Validity and Causality

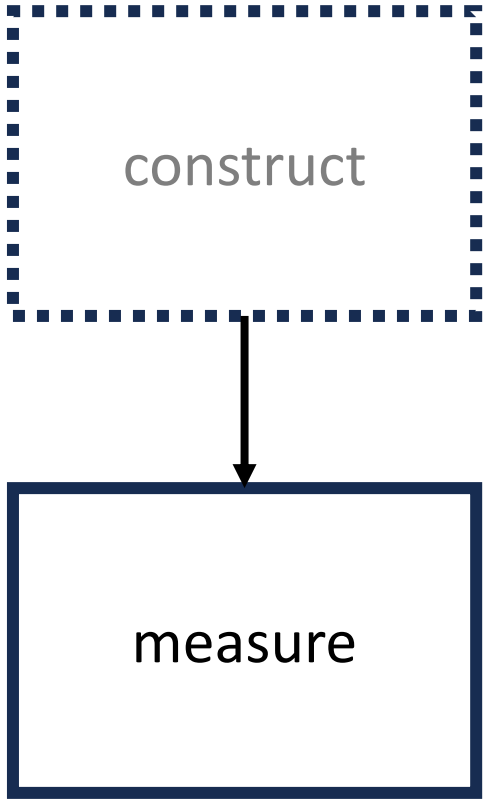
ACC 6300
Advanced Data Analytics

Mason Snow, PhD

What is accounting?

Accounting = measurement

Accountants work with economic constructs and distill them into measures. These measures should be comparable, consistent, decision-useful, and perhaps most importantly – VALID.



Constructs are what “count” and measures are what get counted. As data analysts, we only ever work with measures.

What are the four types of validity?

Internal Validity: does X cause Y?

External Validity: would an observed effect generalize to other settings?

Construct Validity: does the measure sufficiently capture the construct?

Statistical Conclusion Validity: have proper analytical procedures been followed?

Causality is difficult to infer

What are three prerequisites for causal inference?

Temporal precedence: (x must occur before y)

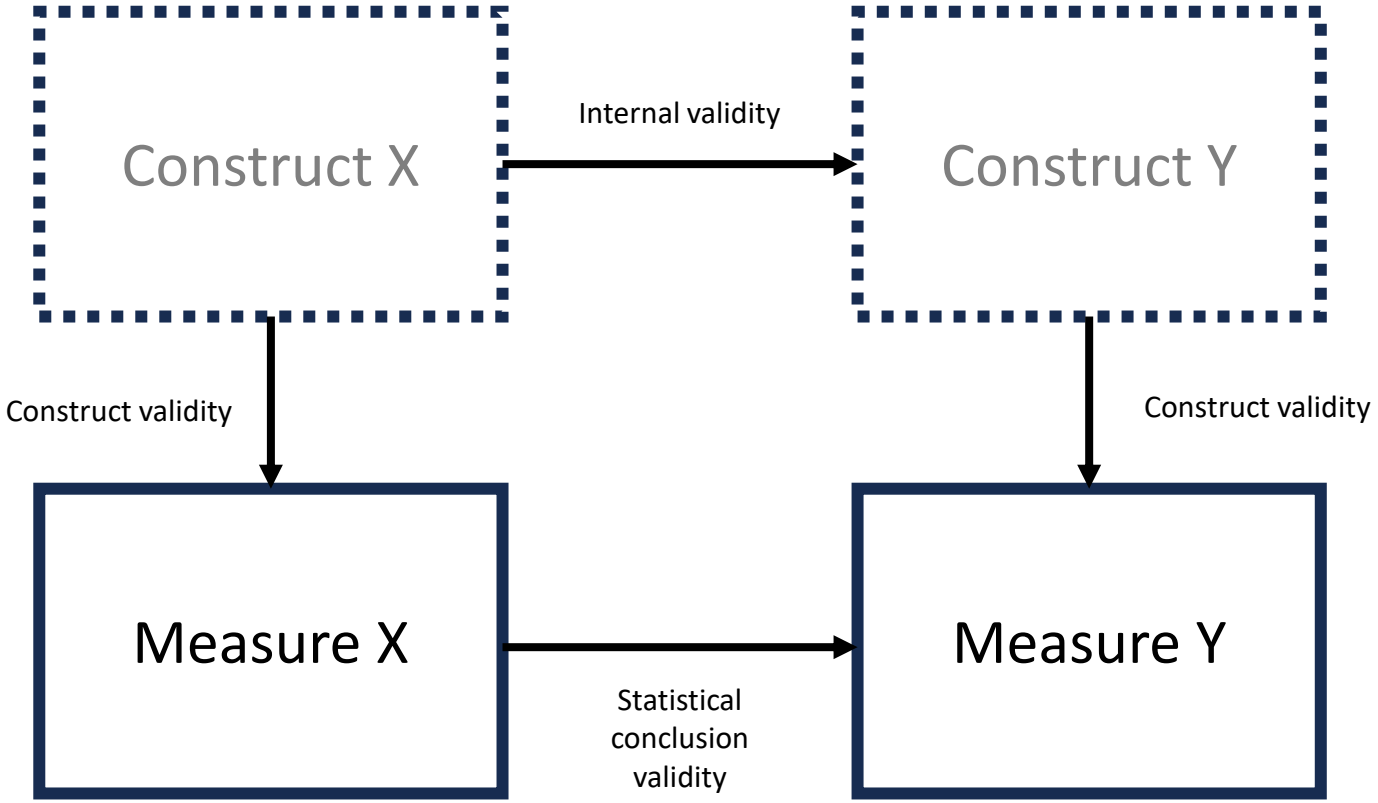
Significant correlation: (x must be related to y)

Alternative explanations must be eliminated ← the difficult part

Five Groups of Alternative Explanations

- **Correlated Omitted Variables**: Assuming that X causes Y when really Z causes both X and Y.
- **Reverse Causality**: Assuming that X causes Y when really Y causes X.
- **Selection Bias**: when the sample of individuals analyzed systematically differs from the population.
- **Measurement Error**: when values of measured observations randomly/systematically differ from true values.
- **Spurious Correlation**: when the correlation between X and Y is really just due to chance.

Libby Box Framework



Weakness in any of these links can cast doubt on a causal relation between X and Y.

Day 2 – Data Modeling and SQL

ACC 6300
Advanced Data Analytics

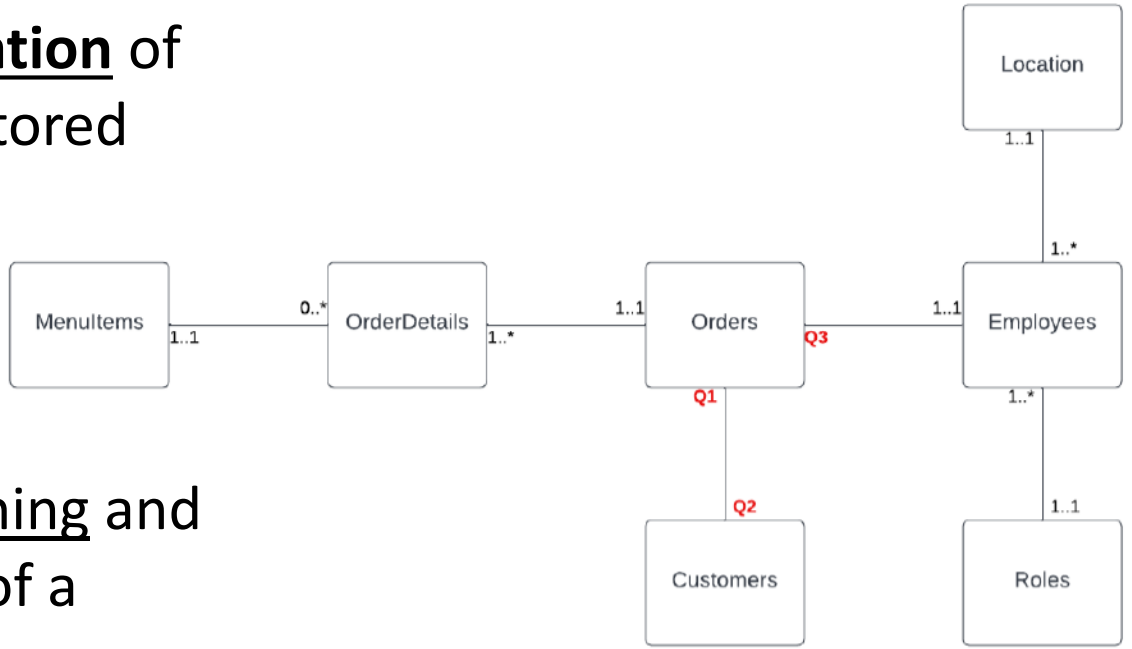
Mason Snow, PhD

What is Data Modeling?

