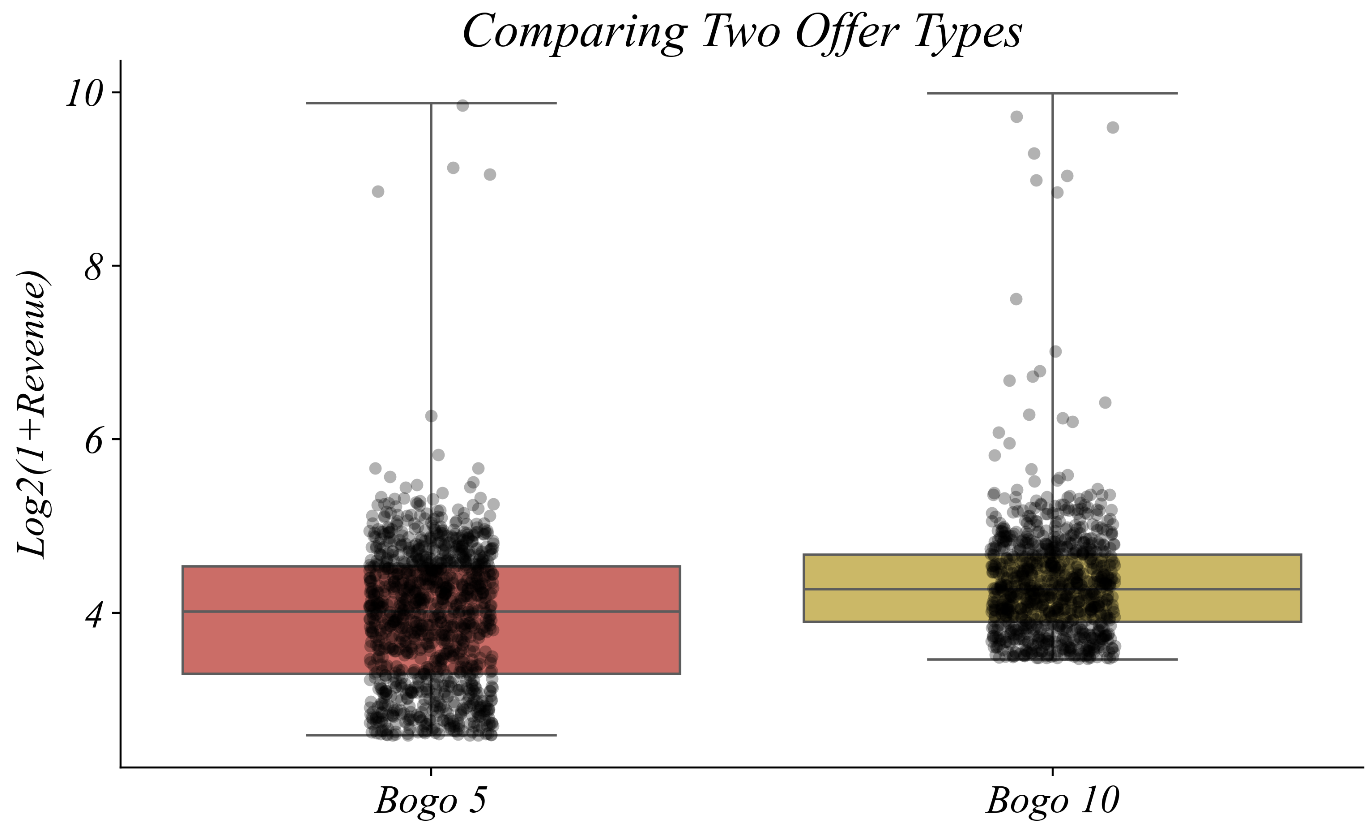
With clean data, we can calculate summary statistics by offer type:

Offer ID	Mean	Std	Count
2off10	4.32	0.89	...
3off7	4.18	0.94	...
5off20	4.58	0.77	...
Bogo 5	4.27	0.91	...
Bogo 10	4.41	0.85	...

