

WEEK 2 COMPANION GUIDE

Data Types → Transformations → Measurement Levels → Visualizations

A practical guide you'll use throughout the course

How to use this guide: This is not just a reference for classifying data. It is a guide for thinking about:

- What questions your data can answer
- What patterns you can reveal
- What decisions your analysis can support

Key idea: The way you transform and classify your data determines what you can see, interpret, and act on.

A. Transformations & Data Preparation

How we shape data to fit the questions we're trying to answer

1 Why Transformations Matter

Transformations are not just about cleaning your spreadsheet. They reshape your data so you can answer the questions that matter. Almost every meaningful visualization relies on transformed data, even if your software does it automatically.

 **Why this matters:** The same dataset can answer completely different questions depending on how you transform it.

2 Types of Transformations

Filtering, grouping, aggregating, binning, and reshaping data each change how the story in the data shows up. Many datasets are too granular or too messy to visualize effectively without these steps.

 **Why this matters:** Raw data rarely answers questions without reshaping first.

Five Core Transformations

Filter

Remove rows that don't match criteria

Group

Organize by categories or segments

Aggregate

Sum, average, count, max, min

Bin

Create ranges (0-10, 11-20...)

Reshape

Pivot, transpose, restructure

3 Derived Variables (New Variables You Create)

Some of the most useful insights come from variables that didn't originally exist. Things like "days since last workout," "age group," or "approval status." These new variables often reveal patterns that raw data can't show on its own. Most of the interesting variables in real organizations aren't sitting in the raw data. Someone has to create them, and that is often where the analysis actually begins.

 **Why this matters:** Many insights only appear after creating derived variables.

Common derived variables:

- Time since event (days since last purchase)
- Categorical bins (age groups, income brackets)
- Status flags (overdue, active, approved)
- Ratios and percentages (conversion rate, profit margin)
- Counts and frequencies (visits per month, orders per customer)

4 Coding Turns Qualitative Data Into Something You Can Compare

Several students worried they "didn't have enough data" because their dataset felt too qualitative. The good news: qualitative data becomes useful once it's coded. Coding lets you organize open-ended responses, categorize behaviors, group patterns, and even assign ordinal structure, without distorting the original meaning. You're not turning text into false numbers. You're organizing it so you can count it, sort it, and compare it.

 **Why this matters:** Essential for qualitative datasets and Week 3-4 visualizations.

Ways to code qualitative data:

- Tag themes or categories
- Group similar responses
- Extract key attributes (e.g., "contains chocolate," "high intensity")
- Convert text to ordinal groups (e.g., "low/medium/high difficulty")
- Count frequencies of themes
- Create flags (Yes/No variables)

</> Coding Example: Open Responses to Categories

Raw Responses

"Dark chocolate bar"
"Milk chocolate truffle"
"Gummy bears"
"Hard candy mint"



Coded Categories

Chocolate
Chocolate
Gummy
Hard Candy



Analysis Ready

Count by category
Compare patterns
Visualize distribution

✓ **Best for:** Students with qualitative-heavy datasets (providers, candy types, workouts, onboarding). This becomes the foundation for meaningful visualization in Weeks 3 and 4.

5 Always Store Data at the Highest Level Possible

Store dates, times, and numbers in their raw form (ratio or interval) whenever possible. You can always transform downward, to categories, bins, or labels, but you can never reconstruct lost detail. This ensures maximum flexibility when building visualizations.

💡 **Why this matters:** Lost detail cannot be recovered later.

↓ You Can Transform Down, Never Up

Store: Full Timestamp

2024-03-15 14:23:07
Maximum flexibility



Derive: Date

2024-03-15

Can still get day/month/year



Derive: Day of Week

Friday

Lost all time detail

Once you aggregate to "Friday," you cannot go back to the original timestamp.

6 Why You Sometimes Transform Variables Downward

Transforming downward (like turning a continuous date into a discrete weekday) helps you answer specific, categorical questions. It can expose patterns that raw values hide, such as peak days, clusters, or segment differences. The trick is knowing when a simpler form reveals more insight.

💡 Why this matters: Simpler forms can reveal hidden patterns more clearly.

Example: Daily sales data might hide the pattern that Fridays are always highest. Transforming to "day of week" immediately reveals this insight through a bar chart.

7 From Metrics to Systems (Pipelines & Journeys)

Individual variables tell you what is happening. Connected variables tell you how a system is working. This is where data becomes much more powerful.

Why this matters: Connected data allows you to identify friction points and improve outcomes.

Examples of connected data:

- Application → Interview → Hire = Hiring pipeline
- Completion → Time → Score = Learning pipeline
- Arrival → Security → Boarding = Experience journey

Key shift: You move from describing performance → to improving performance.

