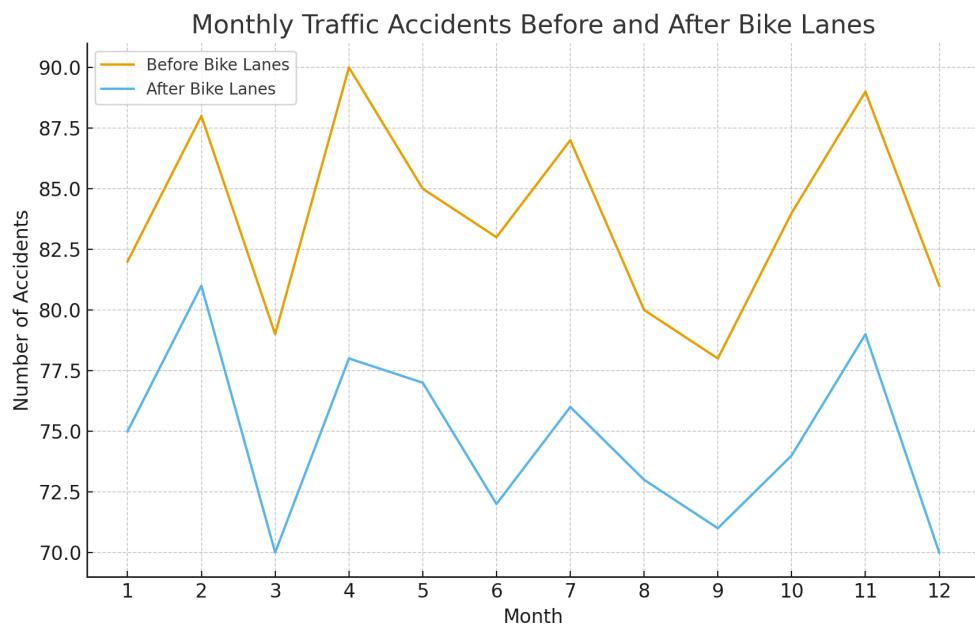


## Data Visualization Exercise

A city claims that adding bike lanes reduced traffic accidents. Two visualizations are provided. Discuss whether the conclusion is justified.

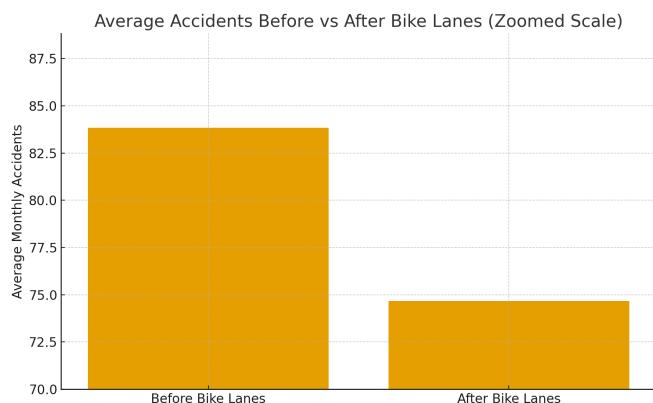
### Visualization A: Monthly Accidents (Before vs After)



Question:

What trends do you observe? Is the change large, small, or unclear?

### Visualization B: Average Monthly Accidents



Question:

Does this visualization make the difference appear larger or smaller? Why?

Compare both visualizations. Which one is more persuasive? Which one is more honest?

### Discussion Tasks

1. List at least 5 other factors that might affect accident rates.
2. Explain how axis scaling can distort perception.
3. Decide whether the claim is justified, unjustified, or uncertain.
4. Propose a new visualization that would more fairly represent the data.

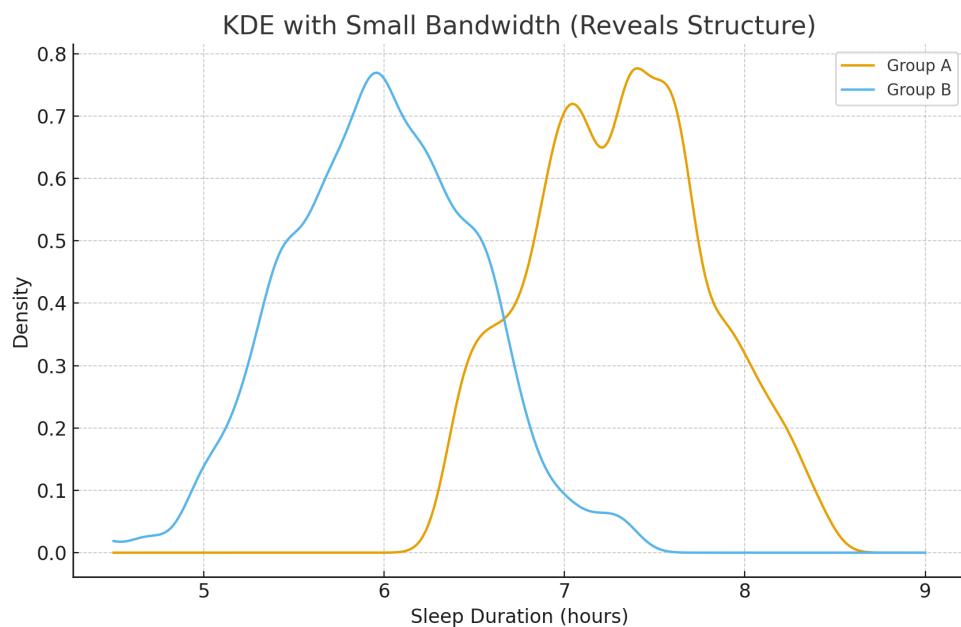
### KDE Exercise

Context: Two groups of adults were studied for nightly sleep duration.

