Stat C131A: Statistical Methods for Data Science

Fall 2025

Lecture 3: Visualization

Sep 4 2025

Book keeping ….(contd)

- **My office hours:** Fridays 11AM-12PM (on Zoom)
- **Relocated course Website:** https://stat131a.berkeley.edu/fall-2025
- **Waitlist update**

Visualization

- Textbook reference for today's topics
  - **2 Data Distributions**
    - **2.1 Basic Exploratory Analysis**

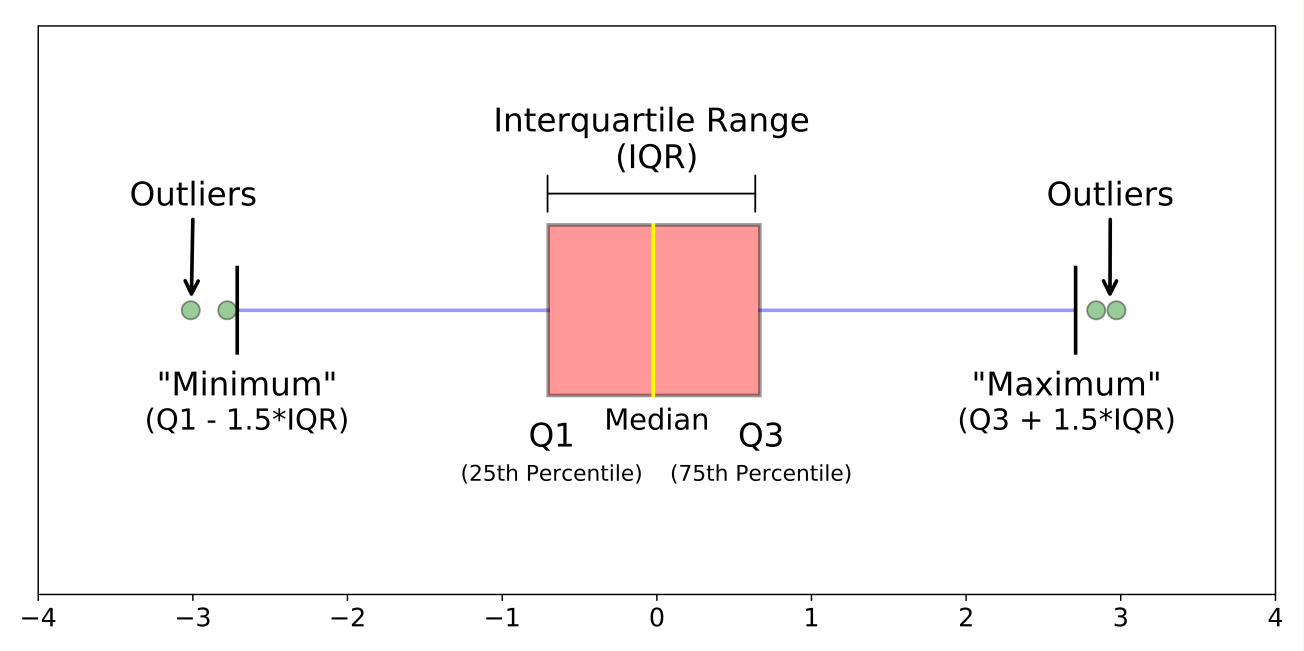
Summarization (of Distribution)

- Q1, Q3, IQR
- Q1, Q3: Quartiles
- IQR: Interquartile range
  - What does it capture ?

Histograms

- **Frequency** histograms
- **Density** histograms
  - Plot the height of rectangles so that the **area** of each rectangle is equal to the **proportion of observations** in the bin
  - Will revisit when we come to distributions