B. Levels of Measurement

The foundation that determines what you can calculate, visualize, and decide

8

Levels of Measurement Overview

Nominal, ordinal, interval, and ratio variables each allow different mathematical and visual operations. Chart types expect certain levels of measurement, so misclassifying data can lead to misleading visuals. Knowing the level of measurement helps you select charts intentionally, not accidentally.

 **Why this matters:** Wrong level means wrong chart and misleading conclusions.

Levels of Measurement Ladder

RATIO

True zero + Equal intervals + Order + Categories | Examples: Height, Weight, Revenue, Age

INTERVAL

Equal intervals + Order + Categories | Examples: Temperature (°C/°F), Credit Score, Year

ORDINAL

Order + Categories | Examples: Rankings, Difficulty (Low/Med/High), Education Level

NOMINAL

Categories only | Examples: Colors, Types, Gender, City, Product Category

You can transform down the ladder but never up. Always store at the highest level possible.

What each level of measurement enables

- **Nominal:**
Compare categories (what, who, where)
- **Ordinal:**
Show progression and ranking
- **Interval:**
Analyze differences and trends
- **Ratio:**
Analyze efficiency, proportions, and performance

Key insight: The level of measurement does not just describe your data — it determines what kinds of insights are possible.

Critical reminders:

- **IDs are always nominal even if numeric** (customer numbers, zip codes)
- **Orders without equal spacing are still ordinal** (small/medium/large)
- **Ratios unlock "twice as much" interpretations** (meaningful zero required)
- **The same variable can be two levels at once** depending on the question you ask of it (a renewal probability is ratio when you multiply it into a forecast, ordinal when you sort accounts by it)

9

The Identifier vs Measurement Test

Not every number in a dataset is actually measuring something. One of the biggest breakthroughs this week was students realizing that many numeric-looking values are really just labels. A number is only quantitative if it reflects a measurable amount or magnitude. If it's simply naming or tagging something, it's qualitative, even if it contains digits. This matters because treating identifiers as true numeric values leads to meaningless statistics and misleading visuals.

💡 **Why this matters:** Prevents mistakes like averaging invoice numbers or charting IDs.

Identifier vs Measurement

🔑 Identifiers (Nominal)

Use numbers as labels, not measurements

- Customer IDs
- Sale or invoice numbers
- Employee numbers
- Reference codes
- Zip codes

Example: ID: 48291 → Nominal

📏 Measurements (Quantitative)

Numbers represent actual amounts

- Height, weight
- Calories burned
- Revenue
- Exam scores
- Temperature

Example: Calories: 482 → Ratio

Critical reminder: Just because a value is numeric does not mean it is quantitative. If the value only identifies, it is Nominal, even if written as "45191."

10 Nominal Variables

Nominal variables represent categories without order. Things like candy types, team names, or departments. These are perfect for bar charts, frequency tables, and categorical comparisons. They help answer questions about "which groups" or "how many in each group."

💡 Why this matters: Categories enable comparisons but not rankings or calculations.

Common Examples

- Product categories
- Department names
- Customer IDs (even if numeric)
- Zip codes
- Colors, brands, types

Best Visualizations

- Bar charts
- Pie charts
- Frequency tables
- Heatmaps (for cross-tabs)

Remember: IDs are always nominal even if they look like numbers. Customer number 45191 is just a label, not a measurement.

11 Ordinal Variables

Ordinal variables have an order but not consistent spacing. Examples include ratings, rankings, or bins like "low / medium / high." You can compare direction and order, but not meaningful numeric differences. These variables work well in sorted bars or comparisons that rely on sequence rather than magnitude.

💡 Why this matters: Order matters but distance between values does not.

Common Examples

- Satisfaction ratings (1-5 stars)
- Education level (HS, BA, MA, PhD)
- Difficulty (Easy/Medium/Hard)
- Rankings (1st, 2nd, 3rd)
- Shirt sizes (S, M, L, XL)

Best Visualizations

- Ordered bar charts
- Slope charts
- Stacked bars

Key distinction: Orders without equal spacing are still ordinal. The gap between "Easy" and "Medium" may not equal the gap between "Medium" and "Hard."

12 Interval Variables

Interval variables have equal spacing but no true zero. Temperature in Fahrenheit, raw dates, and many scaled scores fall here. You can add and subtract these values, but you cannot make ratio statements like "twice as much." Interval values often appear in line charts, trends, and time-series visuals.