The process of creating a **visual representation** of information systems to show how data is stored and connected within a system.

Why?

Data models serve as a blueprint for designing and communicating the information structure of a system.



Relational Data Models

Relational databases dominate the landscape of database management systems.

Class	Attributes	Associations
• A collection of things about which an organization wants to collect and store data	• The specific facts or dimensions of a class for which we will collect and store data	• A formally stated or acknowledged relationship between two classes.

Classes and Attributes

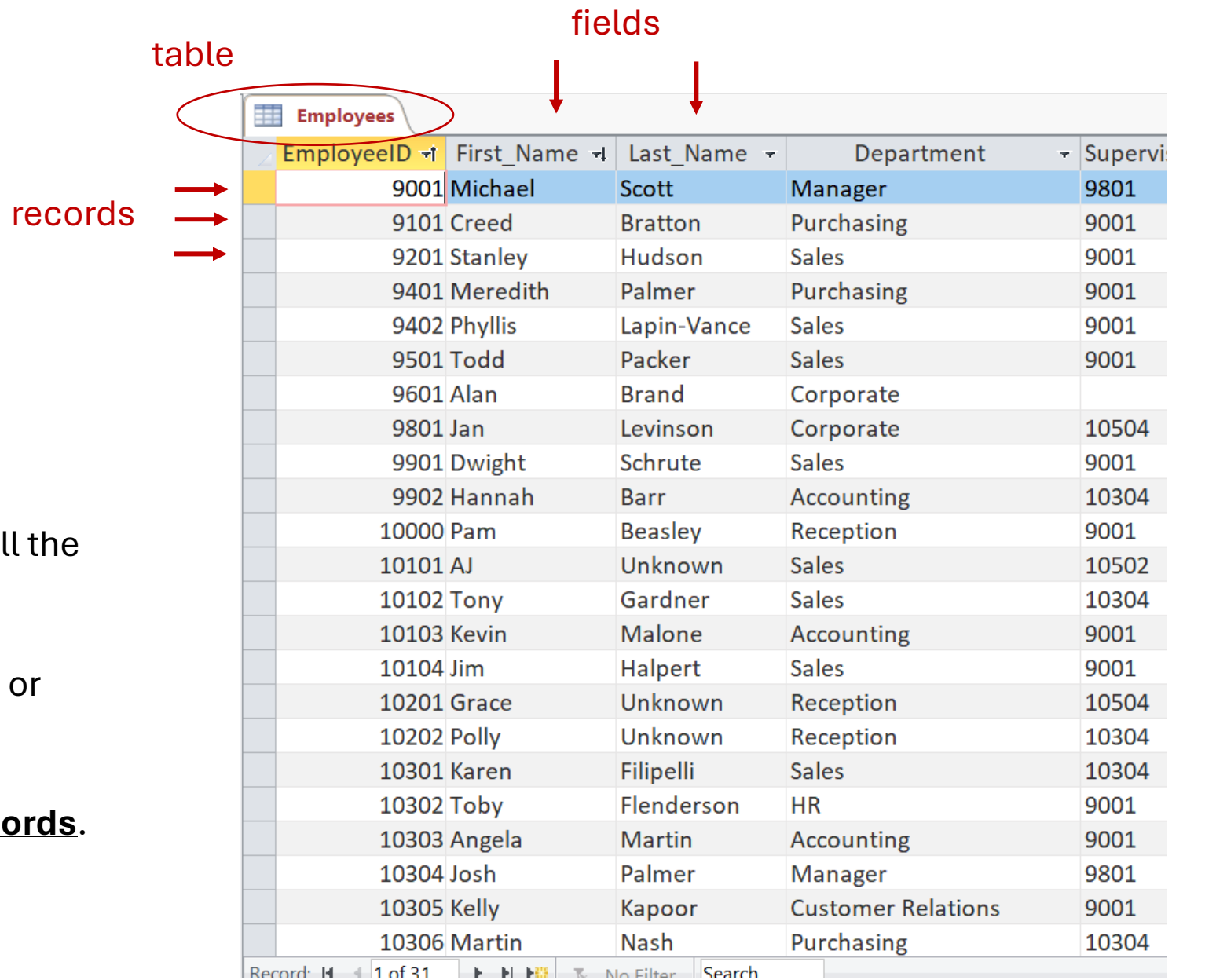
Employees

- EmployeeID
- FirstName
- LastName
- Department
- SupervisorID
- Etc.

The **class** refers to the entire **table** (all the columns and rows)

Attributes are reflected as the **fields** or columns of the table.

Rows of a table are referred to as **records**.

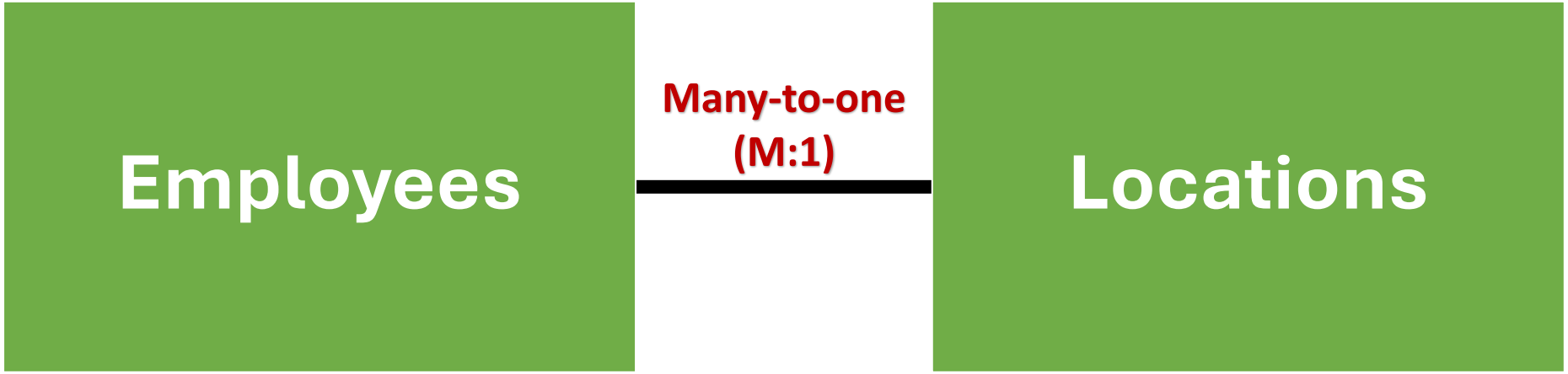


EmployeeID	First_Name	Last_Name	Department	Supervi
9001	Michael	Scott	Manager	9801
9101	Creed	Bratton	Purchasing	9001
9201	Stanley	Hudson	Sales	9001
9401	Meredith	Palmer	Purchasing	9001
9402	Phyllis	Lapin-Vance	Sales	9001
9501	Todd	Packer	Sales	9001
9601	Alan	Brand	Corporate	
9801	Jan	Levinson	Corporate	10504
9901	Dwight	Schrute	Sales	9001
9902	Hannah	Barr	Accounting	10304
10000	Pam	Beasley	Reception	9001
10101	AJ	Unknown	Sales	10502
10102	Tony	Gardner	Sales	10304
10103	Kevin	Malone	Accounting	9001
10104	Jim	Halpert	Sales	9001
10201	Grace	Unknown	Reception	10504
10202	Polly	Unknown	Reception	10304
10301	Karen	Filipelli	Sales	10304
10302	Toby	Flenderson	HR	9001
10303	Angela	Martin	Accounting	9001
10304	Josh	Palmer	Manager	9801
10305	Kelly	Kapoor	Customer Relations	9001
10306	Martin	Nash	Purchasing	10304

Record: 1 of 31 No Filter Search

Associations

Dunder Mifflin is an office supplies distributor with employees working at one of five regional branch locations.



Primary Keys

An attribute that uniquely identifies every instance in a class (i.e., a row of a table).

- Each record must have a primary key
- Values for a table's primary key must NOT repeat.