Boxplots

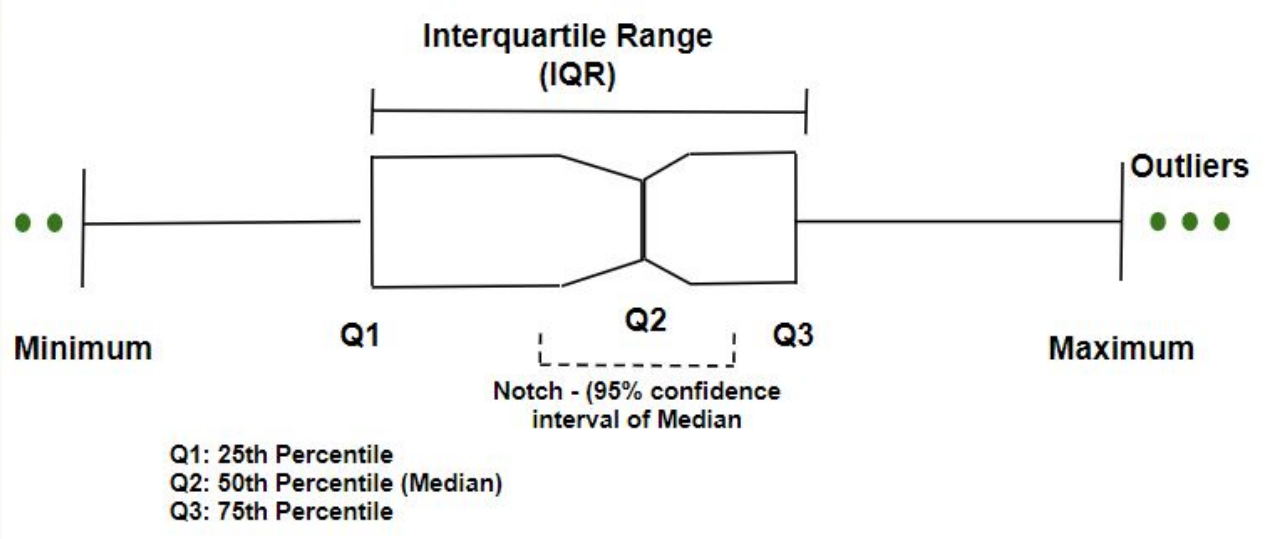


- **Box**
- **Whiskers**
  - lower whisker: max(1st Qu.−(1.5 x IQR), Min)
  - upper whisker: min(3rd Qu.+(1.5 x IQR), Max)
- **Outliers**

R plot

- **Boxplot:** a compact summary of a distribution.
  - Box = **Q1 to Q3** (middle 50%).
  - Line inside = **median** (typical value).
  - Whiskers = most extreme points still considered "typical"
  - Dots beyond whiskers = **outliers**.
- **Markers we added:**
  - **Mean** = solid circle (average; pulled by high/low extremes).
  - **Median** = diamond (middle person; robust to outliers).

Notched Boxplot



- **Notched boxplot:** the notch around the median is a **rough 95% CI (confidence interval)** for the median (approximate)
- **Jitter:** tiny random side-to-side wiggles on points
  - so overlapping values become visible

Violin Boxplot

- **Violin plot:** a **mirrored density** (smoothed histogram)
  - Thick parts = many people; thin parts = few.
- **Thin boxplot** on top: keeps the **median/IQR** visible.
- **Jittered points:** raw observations to show actual values.

Boxplot with outliers labeled

- **Tukey rule:**
  - Compute **IQR = Q3 − Q1**.
  - Whiskers extend to the most extreme points **within 1.5 × IQR** from the quartiles.
  - Points **beyond** those limits are **outliers**
- Outliers are **labeled** with IDs
  - look up

Log1p (log one plus)