 **Why this matters:** Equal spacing allows trends but not meaningful ratios.

Common Examples

- Temperature (Fahrenheit, Celsius)
- Calendar dates (2024-01-15)
- Credit scores
- SAT/GRE scores
- pH levels

Best Visualizations

- Line charts
- Time series
- Scatter plots
- Area charts

Why no true zero: 0°F doesn't mean "no temperature," and the year 0 doesn't mean "no time." This prevents ratio statements like "80°F is twice as hot as 40°F."

13 Ratio Variables

Ratio variables have a true zero and allow meaningful "twice as much" comparisons. Examples include duration, steps, money, age, or counts. These variables unlock more mathematical operations and give you flexibility when choosing visualizations such as line charts, bar charts, comparisons, and trend analysis.

💡 **Why this matters:** True zero unlocks meaningful comparisons like double/half.

🔧 Common Examples

- Revenue, costs, profit
- Age, height, weight
- Duration, distance
- Counts (customers, orders)
- Temperature in Kelvin

📊 Best Visualizations

- All chart types work
- Bar charts, line charts
- Scatter plots
- Histograms
- Any comparison visualization

Quick ratio test: Ask two questions. (1) Is the zero a real absence, or just a point someone picked? (2) Does "twice as much" actually make sense? You need a "yes" to both. A real zero on its own is not enough.

Negative values don't disqualify anything. A bank balance can read -\$2,545 and it's still ratio. Zero dollars is a real zero, and \$100 is genuinely twice \$50. Where the numbers land tells you nothing about the level of measurement. What matters is how the scale is built.

📅 C. Dates, Time, and Derived Categories

Understanding flexibility in temporal data and category creation

14 Dates Can Shift Levels of Measurement

Dates are unique because they change measurement level depending on how you use them. A raw date is interval, but weekday names are nominal, and elapsed time between dates becomes ratio. Understanding this flexibility helps you choose the right chart: line charts for continuous dates, bar charts for discrete categories.

💡 **Why this matters:** Same date, different treatment creates completely different insights.

📅 One Date, Multiple Levels

As Interval

2024-03-15

Use: Timeline, trend

→ Line chart

As Nominal

Friday

Use: Day comparison

→ Bar chart

As Ordinal

March (3rd month)

Use: Ordered periods

→ Ordered bars

As Ratio

14 days since

Use: Duration, elapsed

→ Any numeric viz

✓ Rules of Thumb for Dates

- ✓ Use continuous dates for trends and time series
- ✓ Use discrete dates for ranking, comparisons, or sorting
- ✓ Transform to day-of-week/month when pattern matters more than timeline

15 Continuous vs. Discrete Data

Continuous data can take any value within a range, while discrete data jumps in whole units. This matters because continuous data is suited for lines and trends, while discrete data is best shown using bars, histograms, or tables. The choice directly affects how your audience interprets the shape of the data.

💡 **Why this matters:** Wrong choice creates visual confusion about data behavior.

🔲 Discrete

Whole units you cannot break apart

- Number of people
- Count of events
- Products sold
- Votes received

Best for: Bar charts, histograms

Continuous

Any value along a continuum

- Time
- Revenue (\$245.67)
- Temperature
- Distance

Best for: Line charts, area charts

16 Ordinal Categories Are Often Created, Not Collected

You'll do this frequently: cluster numeric values into categories, rank outcomes, create low/medium/high groupings, assign experience levels. The key point: ordinal data is almost always the result of a transformation.

 **Why this matters:** Creating categories reveals patterns hidden in raw numbers.

Common ways to create ordinal variables:

- Bin continuous data (age 0-18, 19-35, 36-50, 51+)
- Rank performance (top 10%, middle 80%, bottom 10%)
- Categorize by threshold (low/medium/high revenue)
- Assign levels (beginner/intermediate/advanced)

→ From Ratio to Ordinal

Raw Data (Ratio)

\$45,000
\$72,000
\$125,000



Transformed (Ordinal)

Low Income
Medium Income
High Income

Transforming to ordinal simplifies comparisons but loses exact values.

D. Visual Thinking Foundations

How to align data structure with visualization purpose

17 Chart Choice Depends on Levels + Purpose

The level of measurement, combined with your question, determines which charts will work and which will fail. Bar charts, line charts, scatterplots, and heatmaps each expect certain data structures. Good visualization is simply matching chart logic with the nature of the data.

Important: The goal of visualization is not just to show data — it is to support understanding and decision-making.

💡 Why this matters: Right chart makes patterns obvious; wrong chart hides them.