Natural primary keys are derived from existing, real-world data:

- Social Security Numbers (SSNs)
- Phone Numbers
- Vehicle Identification Numbers (VINs)

Often, the best primary keys are assigned, sequential numbers:

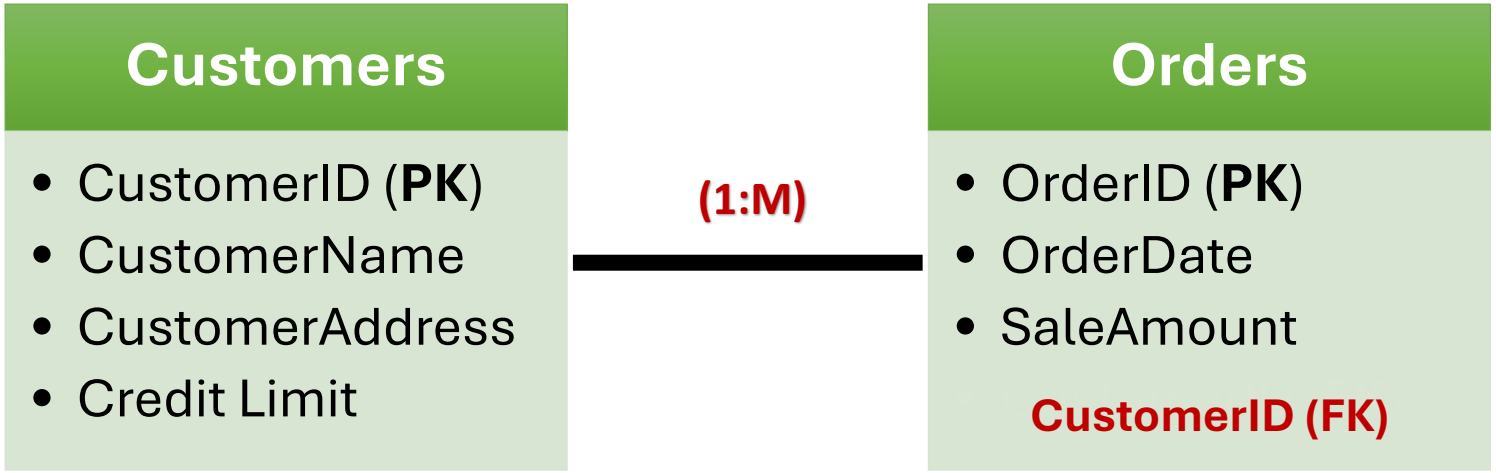
- E.g.) CUST-10023

Foreign Keys

A field in one table that is the primary key of another table in the database.

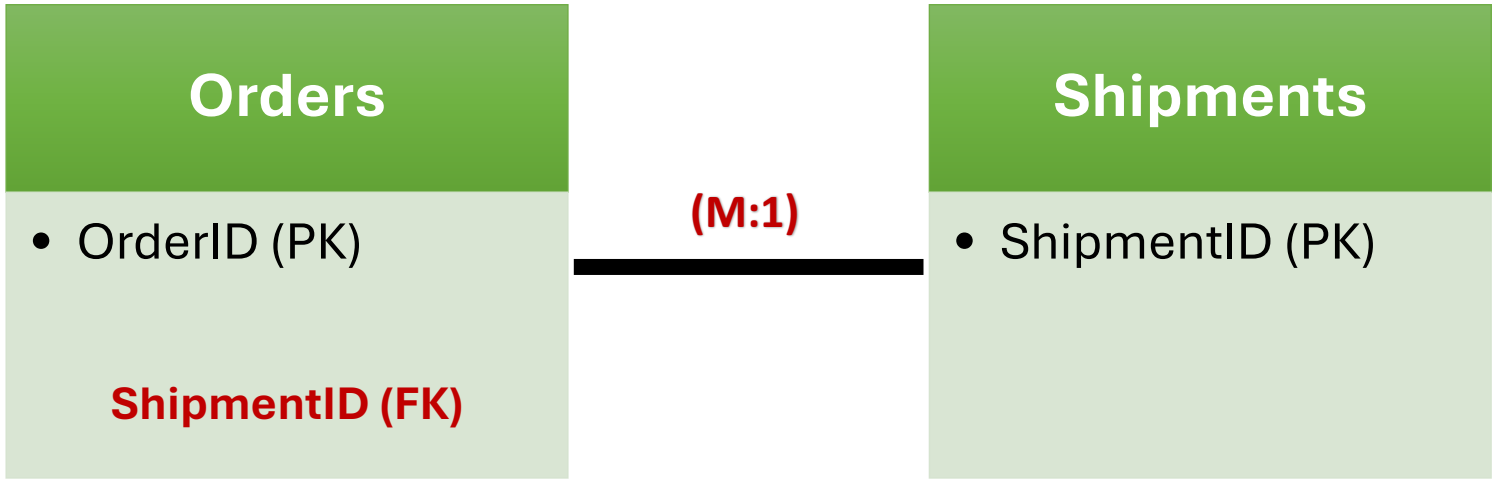
This is what links tables together!

The table that is the “many” side of a one-to-many relationship gets the foreign key.

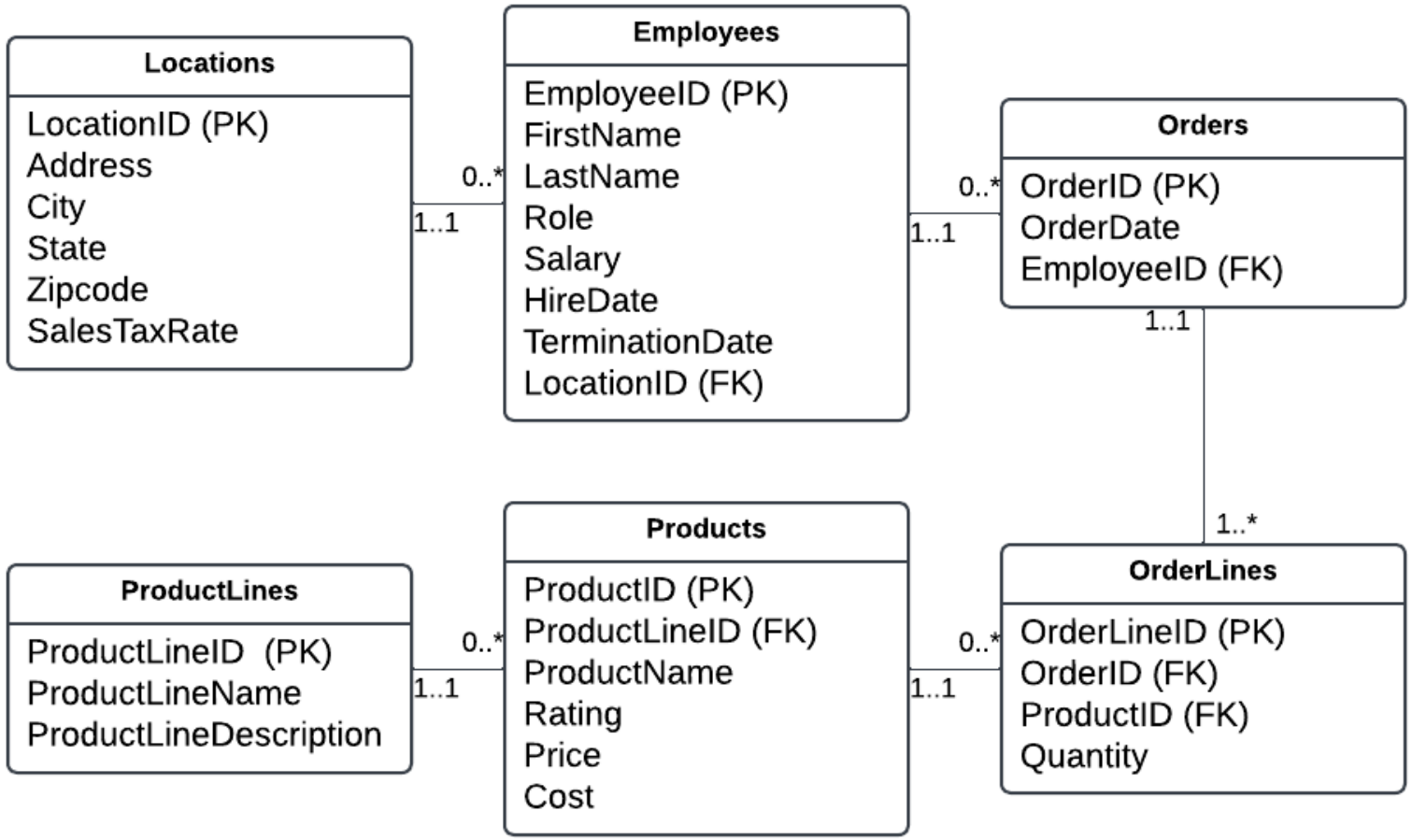


One More Example

A logistics company manages customer orders for various retailers. A given shipment likely relates to several orders. Each order can be linked back to one single shipment.



Getting to know the data



Query Template (multiple tables)

```
SELECT   [table1.field], [table2.field]
FROM     [table1] {JOIN} [table2]
ON       [table1.field] = [table2.field]
WHERE    [field] {meets criteria}
GROUP BY [field]
ORDER BY [field];
```

SQL Skills Checklist

- Write simple SQL Statements using:
  - **SELECT**
  - **FROM**
  - **WHERE**
  - **ORDER BY**
- Aggregate data and use calculated fields:
  - **Count, Sum, Avg**
  - **GROUP BY**
- Query across multiple tables using joins.