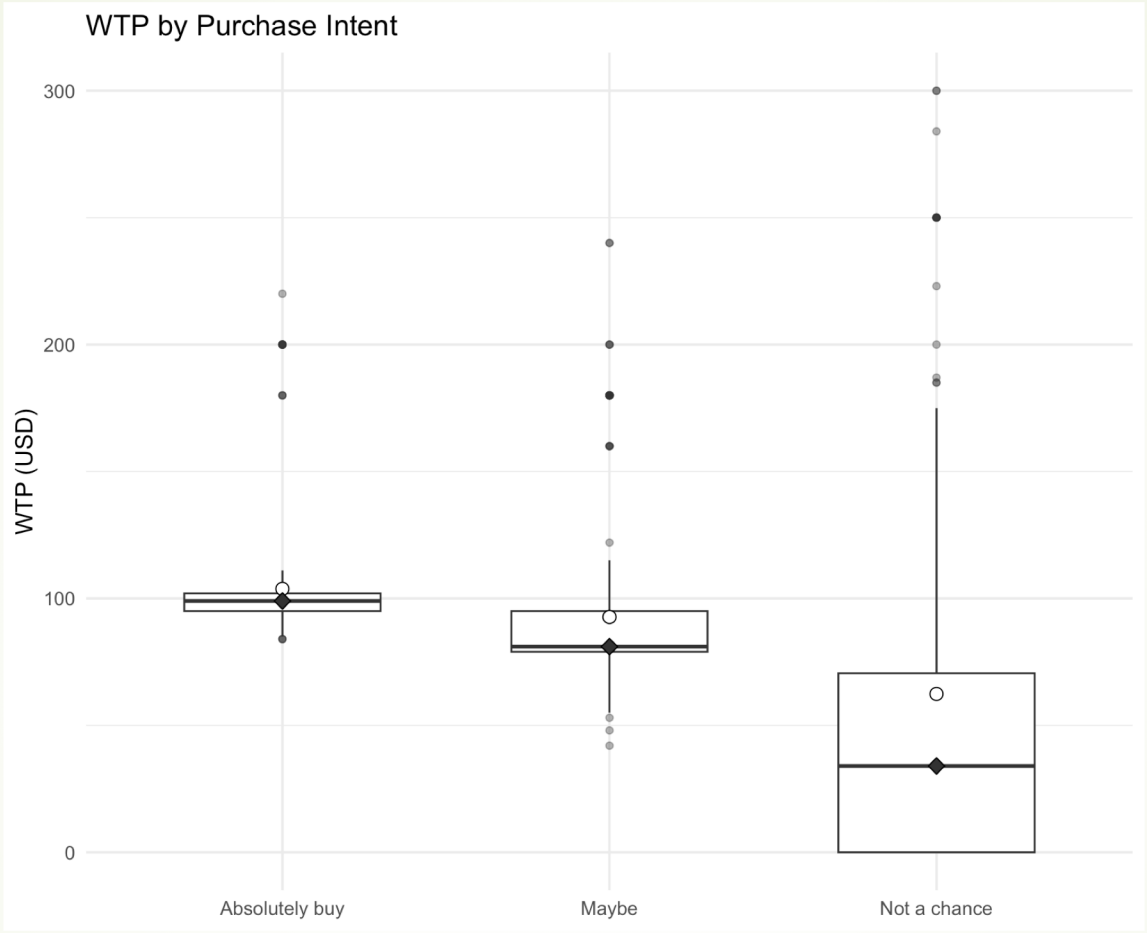
- log(1+X)
- Log1p tames **skew/zeros**; and think in **multiplicative** terms.
  - Adding 1 lets us include **zeros**
    - log(0) is undefined
  - Log scale **compresses the right tail** and **spreads out** small values
    - Making skewed data easier to compare