Chart Selection Guide

Nominal

 Bar chart,  Pie chart,  Heatmap

Ordinal

 Ordered bar chart,  Slope charts

Interval

 Line chart,  Scatter plot,  Area chart,  Histogram

Ratio

All of the above plus any comparison or calculation-based visualization

The zero baseline rule is about the chart, not the data. A bar says what it means with length, so it has to start at zero. That's why bars only work honestly for ratio data. A line says what it means with position,

so it doesn't need a zero at all, which is why lines work for interval and ratio both. Ratio data earns a zero baseline. It doesn't force you to use one. A stock price is ratio and nobody charts it from zero.

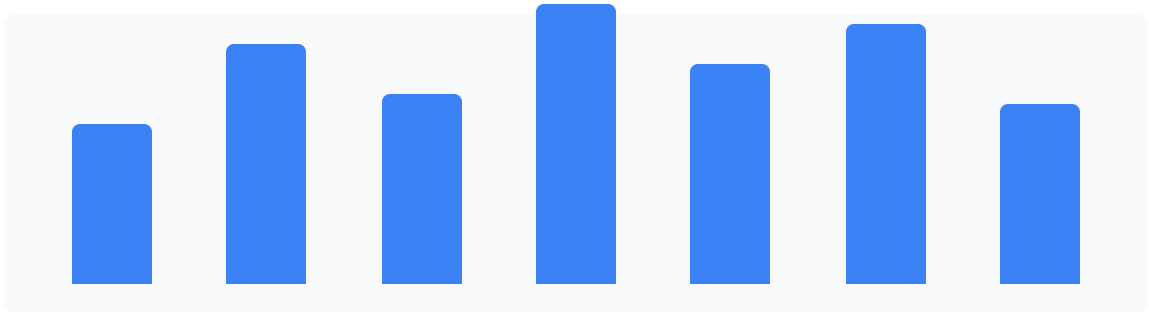
18 Visual Patterns Depend on How You Slice the Data

Changing the transformation, such as grouping daily data into weeks, or splitting users by category, can completely change what patterns appear. Slicing correctly requires understanding the story you're trying to uncover. Sometimes one small shift in how the data is aggregated reveals something you would have missed otherwise.

💡 **Why this matters:** Different slices reveal completely different insights from same data.

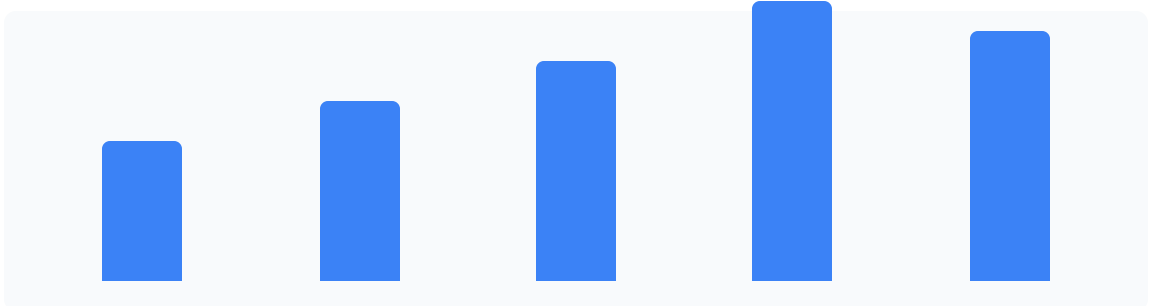
🔗 Same Data, Different Slices

By Day of Week



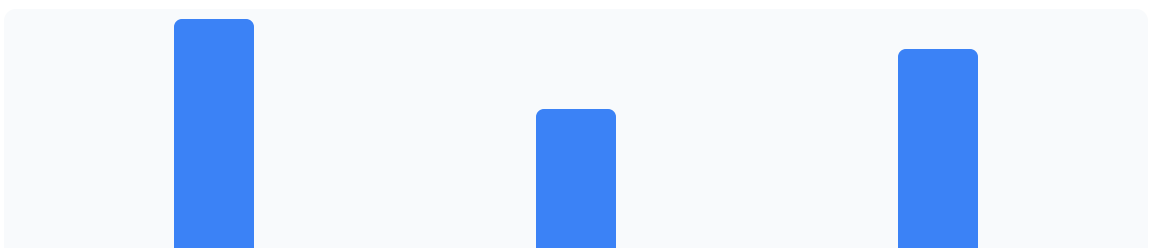
Reveals: Friday peak pattern

By Month



Reveals: Seasonal trend upward

By Customer Segment



Reveals: Segment A dominates

19 Choosing How to Treat Variables in Visualizations

Visualization tools often let you choose whether to treat a variable as continuous or discrete. This choice dramatically changes the chart's design and the story it tells. Knowing the difference helps you align the chart to the actual question you're answering.