SQL Syntax

SELECT

- operator used to begin a query
- this statement specifies which columns/attributes should ultimately be returned
- type the column names individually separated by commas.
- to select all fields in a table, use an asterisk (*).
- SQL commands, but you will often see them in all caps
- SELECT DISTINCT returns unique instances of fields selected.

Practice

SQL Syntax

FROM

- clause added after the **SELECT** command.
- designates from which tables data is to be retrieved.
- the indicated table *must* contain the columns requested in the **SELECT** command (duh).
- it is best practice (but not required) to type each clause in a new line.
- Semicolons are used to indicate the end of an individual query (not necessary if running singular queries at a time).

SQL Syntax

WHERE

- Use to impose criteria in a query.
  - restricts results to only records that meet the designated criteria.
  - impose multiple criteria by separating with “AND” or “OR”

SQL Syntax

WHERE (Examples)

Exact Match (Text)

WHERE OrderID = “PO-01848”

Exact Match (Numeric)

WHERE Quantity = 2

Not Equal (Text)

WHERE FirstName != “Roger”

SQL Syntax

WHERE (Examples)

Greater Than / Less Than (Numeric/Date)

WHERE Salary > 75000

Null Values (Text/Numeric/Date)

WHERE TerminationDate Is Null

SQL Syntax

ORDER BY

- used to sort the data.
- by default, assumes to sort ascending, but you can manually select this by typing “ASC” after the column(s).
- to sort descending, type “DESC” after the column(s).

Calculated Fields

To create a new field based on a calculation:

- Within the SELECT command line, do the calculation followed by "AS" and whatever you wish to name the new field.
- We will only use Count, Sum, Avg, Min, and Max

Examples:

```
SELECT    Avg(SalesTaxRate) AS AverageRate
SELECT    Sum(Cost) AS TotalCost
SELECT    Count(CustomerID) as CustomerCount
```

GROUP BY

To perform a calculation within certain groups:

- Use the **GROUP BY** clause and specify the desired level of aggregation (i.e., the field name(s))
- **GROUP BY** comes before **ORDER BY** but after **WHERE**

Joins

- Join clauses combine rows from two or more tables in SQL queries.
- Multiple tables are indicated in the FROM clause.

  On Table.Field = Table.Field

- Use matching values between tables' columns (primary and foreign keys).

- For now, we will only use JOIN (inner joins)

Day 3 – Advanced Topics in Excel

Please log in to Canvas and download AdvExcel.xlsx as you arrive

ACC 6300
Advanced Data Analytics

Mason Snow, PhD

What is the role of descriptive statistics?

What is ‘dispersion?’

How is it often measured?

Why should we care?

What is the normal distribution?

What’s so great about it?

Apply Wheelan's example about assessing school quality to the predictive validity framework.

Today’s topic:
“Advanced” Topics in MS Excel

What are these topics?

- Dynamic array functions
- Pivot Tables (yes, Pivot Tables)
- Analyze Data
- Power Pivot (Data Model)
- Lambda functions

Dynamic Arrays

- Introduced with Excel 2021
- A dynamic array is a feature in Excel that allows a single formula to return multiple values in multiple cells.
  - These typically “spill” into adjacent cells.

- Why? They automatically expand/contract to fit results, which often eliminates the need for manual tasks (copying formulas, performing sorts, etc.)

Dynamic Array Functions

- **UNIQUE**
  Returns a list of unique values from a range, eliminating duplicates.
- **SORT**
  Sorts data in ascending or descending order automatically.
- **SEQUENCE**
  Generates a sequence of numbers, with an optional starting value and step increment.

Dynamic Array Functions

- **XLOOKUP**
  Searches for a value in a range and returns the corresponding value from another range, with additional options for matching and handling missing data.
- **TEXTSPLIT**
  Splits text into multiple cells based on a delimiter, with an option to ignore empty results.
- **FILTER**
  Filters data based on a specified condition and returns the matching results.

Analyze Data

- Newly introduced button in Excel.

- Described as offering “users a quick way to generate insights and summaries from data using AI-powered features. Complex formulas or advanced knowledge of data analysis tools is no longer necessary”

- Prof Snow’s opinion: not that useful today, but still a preview of how we will likely interact with most platforms in the future.

Power Pivot / Data Model

- Power Pivot is a COM add-in for Excel that enables relational data modeling and analysis.

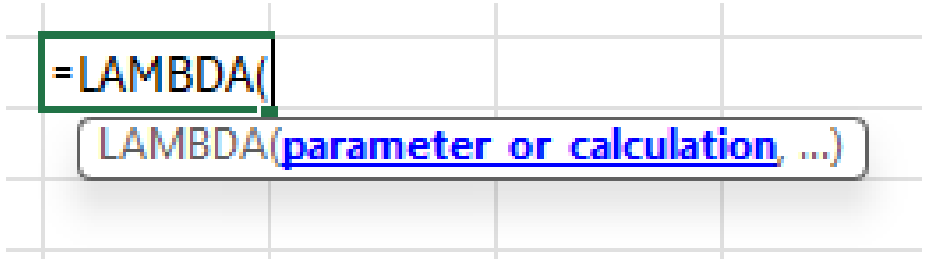
- Features:
  - Building data models in Excel using internal/external data sources (i.e., can connect to tables in local databases, servers, etc.)
  - This enables Excel to analyze larger datasets
  - Multi-table calculations using DAX formulas

Lambda functions

- Lambda functions allow users to create custom, flexible functions within Excel.

- These functions can be stored in the Name Manager and invoked like any other built-in function in excel.

Lambda functions – How To Implement



1. List out the variables or parameters that will be used in a calculation, separated by parentheses.
2. Once all parameters are indicated, provide a calculation that references the parameters.
3. To immediately perform the calculation:
   a) close the parentheses
   b) open a new set of parentheses and provide values for the parameters, separated by commas.
4. To save as a new custom function, copy the formula through the parameters and calculations and paste in the Names Manager in the Formulas tab.

Day 4 – ETL Process Overview

Please log in to Canvas and download and extract the ETL.zip folder as you arrive

ACC 6300
Advanced Data Analytics
Mason Snow, PhD

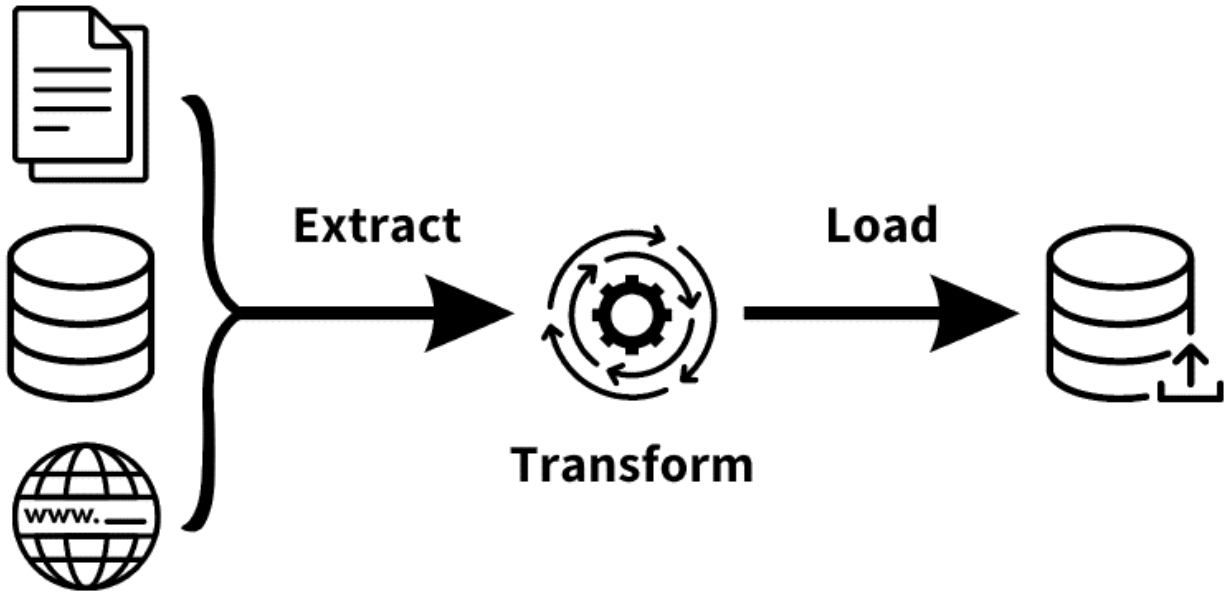
AMPS Model Revisited

ETL Process – Big Picture

Consists of all the activities needed to prepare the data for analysis. This could include:

