Seneca | SDDS

Business Analytics

Introduction to Data Visualization

BAN140 - Section NBB /NCC

Mufleh Al-Shatnawi, Ph.D., P.Eng.,

Week 2

Week Topics

Previous Week

- Introduction
- Communicating Visually

Current Week

- Basic Data Types
- Understanding Data
- Discrete and continuous data

Basic Data Types

Introduction

- Research questions are ultimately answered using data.
- Data is collected through **observation** and **measurement**.
- Data elements can be classified into different types. **The choice of statistical procedure used to analyze data depends on the data type.**

Levels of Measurement

Levels of measurement include:

- **Nominal**
- **Interval**
- **Ordinal**
- **Ratio**

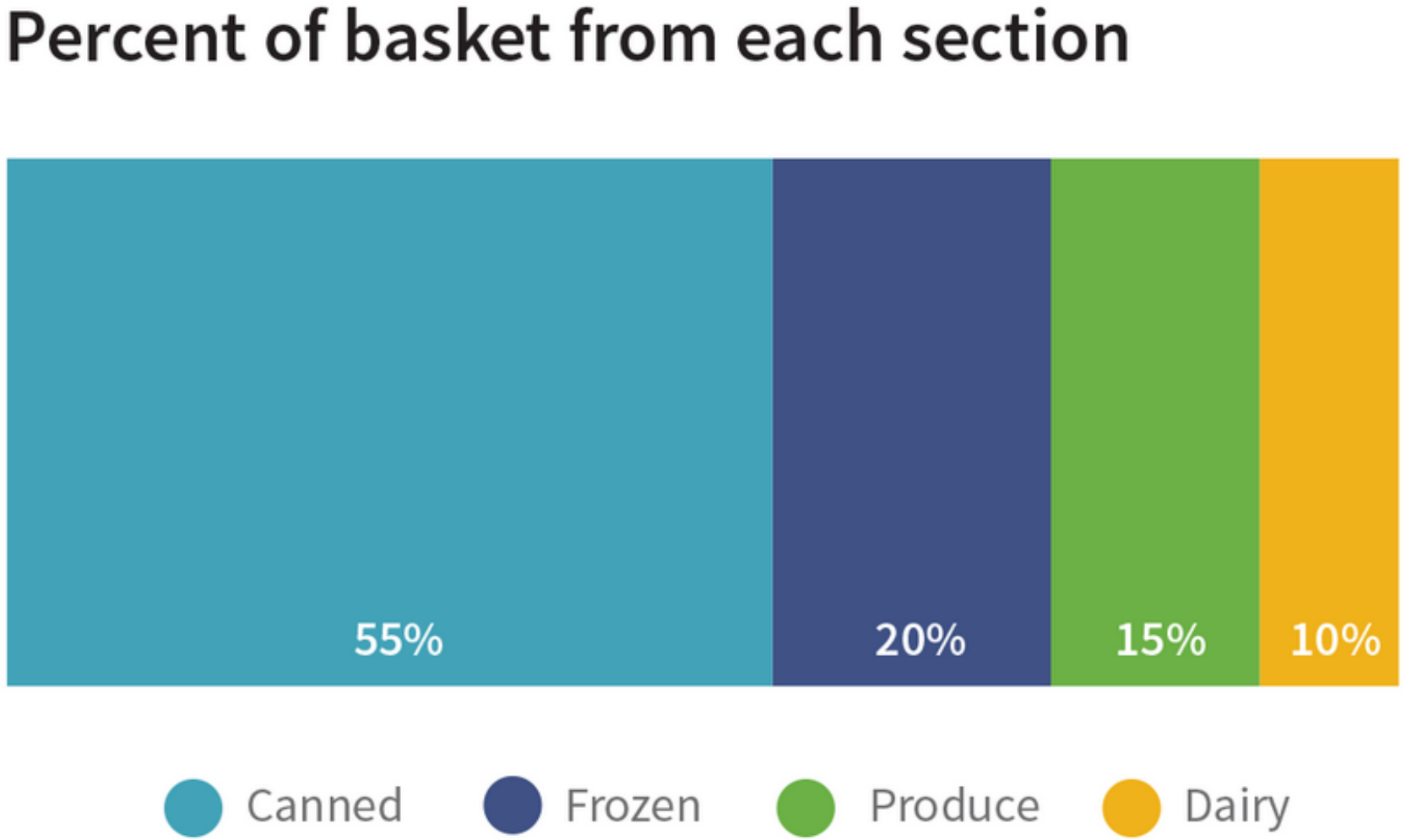
The level determines the amount of information contained in the data.

The level indicates the data summarization and statistical analyses that are most appropriate.

Nominal Data type

- Nominal data type represents data that can be categorized, and it is inherently unordered.
  - Data are labels or names used to identify an attribute of the element
- Nominal data can be counted and used to calculate percentages, but you can't take the average of nominal data.
- When there are only two categories available, the data is referred to as **dichotomous**. The answers to **yes/no** questions are dichotomous data.