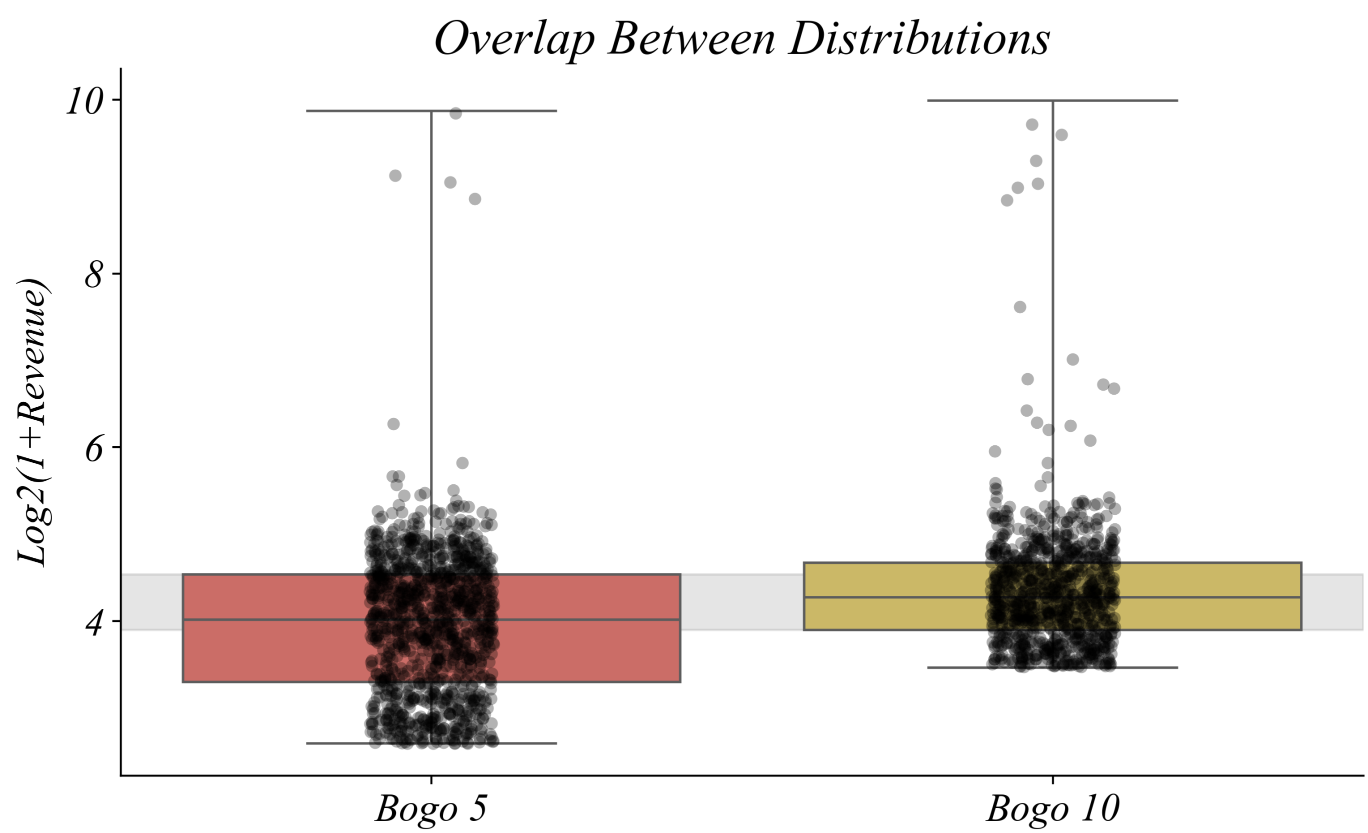
The means differ. 5off20 has the highest average log spending. Bogo 10 beats Bogo 5. But look at those standard deviations — there's substantial variation within each group.

Testing the Hypothesis: Bogo 5 vs Bogo 10

Back to our hypothesis. Economic intuition says Bogo 10 (a bigger incentive) should lead to more spending than Bogo 5. The means support this — Bogo 10 is higher. But let's look at the full distributions.



Bogo 10 does have higher average spending. But look at the overlap. Many Bogo 5 customers spent more than many Bogo 10 customers.



The gray region shows where the distributions overlap. It's substantial.

The Key Question

This is where Card and Krueger faced the same challenge. They found that New Jersey employment was slightly higher than Pennsylvania after the minimum wage increase. But was that difference real, or just noise?

We're in the same position. Bogo 10 spending is higher than Bogo 5 on average. But:

- There's lots of variation within each group
- The distributions overlap considerably
- Some Bogo 5 customers outspent Bogo 10 customers

Is the difference we observe actually meaningful, or could it have happened by chance?

This is the fundamental question of statistical inference. We can describe the data — we've done that. We can visualize it — we've done that too. But to answer whether the difference is real, we need new tools.

Looking Ahead

In Part 3, we'll learn how to answer this question formally. Just like Card and Krueger needed statistical methods to conclude that the minimum wage effect was "not significantly different from zero," we need methods to determine if Bogo 10 really outperforms Bogo 5, or if we're just seeing noise.

For now, the takeaway is this: *means can hide important information.* Always visualize your distributions. And when you see overlap, be skeptical of simple comparisons.

Summary

- **Summary statistics can hide problems** — always visualize
- **Log transformation** helps with skewed data
- **Filter your data —** make sure you're analyzing what you think
- **Boxplots by category** show distributions, not just means
- **Overlapping distributions** raise inference questions
- **Testing hypotheses** requires more than comparing means