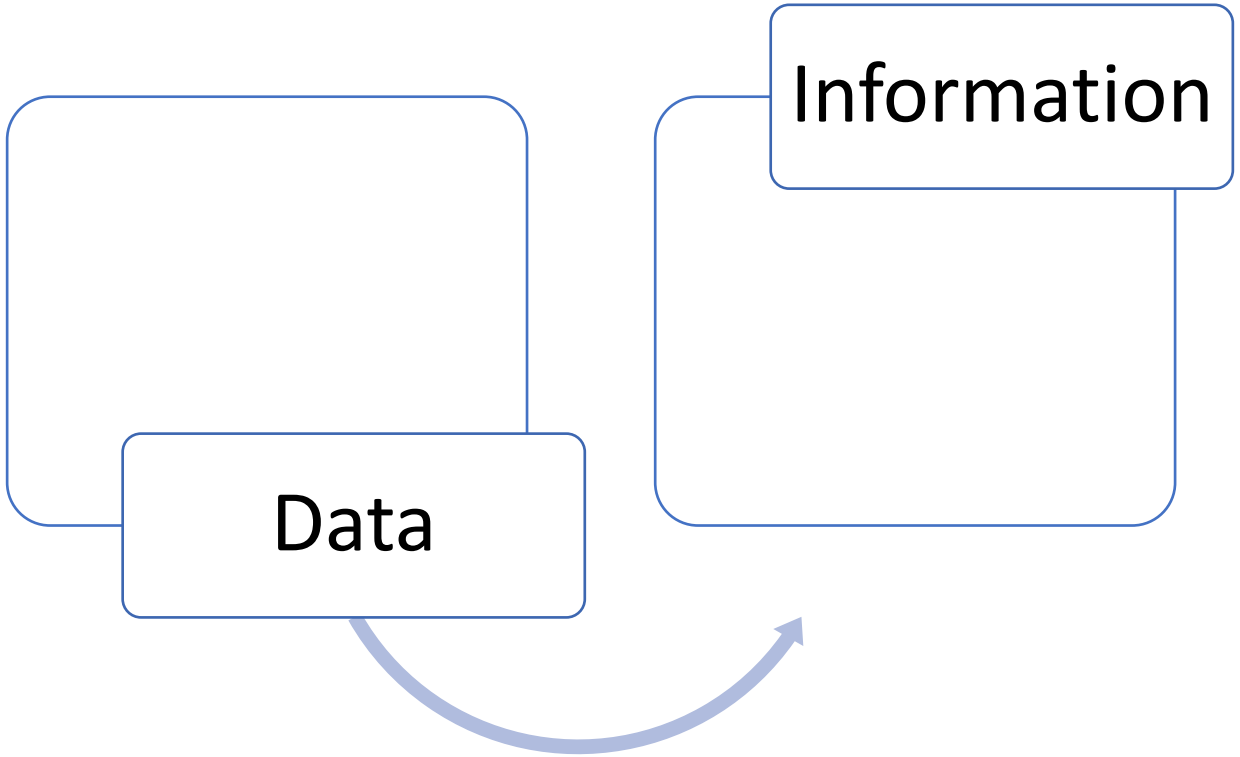
collection, cleaning, combining, structuring, transformation, profiling, formatting, etc.

Often the most time-consuming part of the whole data analytics process.

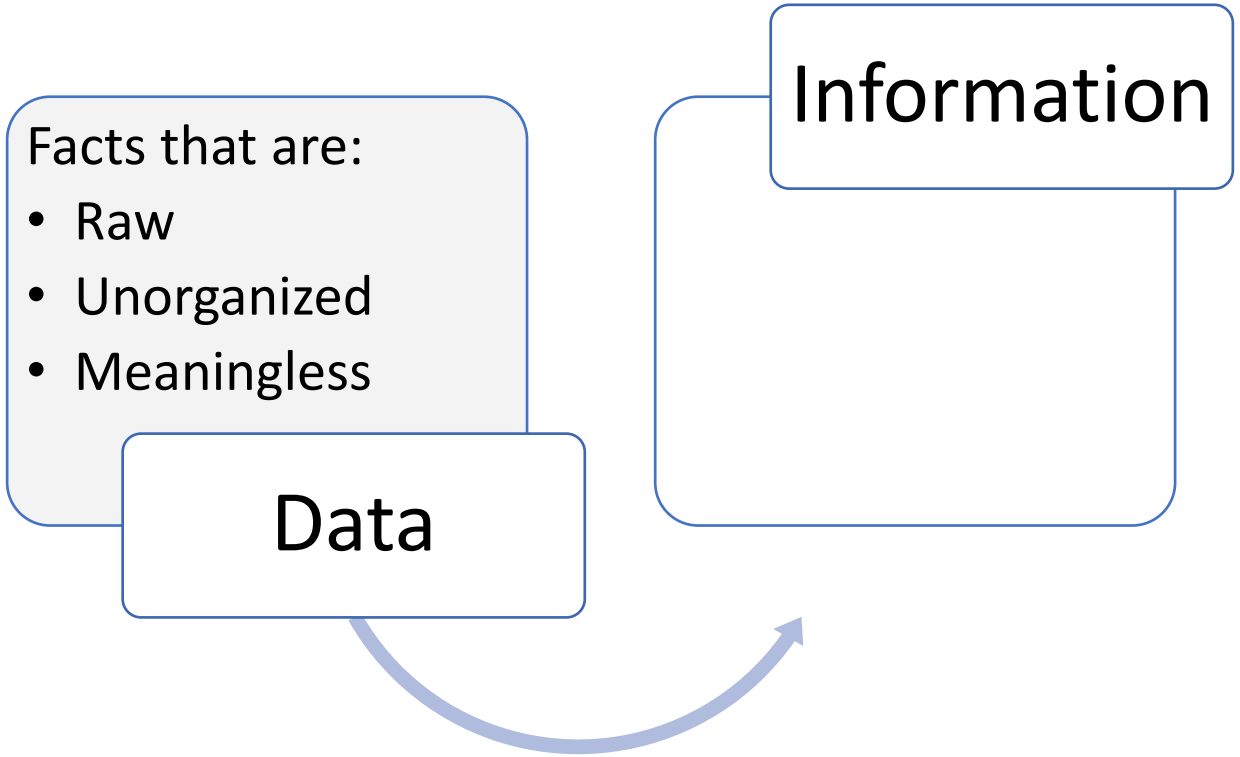


Often referred to as the “ETL” process.

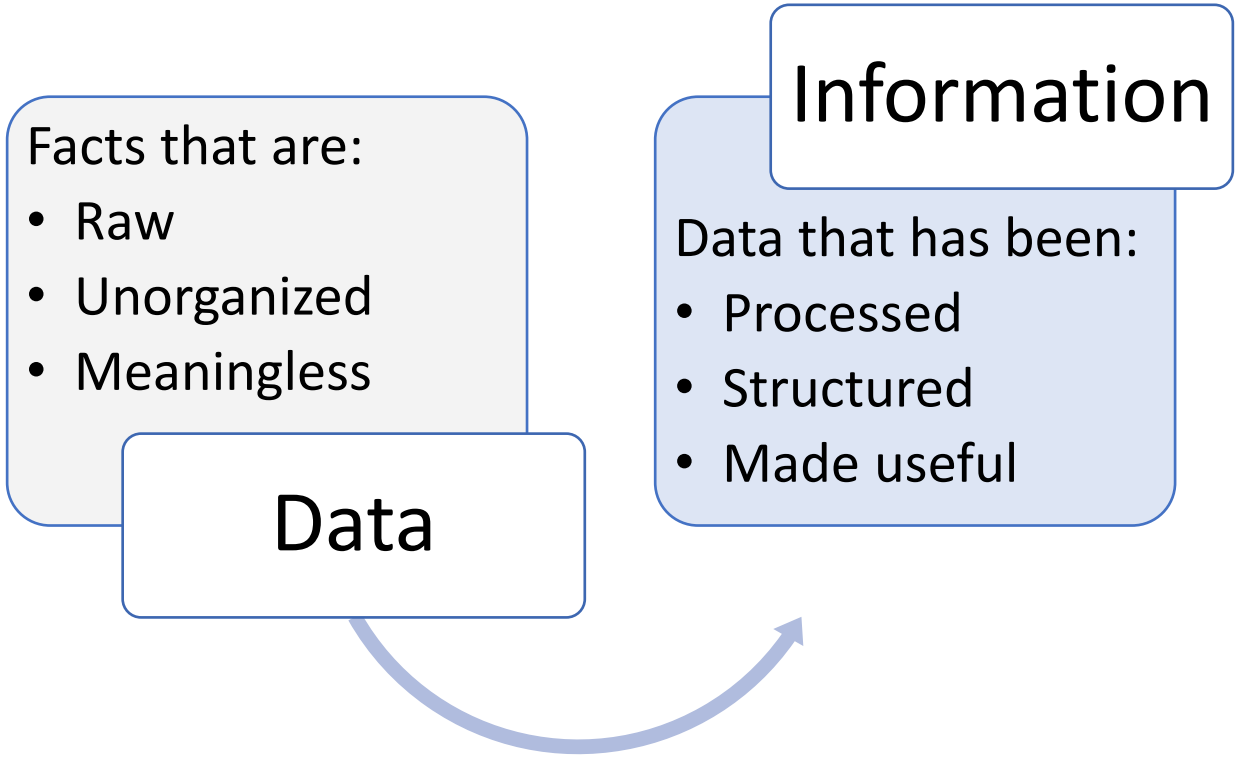
Data vs Information



Data vs Information



Data vs Information



High Quality Information

Characteristic	Description
Accurate	Correct; free from error; accurately represents underlying facts
Available	Accessible to users when needed
Complete	Does not omit needed data; sufficient in depth and breadth
Consistent	Presented in the same format every time
Current	Includes data that is up-to-date to the present
Objective	Unbiased; impartial
Relevant	Appropriately pertains to the requested situation
Timely	Provided in time for users to make decisions
Understandable	Easily comprehended and communicated
Verifiable	Can be checked/confirmed by others; supported with documentation.