Example about Nominal Data Type



Ordinal Data Type

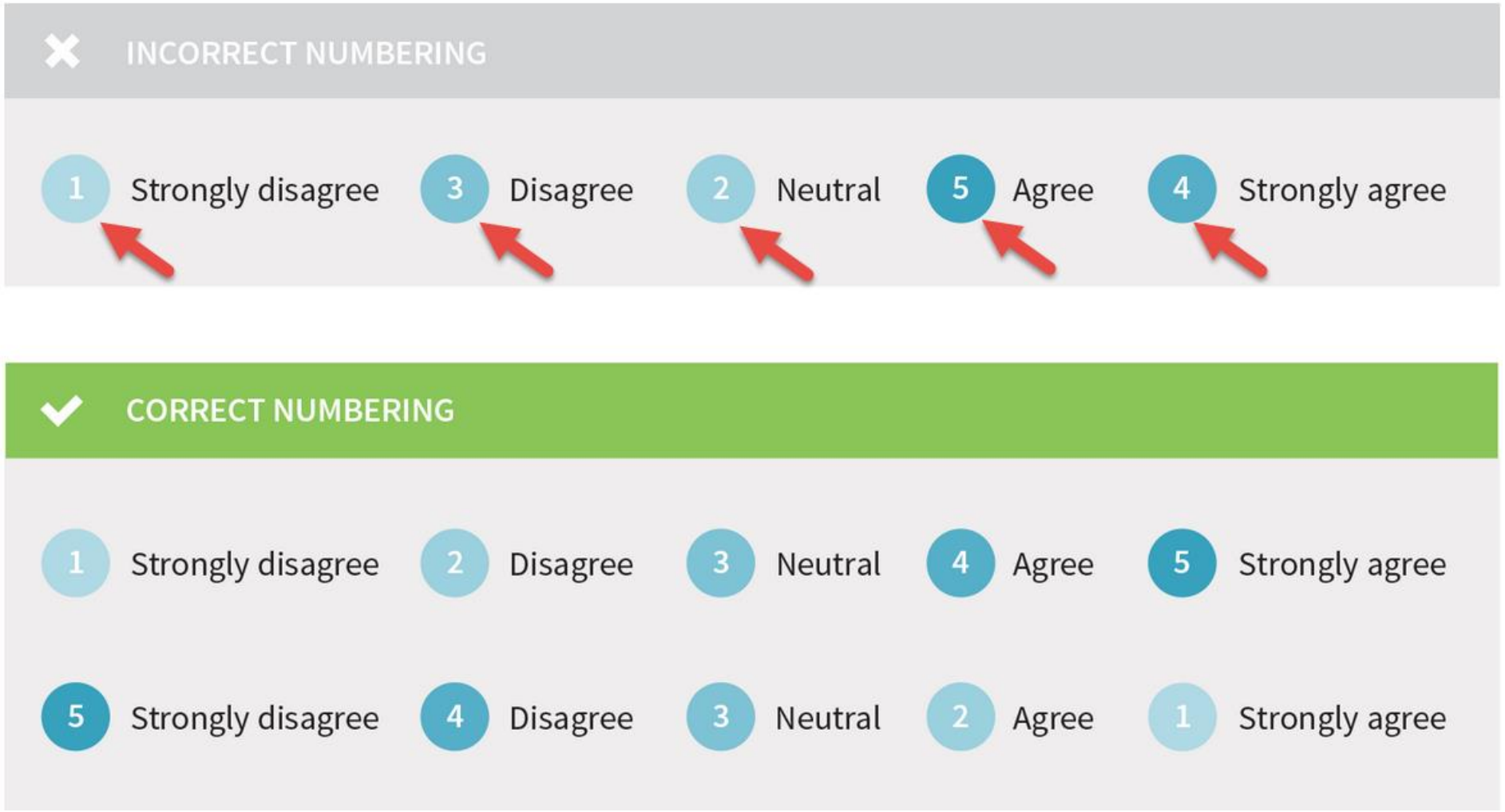
- Ordinal data type represents data has natural ordering to the categories.
  - The data have the properties of nominal data and the order or rank of the data is meaningful.
  - For example: Survey questions that have answer scales like “strongly disagree,” “disagree,” “neutral,” “agree,” “strongly agree” are collecting ordinal data.

More about Ordinal Data Type

- No category on an ordinal scale has a true mathematical value. we can say
  - 1 = strongly disagree, 5 = strongly agree
  - 5 = strongly disagree, 1 = strongly agree
- The numbers you select to represent ordinal categories do change the way you interpret your end analysis, but you can choose any set you wish as long as you keep the numbers in order.

Example about Ordinal Data Type

We can count ordinal data and use them to calculate percentages



Interval Data Type

- Interval data type represents numeric data that you can do mathematical operation on it.
  - The data have the properties of ordinal data, and the interval between observations is expressed in terms of a fixed unit of measure.
  - Any interval between each consecutive point of measurement is equal to every other
- For example:
  - the difference between 11:15 and 11:30 has the exact same value as the difference between 12:00 and 12:15.