In Summary

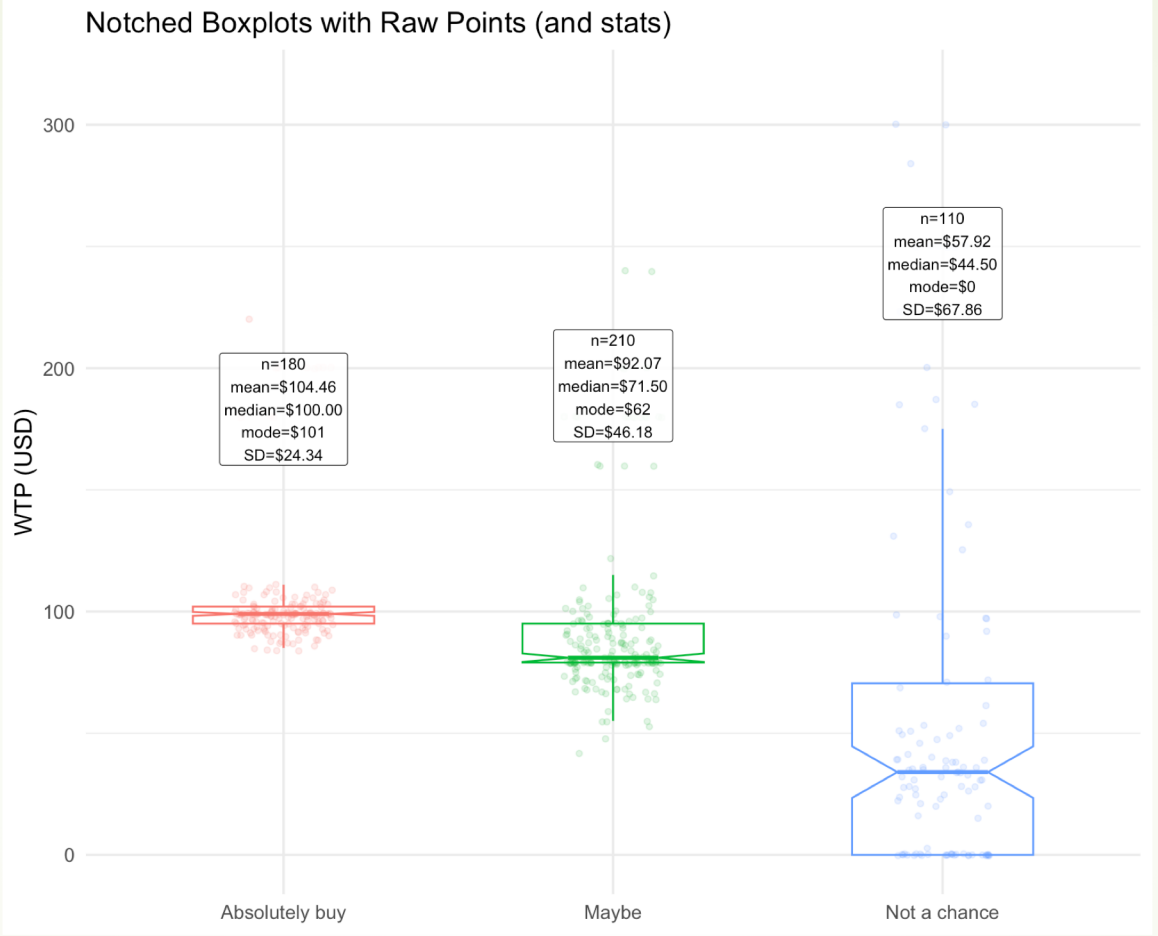
- **Boxplot:** fast **center & spread** comparison.
- **Mean vs Median markers:** mean = **AOV planning**; median = **typical buyer**.
- **Notch:** rough **median CI**; non-overlap ≈ medians differ.
- **Jitter:** shows **stacked/duplicate** values and **clumps** (modes).
- **Violin:** shows **shape** (skew, multimodality)
  - Can't see from a box alone.
- **Tukey outliers:** surface **niches** or **data errors** to inspect.
- **Horizontal:** readability for presentations.
- **Log1p:** tames **skew/zeros**; think in **multiplicative** terms.

Various Boxplots we went through (in R)