Day 5 – Workflow Automation

Please log in to Canvas and download and extract the ETL2.zip folder as you arrive

ACC 6300
Advanced Data Analytics

Mason Snow, PhD

Power Query Skills Checklist

Where we've been:

- Connect to data
- Profile data
- Filter data
- Add new columns
- Replace values
- Aggregate data
- Append data

Where we're headed today:

- Restructure data
- Join data
- Setting up for automation

Data Structuring

Transposing

- Switching the rows and columns of the data

Pivoting

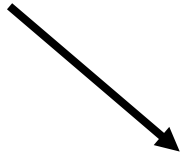
- Reshaping the data from tall to wide format. The unique values of one column are converted to columns, values are specified for aggregation.

Unpivoting

- Reshaping the data from wide to tall format. Takes multiple related columns and transforming them into a single column of values.

Transposing

County	Utah	Salt Lake	Davis	Summit
Jan	141	126	114	148
Feb	111	61	103	131
Mar	144	110	153	145
Apr	112	172	123	117
May	95	126	137	73
Jun	105	109	106	150



Changing the orientation of the dataset, switching columns and rows

County	Jan	Feb	Mar	Apr	May	Jun
Utah	141	111	144	112	95	105
Salt Lake	126	61	110	172	126	109
Davis	114	103	153	123	137	106
Summit	148	131	145	117	73	150

Pivoting

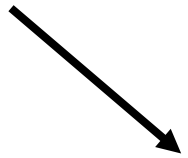
Date	Product	Sales
1/1/2024	A	$100
1/1/2024	A	$120
1/1/2024	B	$150
1/2/2024	A	$200
1/2/2024	B	$100
1/2/2024	B	$130

Converting from tall to wide format, aggregating repeat values for a designated column.

Date	A	B
1/1/2024	$220	$150
1/2/2024	$200	$230

Unpivoting

Country Name	2016	2017	2018
Argentina	0.7	0.6	1
Canada	0.2	0.2	0.2
Peru	4.6	4.5	3.6



Country Name	Attribute	Value
Argentina	2016	0.7
Argentina	2017	0.6
Argentina	2018	1
Canada	2016	0.2
Canada	2017	0.2
Canada	2018	0.2
Peru	2016	4.6
Peru	2017	4.5
Peru	2018	3.6

Converting from wide to tall format, taking related columns and collapsing them into a single column

Combining Data

Appending

- Stacking datasets vertically (adding rows).

Connecting to a Folder

- One of the options to “Get Data.” When importing a folder containing multiple files, combining the files is equivalent to “appending.”

Merging (Joining)

- Combines datasets horizontally based on a common primary/foreign key match (adding columns).

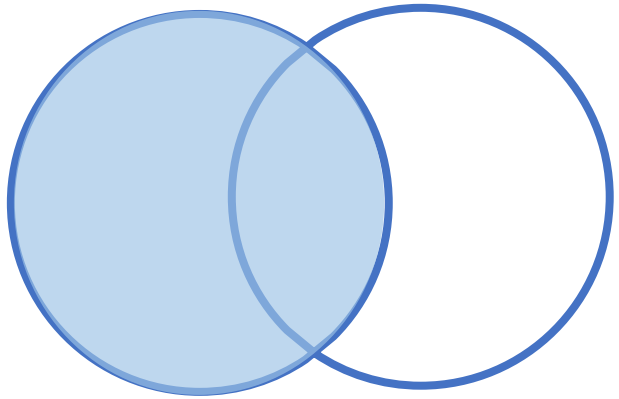
Appending

Location	OrderID	OrderDate	SaleAmount
Lindon	PO-2408	8/1/2026	$ 1,944.96
Lindon	PO-2412	8/1/2026	$ 2,064.61
Lindon	PO-2413	8/3/2026	$ 1,872.19
Lindon	PO-2414	8/4/2026	$ 1,399.68

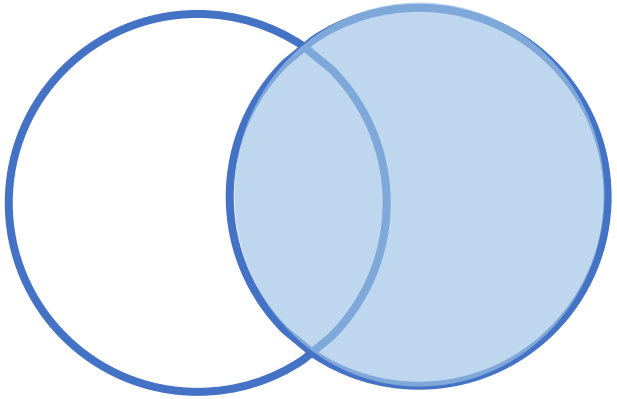
Location	OrderID	OrderDate	SaleAmount
Orem	PO-2409	8/1/2026	$ 1,739.61
Orem	PO-2410	8/2/2026	$ 2,038.60
Orem	PO-2411	8/3/2026	$ 1,659.36
Orem	PO-2415	8/4/2026	$ 1,814.87

Location	OrderID	OrderDate	SaleAmount
Lindon	PO-2408	8/1/2026	$ 1,944.96
Lindon	PO-2412	8/1/2026	$ 2,064.61
Lindon	PO-2413	8/3/2026	$ 1,872.19
Lindon	PO-2414	8/4/2026	$ 1,399.68
Orem	PO-2409	8/1/2026	$ 1,739.61
Orem	PO-2410	8/2/2026	$ 2,038.60
Orem	PO-2411	8/3/2026	$ 1,659.36
Orem	PO-2415	8/4/2026	$ 1,814.87

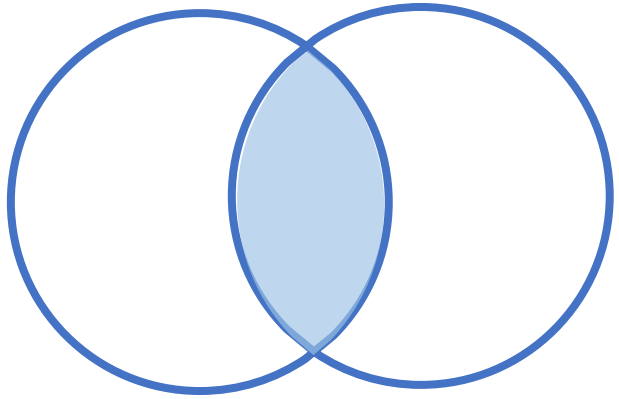
Four Major Types of Joins



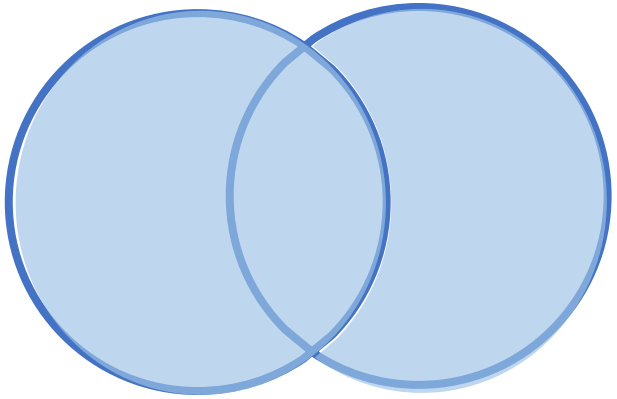
Left Join
Retains all records from the first table, and only matches from the second.



Right Join
Retains all records from the second table, and only matches from the first.

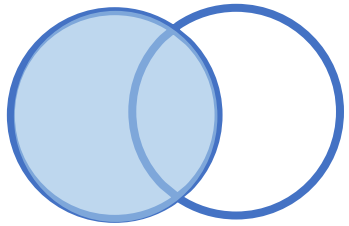


Inner Join
Retains only matches between the two tables.