More about the interval data type

- Interval data type doesn't have a "meaningful" zero point
  - the value of zero doesn't indicate the absence of the thing you're measuring.
  - For example: 0:00 am isn't the absence of time, it just means it's the start of a new day.

Ratio Data Type

- Ratio data is numeric and have all the properties of interval data, except it *does* have a meaningful zero point.
- **This ration scale must contain a zero value that indicates that nothing exists for the variable at the zero point**
  - For example: zero minutes, zero people in line, zero dairy products in your basket.

Discrete or Continuous Data

Discrete Data

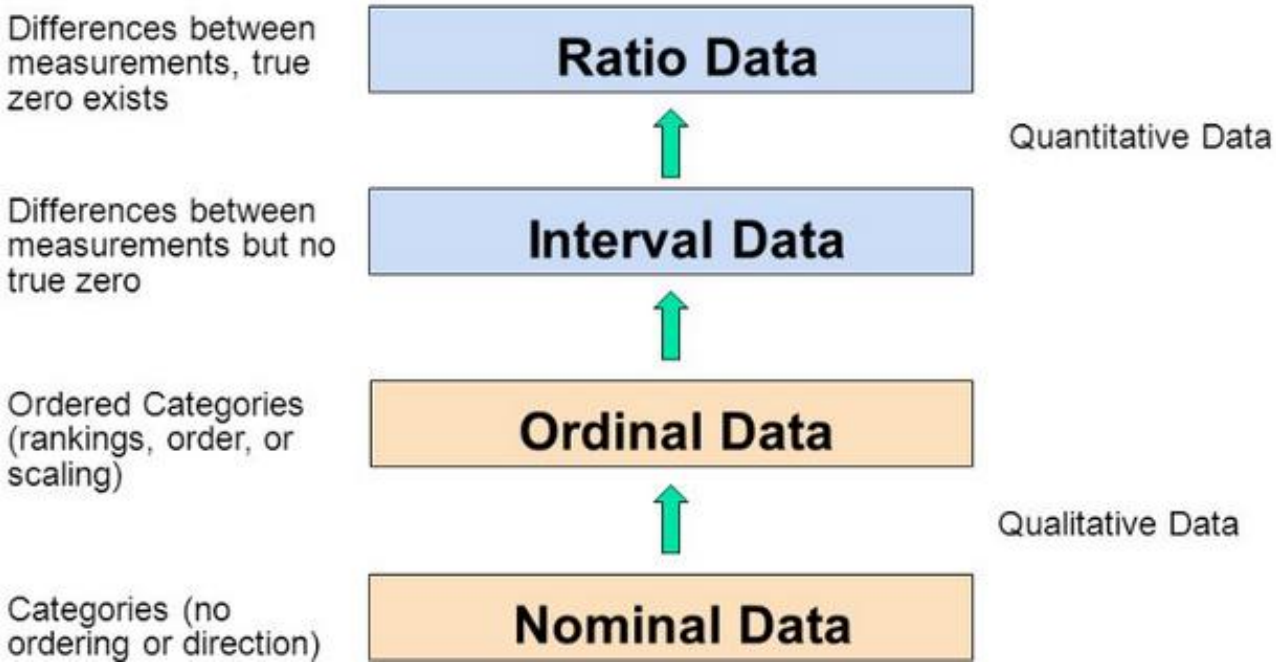
- Discrete means that you can only have specific amounts of the thing you are measuring (typically integers) and no values in between those amounts.
- There must be a whole number of people in line; there can't be a third (1/3) of a person.
- You can have an *average* of, say, 4.25 people per line, but the actual count of people must be a whole number.

Continuous Data

- Continuous means that the data can be any value along the scale.
  - You can buy 1.25 lbs of cheese or be in line for 7.75 minutes.
- This doesn't mean that the data must be able to take all possible numerical values – only all the values within the bounds of the scale.
- You can't be in line for a negative amount of time and you can't buy negative lbs of cheese, but these are still continuous.

Data Type - Summary

- **Nominal** data are used to "*name*," or label a series of values.
- **Ordinal** scales provide good information about the *order* of choices, such as in a customer satisfaction survey.
- **Interval** scales give us the order of values plus the ability to quantify *the difference between each one*.
- **Ratio** scales give us the ultimate– order, interval values, plus the *ability to calculate ratios* since a "true zero" can be defined.



Reference

Data Type - Summary

- In any given study, the data type determine by specifying how the data will be collected
- A single variable can have two different data collection methods therefore two different data types.
- For example:
  - Age can be collected as ratio data or ordinal data. “What age group do you fall in?”

Data Type Rule

- **Ratio is 4th Level of Measurement**
- **Interval is 3rd Level of Measurement**
- **Ordinal is 2nd Level of Measurement**
- **Nominal is 1st Level of Measurement**
- There general rule is that you can go down in level of measurement but not up. 4→3→2→1
- Variables that are naturally ordinal can't be captured as interval or ratio data but can be captured as nominal.