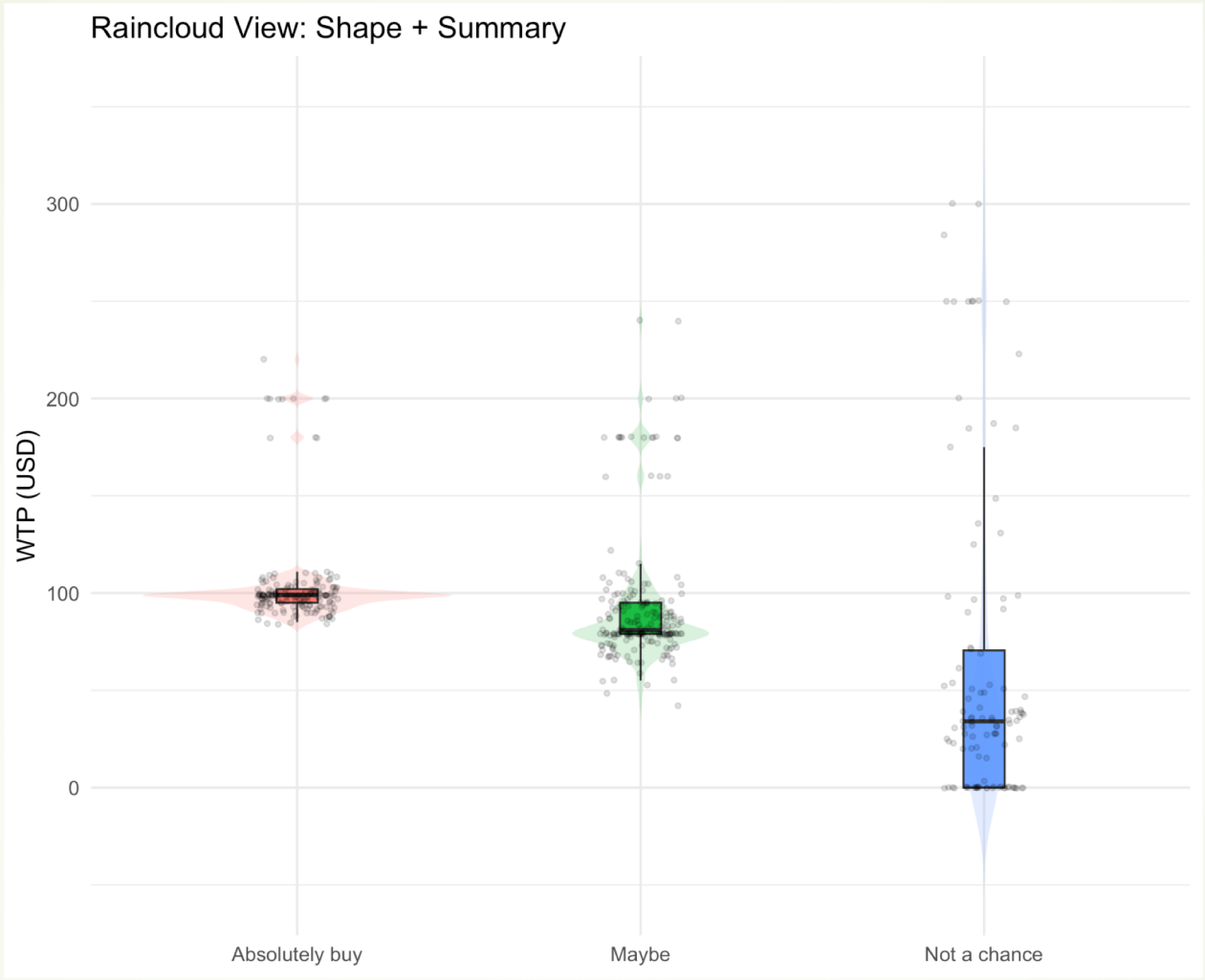
The Q&A we had is also included



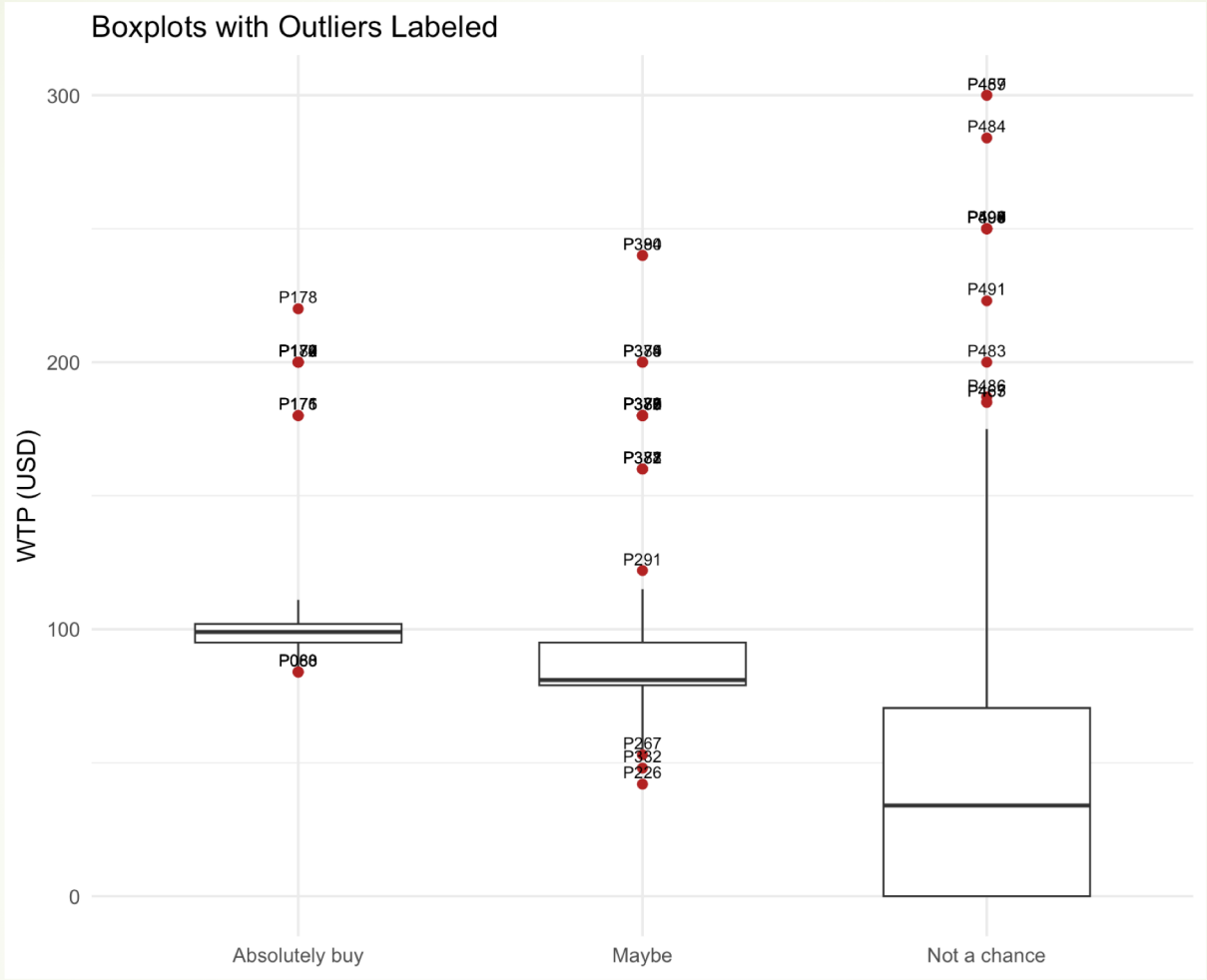
- **Which group has the highest *typical* WTP? How do you know?**
  Answer: *Absolutely buy* - its **median** line is highest.
- **In which group is the mean noticeably above the median, and what does that imply?**
  Answer: *Absolutely buy* (and often *Maybe*). **Mean > median** ⇒ **right-skew** (a high-paying tail).
- **Which group is most homogeneous in WTP?**
  Answer: *Absolutely buy* - **smallest IQR** and tighter whiskers.