Group A = 1 hour of screen use before sleep.

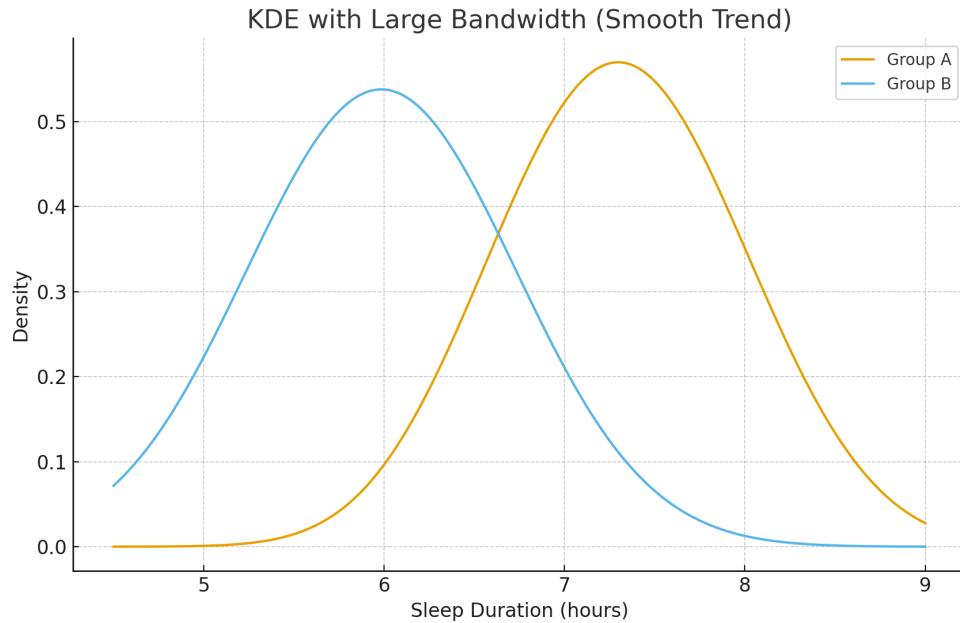
Group B = 3+ hours of screen use before sleep.

### Visualization 1: Small Bandwidth KDE



Discuss what patterns and variability you observe.

## Visualization 2: Large Bandwidth KDE



Discuss what overall trend this visualization suggests.

### Discussion Questions

- 1) How does changing the bandwidth affect what we learn from the distribution?
- 2) Which plot would a researcher prefer? Which might a marketer prefer?
- 3) Does either visualization alone justify the claim that screen use reduces sleep?
- 4) What additional data would be needed to test causality?

## **Solution:**

### **Key concept:**

Both charts show fewer accidents after bike lanes were added, but the bar chart exaggerates the difference by using a truncated y-axis. Thus, the data suggest a decrease, but do not prove that bike lanes caused it.

### **Interpretation of the Charts**

- Line chart: Shows a consistent but moderate month-to-month decrease.
- Bar chart: Shows averages only and visually exaggerates the drop because the y-axis does not start at zero.

### **1. Confounding Factors (Examples)**

Weather patterns, police enforcement, road construction, tourism/traffic volume changes, data reporting changes, economic/work-from-home shifts.

### **Conclusion**

The observed decrease (~11%) indicates an association, but causality is uncertain. To make a sound claim, we would need:

- Accident rates per number of road users (not raw counts)
- Control areas without bike lanes
- Multiple years to check seasonal effects

### **2. How axis scaling can distort perception**

When the y-axis does not start at zero, small numerical differences appear much larger visually. In the bar chart, starting the axis near 70 instead of 0 exaggerates the difference between 78 and 72 accidents. This can mislead the viewer into believing the change is larger than it actually is.

### **3. Is the claim justified?**

Conclusion: Uncertain.

Although accidents decreased after the installation of bike lanes, we cannot conclude that the lanes caused the change. Other variables such as weather, traffic volume, enforcement, or reporting changes may also explain the reduction. The data show an association, not proof of causation.

### **4. A fairer visualization**

A bar chart with a zero-based y-axis and error bars, or a line chart with multiple years to account for seasonality, would more accurately reflect the data. Ideally, use accident rates (e.g., accidents per 10,000 trips) instead of raw counts.

**Solution:**

**Key Concept: Bandwidth Interpretation**

Small bandwidth reveals fine structure and possible subgroups.

Large bandwidth reveals overall distribution trends but hides detail.

**Answers to Discussion Questions**

- 1) The small bandwidth KDE suggests clusters or sub-patterns in each group. The large bandwidth KDE smooths these into general peaks, making the groups look more uniformly different.
- 2) A scientific researcher would prefer the small-bandwidth version to see structure. A marketer promoting 'sleep improvement products' might prefer the large-bandwidth version to present a clearer separation.
- 3) No. The KDEs show distribution differences, but do not prove causality. Confounders like stress, workload, caffeine, or lifestyle may explain the differences.
- 4) Needed: Randomized controlled assignment of screen time, measurement of confounders, and possibly longitudinal tracking to observe consistent effects.