In class Question

A. Nominal B. Ordinal C. Interval D. Ratio

In class Question





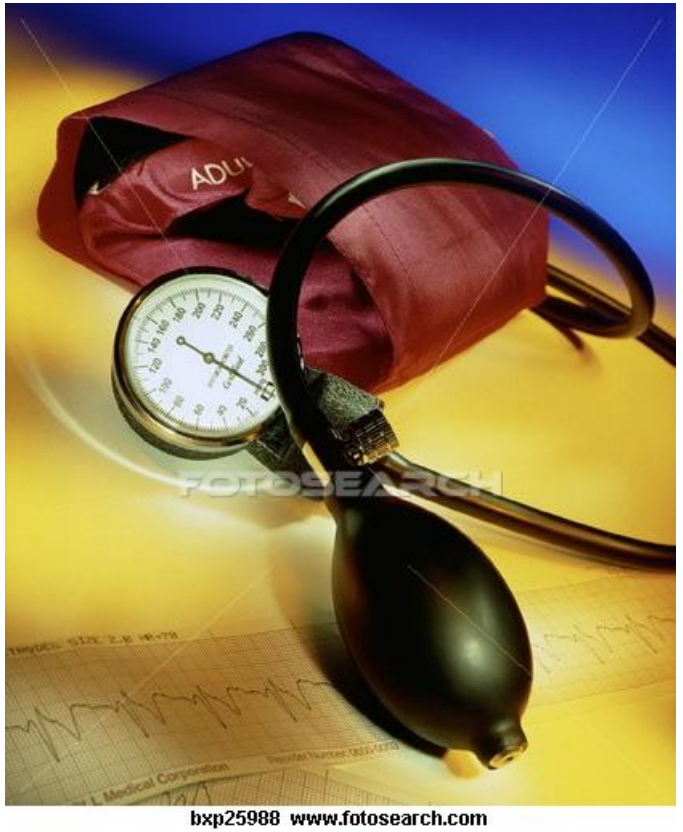

A. Nominal B. Ordinal C. Interval D. Ratio

In Class Question

Finishing order of horse.

A. Nominal B. Ordinal C. Interval D. Ratio

In Class Question



A. Nominal B. Ordinal C. Interval D. Ratio

In Class Question

Age

A. Nominal B. Ordinal C. Interval D. Ratio

What Data Represents

What Data Represents

- Data is more than numbers, and to visualize it, you must know what it represents. **Data represents real life**.
  - It's a snapshot of the world in the same way that a photograph captures a small moment in time.
- A single data point can have a **who**, **what**, **when**, **where**, and **why** attached to it.
  - So, for each single point there are dismissions that can add more annotation about it

More about Data Representation

- Visualization can help detach your focus from the individual data points and explore them from a different angle
- **Interpretation of the data changes based on the visual form it takes on.**
- Computers do a bulk of the work to turn numbers into shapes and colors, but you must make the **connection** between data and real life, so that you or the people you make graphics for extract something of value.