Full Outer Join
Retains all records of both tables, with matching records being joined where possible.

Left Join

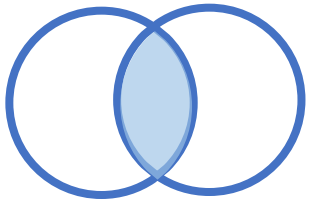


LocationID	OrderID	OrderDate	SaleAmount
1	PO-2408	8/1/2026	$ 1,944.96
3	PO-2412	8/1/2026	$ 2,064.61
1	PO-2413	8/3/2026	$ 1,872.19
2	PO-2414	8/4/2026	$ 1,399.68
3	PO-2409	8/1/2026	$ 1,739.61
1	PO-2410	8/2/2026	$ 2,038.60
2	PO-2411	8/3/2026	$ 1,659.36
2	PO-2415	8/4/2026	$ 1,814.87

LocationID	LocationName
1	Lindon
2	Orem
4	Provo

LocationID	OrderID	OrderDate	SaleAmount	LocationName
1	PO-2408	8/1/2026	$ 1,944.96	Lindon
3	PO-2412	8/1/2026	$ 2,064.61	
1	PO-2413	8/3/2026	$ 1,872.19	Lindon
2	PO-2414	8/4/2026	$ 1,399.68	Orem
3	PO-2409	8/1/2026	$ 1,739.61	
1	PO-2410	8/2/2026	$ 2,038.60	Lindon
2	PO-2411	8/3/2026	$ 1,659.36	Orem
2	PO-2415	8/4/2026	$ 1,814.87	Orem

Inner Join

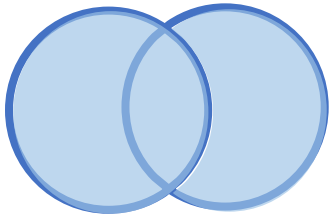


LocationID	OrderID	OrderDate	SaleAmount
1	PO-2408	8/1/2026	$ 1,944.96
3	PO-2412	8/1/2026	$ 2,064.61
1	PO-2413	8/3/2026	$ 1,872.19
2	PO-2414	8/4/2026	$ 1,399.68
3	PO-2409	8/1/2026	$ 1,739.61
1	PO-2410	8/2/2026	$ 2,038.60
2	PO-2411	8/3/2026	$ 1,659.36
2	PO-2415	8/4/2026	$ 1,814.87

LocationID	LocationName
1	Lindon
2	Orem
4	Provo

LocationID	OrderID	OrderDate	SaleAmount	LocationName
1	PO-2408	8/1/2026	$ 1,944.96	Lindon
1	PO-2413	8/3/2026	$ 1,872.19	Lindon
2	PO-2414	8/4/2026	$ 1,399.68	Orem
1	PO-2410	8/2/2026	$ 2,038.60	Lindon
2	PO-2411	8/3/2026	$ 1,659.36	Orem
2	PO-2415	8/4/2026	$ 1,814.87	Orem

Full Outer Join



LocationID	OrderID	OrderDate	SaleAmount
1	PO-2408	8/1/2026	$ 1,944.96
3	PO-2412	8/1/2026	$ 2,064.61
1	PO-2413	8/3/2026	$ 1,872.19
2	PO-2414	8/4/2026	$ 1,399.68
3	PO-2409	8/1/2026	$ 1,739.61
1	PO-2410	8/2/2026	$ 2,038.60
2	PO-2411	8/3/2026	$ 1,659.36
2	PO-2415	8/4/2026	$ 1,814.87

LocationID	LocationName
1	Lindon
2	Orem
4	Provo

LocationID	OrderID	OrderDate	SaleAmount	LocationName
1	PO-2408	8/1/2026	$ 1,944.96	Lindon
3	PO-2412	8/1/2026	$ 2,064.61	
1	PO-2413	8/3/2026	$ 1,872.19	Lindon
2	PO-2414	8/4/2026	$ 1,399.68	Orem
3	PO-2409	8/1/2026	$ 1,739.61	
1	PO-2410	8/2/2026	$ 2,038.60	Lindon
2	PO-2411	8/3/2026	$ 1,659.36	Orem
2	PO-2415	8/4/2026	$ 1,814.87	Orem
4				Provo

Day 6 – Process Mining

Please log in to Canvas and download and extract the PM.zip folder as you arrive

ACC 6300
Advanced Data Analytics
Mason Snow, PhD

The Changing Role of Accountants

In the past, accountants focused primarily on:

- Preparing financial reports
- Auditing
- Compliance (e.g., tax)

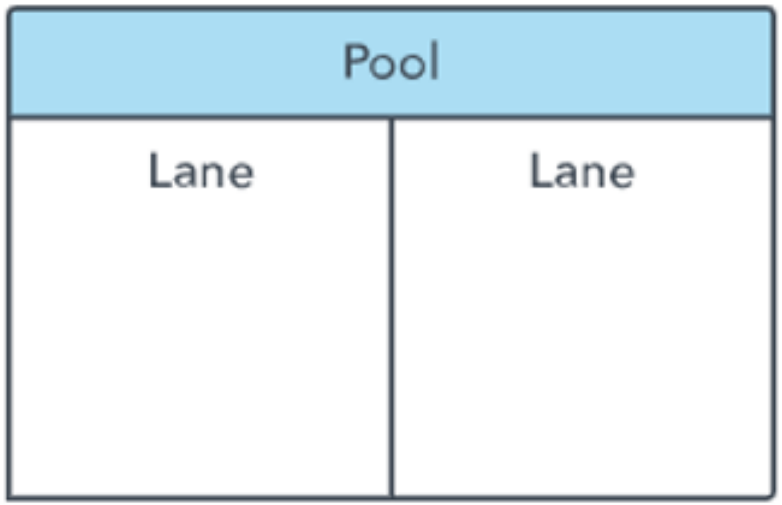
Now, accountants also take an active role in:

- Helping organizations **optimize processes**
- Achieving **competitive advantages**
- Maximizing **shareholder value**

"Given accountants' knowledge of business processes and how information flows through the organization, accountants can help management create specific questions that get to the heart of the analysis at hand."

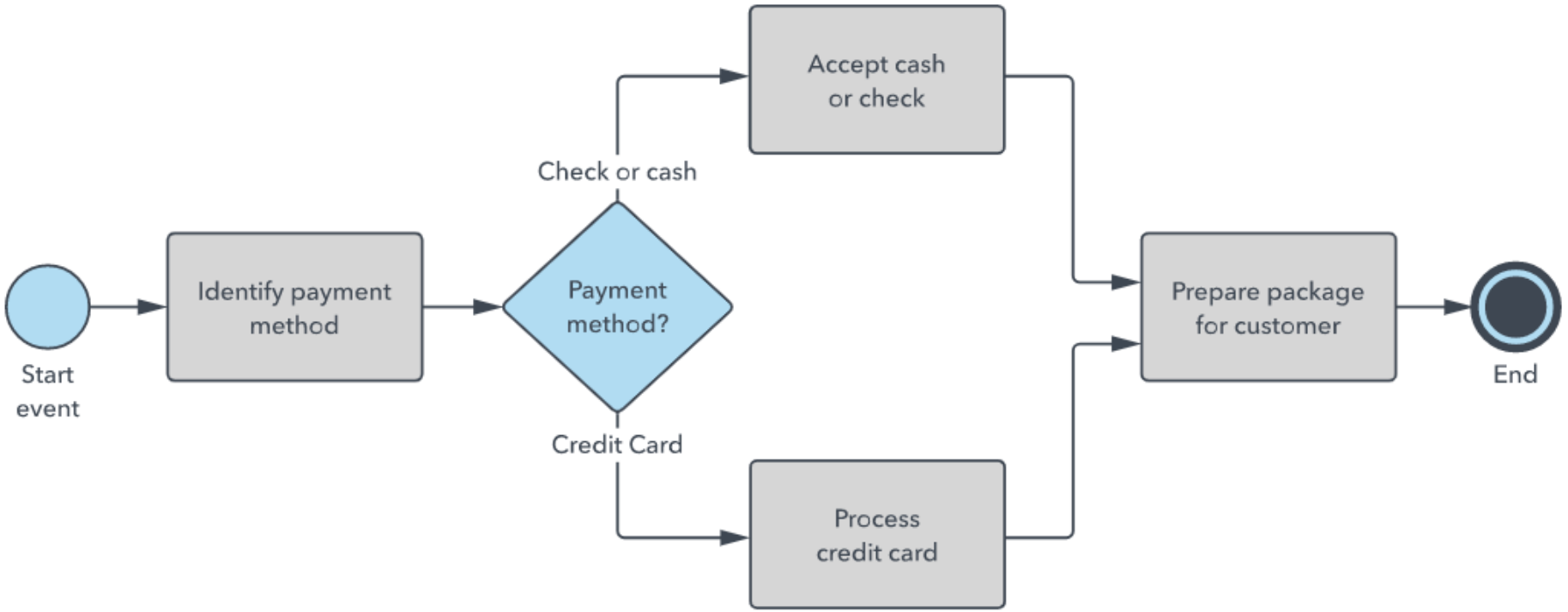
(Richardson and Watson, 2021)