💡 **Why this matters:** Same variable, different treatment produces entirely different charts.

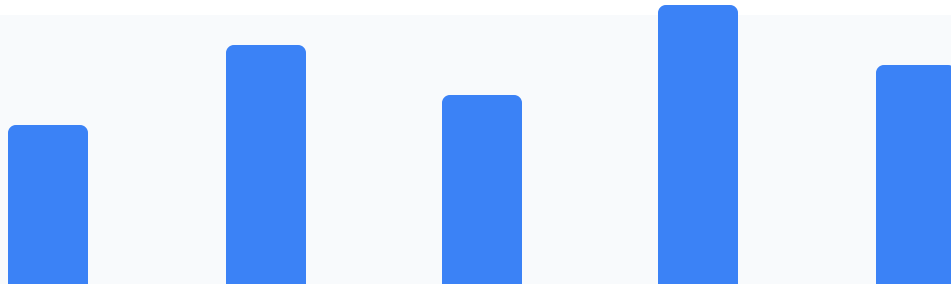
↔ Continuous vs Discrete Treatment

Date as Continuous



Shows: Trend over time

Date as Discrete



Shows: Comparison by period

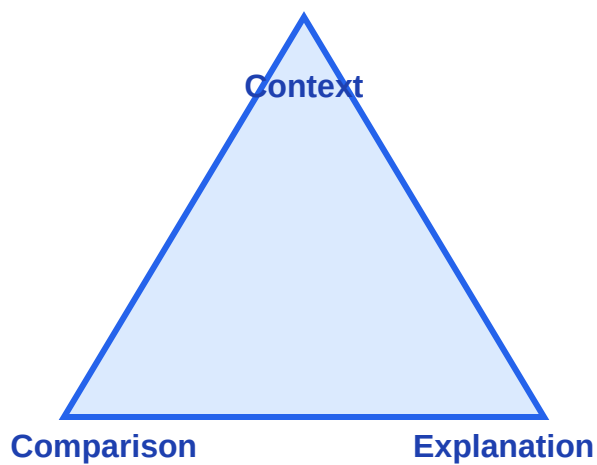
20

"Do I Have Enough Data?" (A Pattern You Can Apply Anywhere)

A common question this week was, "Do I have enough data?" Not meaning number of rows, but number of variables. The way to answer that is simple: look at whether your dataset allows you to add context, explain variation, or compare groups. If you can do all three, your dataset is strong. If you're missing one or two, consider adding variables that strengthen those dimensions.

💡 Why this matters: Works for every dataset and guides what to collect.

Three Dimensions of Strong Data



🔍 Three Key Questions to Assess Your Dataset

- ★ **Does this add context?** Helps you understand who/what/when
Examples: category, type, region, customer segment
- ★ **Does this help explain variation?** Helps you understand why patterns occur
Examples: revenue, size, tenure, intensity, score
- ★ **Does this allow comparison?** Helps you contrast groups or time periods
Examples: month, region, provider type, workout type

💡 Universal Pattern: This pattern works for every dataset in the course: exercise logs, candy ratings, HR onboarding, sports, lending, sales, healthcare. It also helps you know what additional variables to collect if you want stronger insights.

21

Visualization Tools Do a Lot Automatically, But You Need to Guide Them

Most modern tools detect levels of measurement and choose continuous or discrete settings automatically. But software is guessing. You understand your context better than the tool. Knowing the underlying concepts ensures the visualization

reflects your intent, not default settings.

💡 **Why this matters:** Tools guess. You decide, based on the question and the context. When a chart looks wrong, it is often because the tool assumed something different than you intended.

What tools typically auto-detect:

- Numbers with decimals → continuous
- Dates → continuous timeline
- Text values → discrete categories
- Whole numbers → could be either (you should specify)

💡 **Best Practice:** Always check how your tool is treating each variable. Look at the axis type, the aggregation, and the sort order, and make sure they match what you actually meant.

22 Start With the Question, Not the Data

One of the most important habits in analytics is letting the question lead the process. Instead of scrolling through data and wondering what to chart, start with what you want to understand. That question then tells you what variables you need, how to transform them, how to classify them, and which visualization best reveals the answer. This mindset shifts you from chart-making to problem-solving.

💡 **Why this matters:** This approach guides all your work from Week 3 onward.

The Question-Driven Workflow



Choose the chart that reveals the pattern



5. Insight → Decision

Interpret and apply meaning

This moves you from chart-creating to problem-solving.

Professional practice: This approach is how analysts, researchers, and data-driven organizations operate. It ensures every visualization has a purpose and every insight drives action.

🔖 Keep this guide handy. You'll reference these concepts throughout the course.