Example

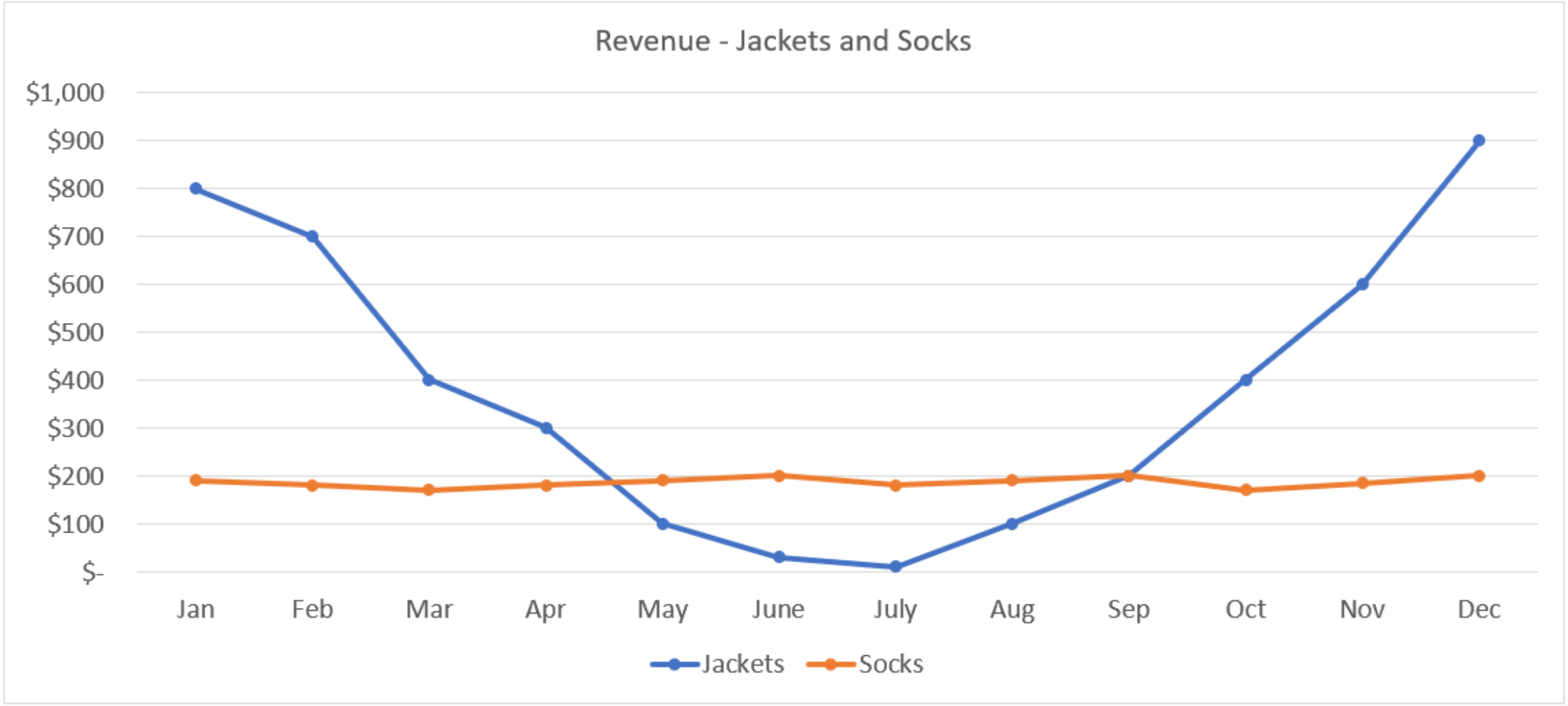
Revenue - Jackets and Socks (Thousands of U.S. $)

	Jan	Feb	Mar	Apr	May	June	July	Aug	Sep	Oct	Nov	Dec
Jackets	$ 800	$ 700	$ 400	$ 300	$ 100	$ 30	$ 10	$ 100	$ 200	$ 400	$ 600	$ 900
Socks	$ 190	$ 180	$ 170	$ 180	$ 190	$ 200	$ 180	$ 190	$ 200	$ 170	$ 185	$ 200

A compare sales of jackets to sales of socks over the course of a year

The table above does an excellent job showing precise if this information is needed. However, it's difficult to instantaneously see trends and the story the data tells

Example


Revenue - Jackets and Socks
$1,000
$900
$800
$700
$600
$500
$400
$300
$200
$100
$-
Jan
Feb
Mar
Apr
May
June
July
Aug
Sep
Oct
Nov
Dec
Jackets
Socks