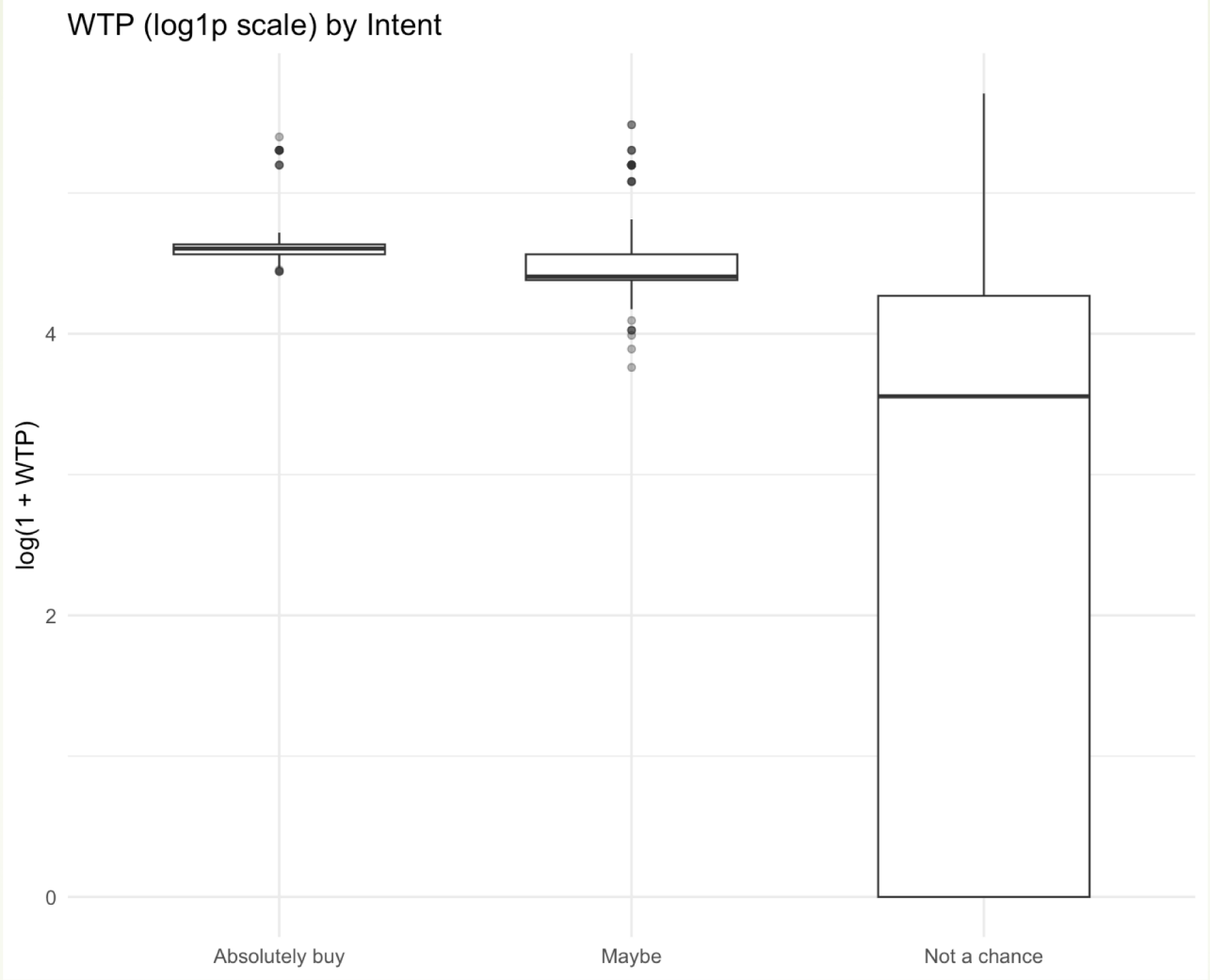
- **Do the notches for "Absolutely buy" and "Maybe" overlap? What's the implication?**
  Answer: Typically **no**; **non-overlapping notches** suggest **different medians** (rough ≈95% evidence).
- **What do clusters of jittered points near $79 and $99 suggest?**
  Answer: Modes/price magnets - $79 (entry) and $99 (flagship).
- **If "Maybe" shows many points beyond the upper whisker, what does that say about dispersion?**
  Answer: Large spread / heavier tail - segment is **heterogeneous**; consider multiple price tiers.



- **If "Maybe" has two visible bulges (≈$79 and ≈$99), what kind of distribution is that?**
  Answer: Bimodal - two distinct preference clusters.
- **What does a long skinny upper "cloud tail" indicate?**
  Answer: A **premium** niche (few buyers willing to pay much more).



- **Pick one high outlier in "Absolutely buy." Is that an error or a niche? How would you check?**
  Answer: Could be
  - a **niche** (check the text feedback)
  - an **error !** (check the text for that as well, maybe says "*I'm just kidding!* ☺ "



- **log1p**: log **1 PLUS** x
- **Merits**
  - Multiplicative differences (percent style)
  - 0 values can be handled (log of 0 is undefined)
  - Tames the skew: compresses the right/upper tail