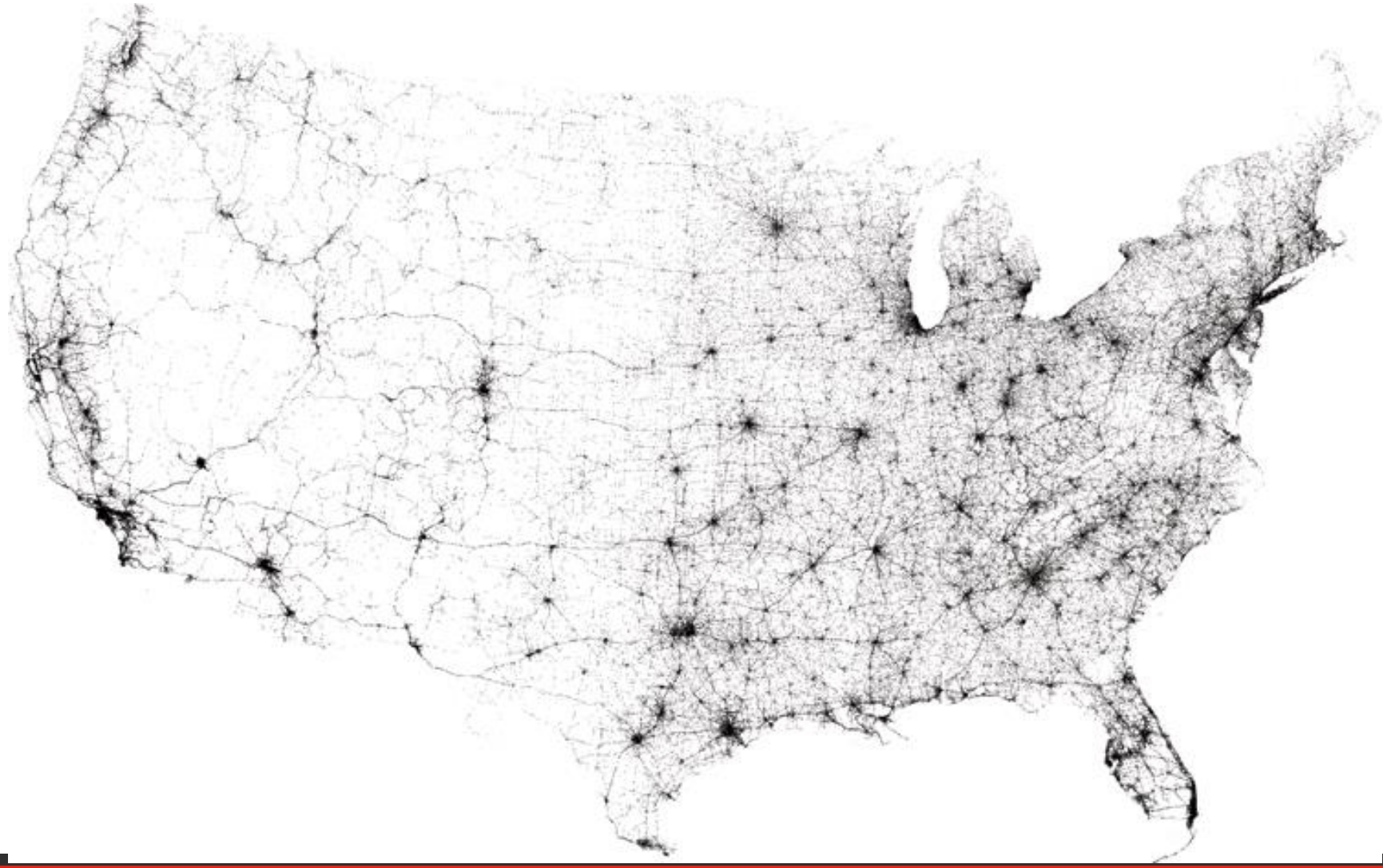
Variability in the Data

- Fluctuations in the data
- Do not look at single value that summarize the data
- With data, you can find patterns, trends, and cycles, but it's not always (rarely, actually) a smooth path from point A to point B.
- Total counts, means, and other aggregate measurements can be **interesting**, but they're only part of the story, whereas the fluctuations in the data might be the most interesting and important part.

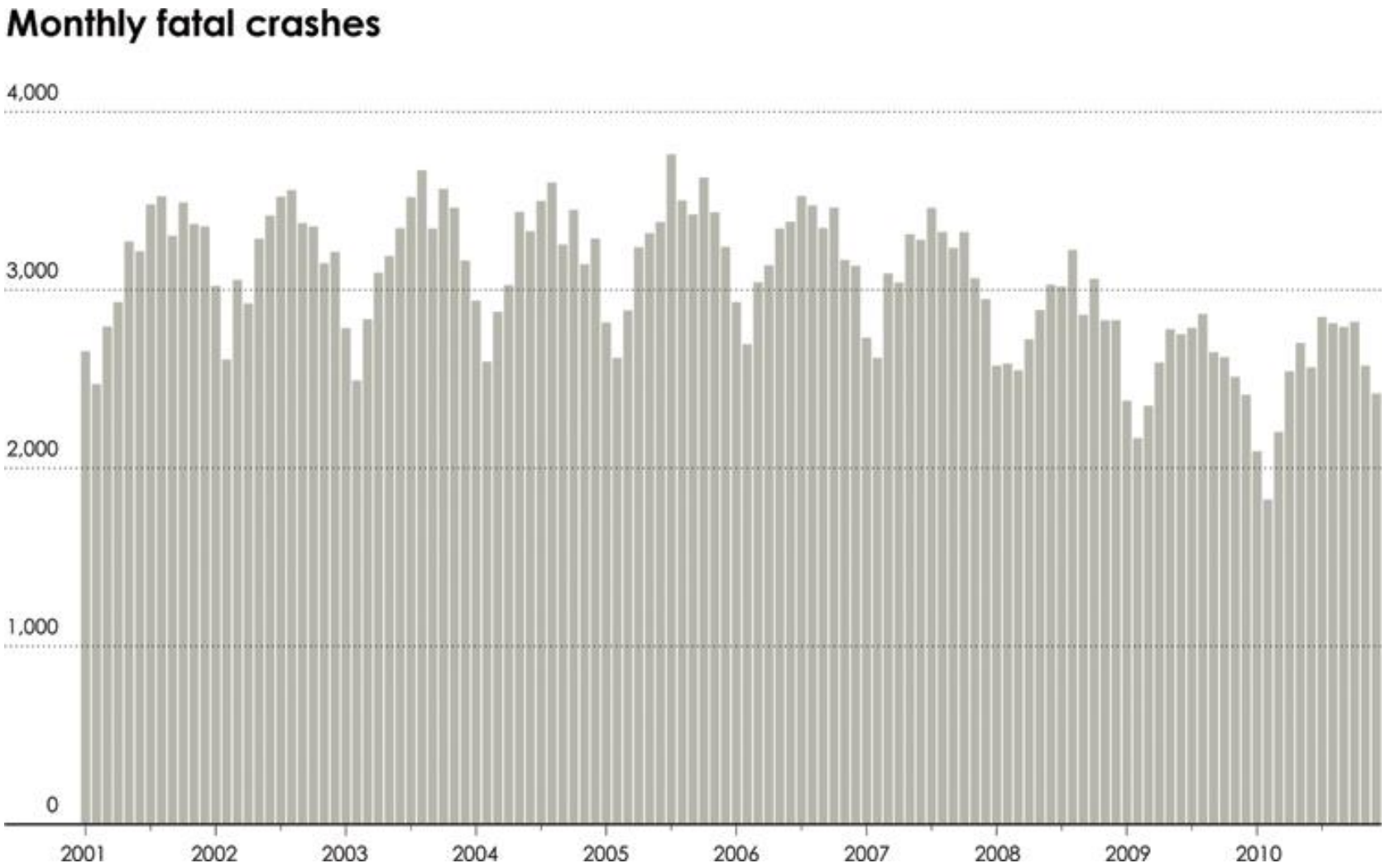
Example

- Between 2001 and 2010, according to the National Highway Traffic Safety Administration, there were 363,839 fatal automobile crashes in the United States.

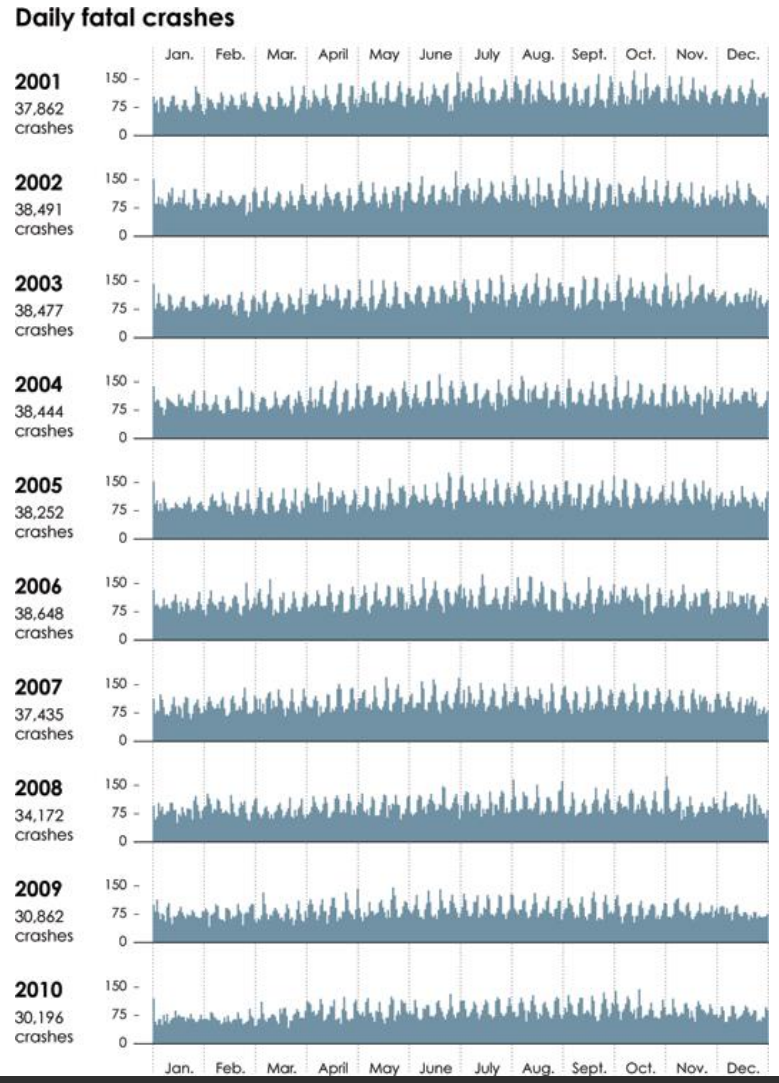
Crashes are Mapped



Crashes vs Months



Can you see a pattern?



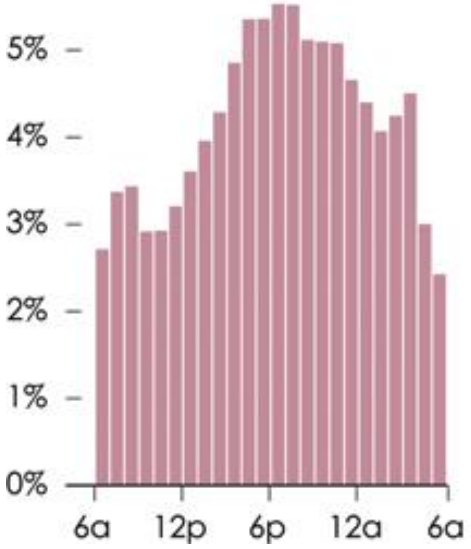
Now, can you see a pattern?

2001–2010

Fatal crashes by…

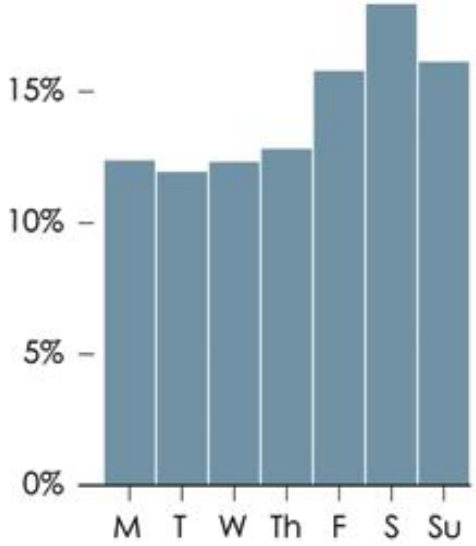
Time of day

Most in the evening and least early morning



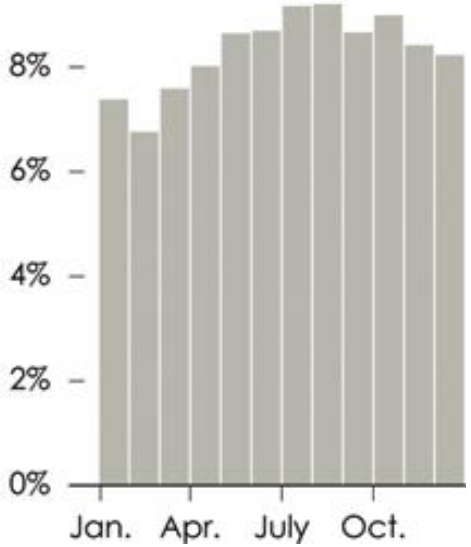
Day of the week

Most on weekends and least middle week



Month

Most in the summer and least during the winter



Variability - summary

- The main point is that there's value in looking at the data beyond the mean, median, or total because those measurements tell you only a small part of the story.
- A lot of the time, aggregates or values that just tell you where the middle of a distribution is hide the interesting details that you should focus on, for both decision making and storytelling.

Uncertainty in the Data

- One of the most challenging aspects of data visualization is the visualization of uncertainty.
- An analyst considers the evidence (such as a sample) and makes an education guess about a full population. That educated guess has uncertainty attached to it.
  - When we see a data point drawn in a specific location, we tend to interpret it as a precise representation of the true data value.
  - Two commonly used approaches to indicate uncertainty are **error bars** and **confidence bands**.

Uncertainty exist in our life

- Take the weather for example. How many times have you looked up the forecast for the next day or for the next week as you pack for a trip, only to find, when the time comes, that the weather isn't how you expected it to be?
- The subway announcement says a train will arrive in 10 minutes, but it comes in 11, or a delivery is estimated to arrive on Monday, but it comes on Wednesday instead.

Question you should ask

- How well does a sample represent a full population?
- How likely is it that a dataset represents the truth?
- How much do you trust the numbers?
- You should always wonder about the uncertainty.

Context in the Data

- Context can completely change your perspective on a dataset, and it can help you decide what the numbers represent and how to interpret them.
- After you do know what the data is about, your understanding helps you find the fascinating bits, which leads to worthwhile visualization.
- You must know the **who**, **what**, **when**, **where**, **why**, and **how**—the metadata, or the data about the data—before you can know what the numbers are about.

Context is important

- Without context attached to your data visualizations, you simply can't make the best business decisions.
  - Remember that when it comes to business decisions, the best decision can only be the outcome of having the **most accurate data** and that data needs to be framed in the bigger picture. ***It needs context.***
- Context comes in the form of text.
  - labeling your axes and
  - providing color keys, to using data point labels and annotations on the visualization or
  - ***adding explanatory paragraphs in an article.***
- **Without these, the data visualization can be meaningless and useless**

References and Resources

- [Knaflic] Cole Nussbaumer Knaflic, **Storytelling with Data: A Data Visualization Guide for Business Professionals**, Wiley, 2017
  - Available online through Seneca Libraries: https://senecacollege-primo.hosted.exlibrisgroup.com/permalink/f/t3376v/01SENC_ALMA5146374280003226
- [Ryan] Lindy Ryan, **Visual Data Storytelling with Tableau**, Pearson Addison-Wesley, 2018
  - Available online through Seneca Libraries: https://senecacollege-primo.hosted.exlibrisgroup.com/permalink/f/t3376v/01SENC_ALMA5167006190003226
- [Healy] Kieran Healy, **Data Visualization: A Practical Introduction**, Princeton University Press, 2018.
  - Available (hardcopy) at Seneca Libraries: https://senecacollege-primo.hosted.exlibrisgroup.com/permalink/f/t3376v/01SENC_ALMA2172469250003226
- **A Reader on Data Visualization:** https://mschermann.github.io/data_viz_reader/
- **Data visualization:** https://en.wikipedia.org/wiki/Data_visualization