Review of BPMN

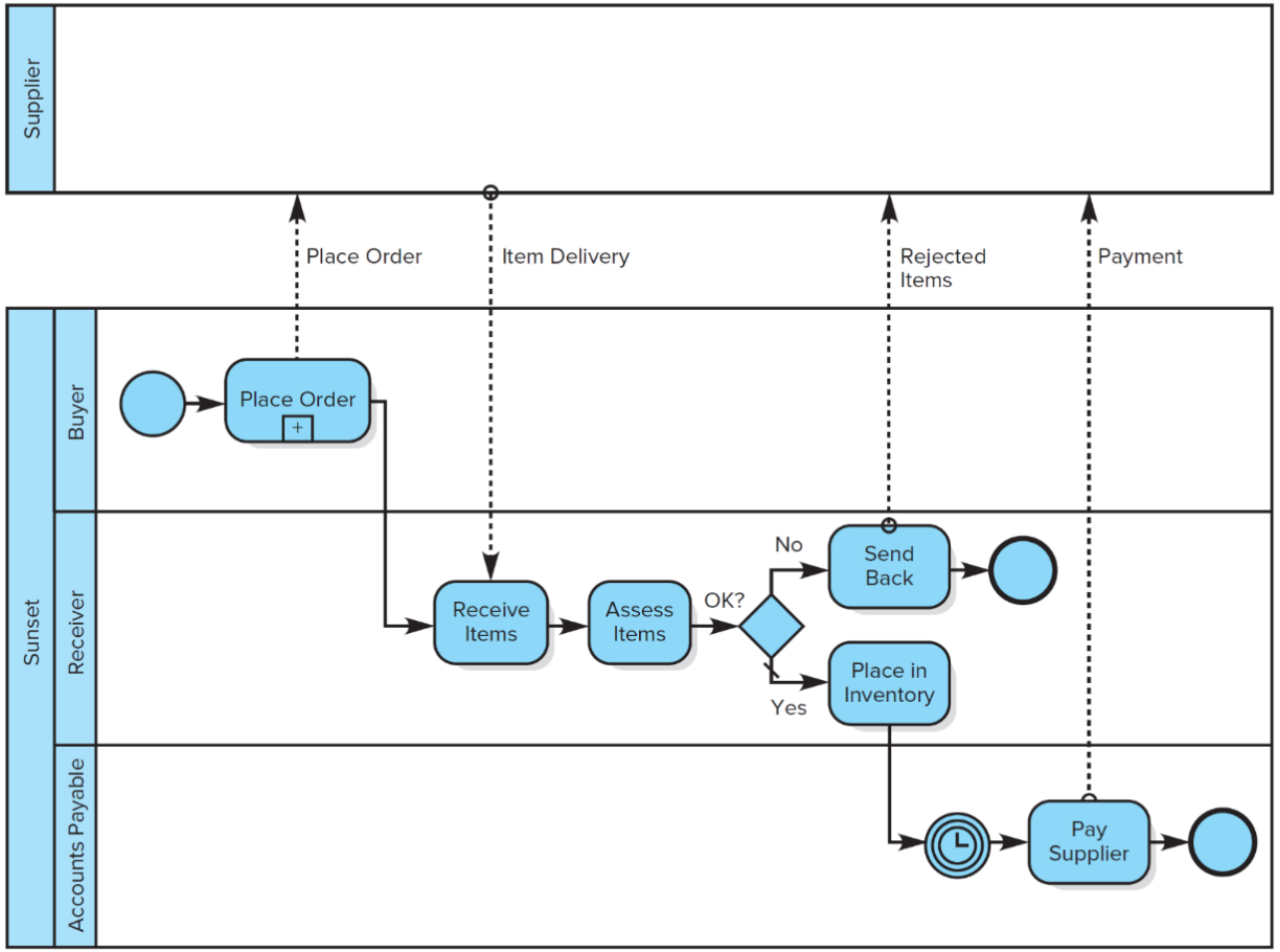


Element	Description	Symbol
Event	Incidents that affect the flow of the business process	start, intermediate, end
Activities	The processes themselves, the work that gets done.	Activity
Sequence Flow	Indicates the order of activities	Sequence flow
Gateways	Decision points that control the sequence of subsequently performed activities	Gateway
Annotations	Provide descriptive information to the diagram	Textual annotation
Message Flow	Interactions between pools	Message flow

Review of BPMN



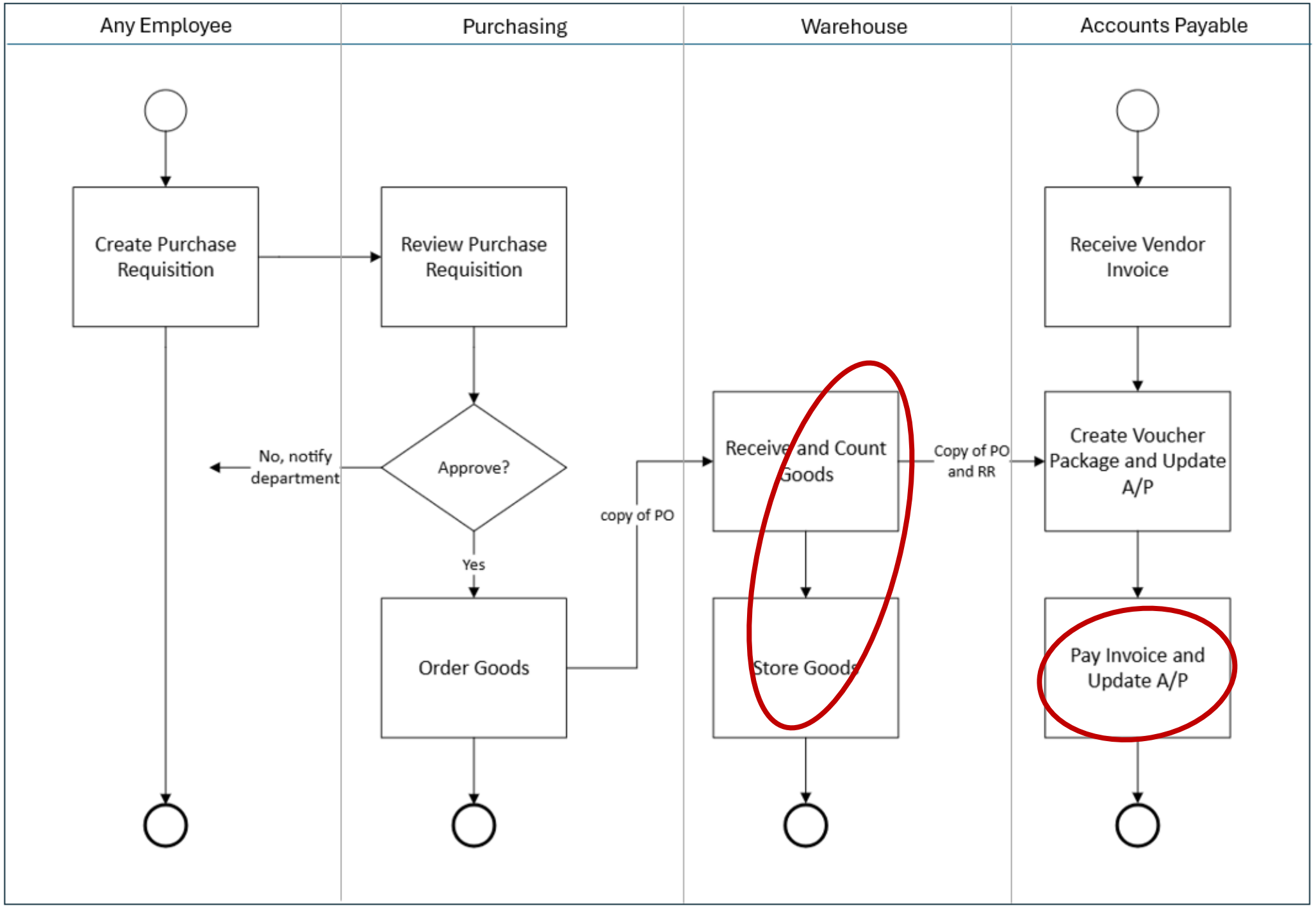
Review of BPMN



Segregation of Duties

Requiring certain tasks be performed by separate individuals. Within a business process, individuals should not perform duties across more than one of the following areas:

Custody (of assets)
Authorization (of procedures)
Recording (of information)

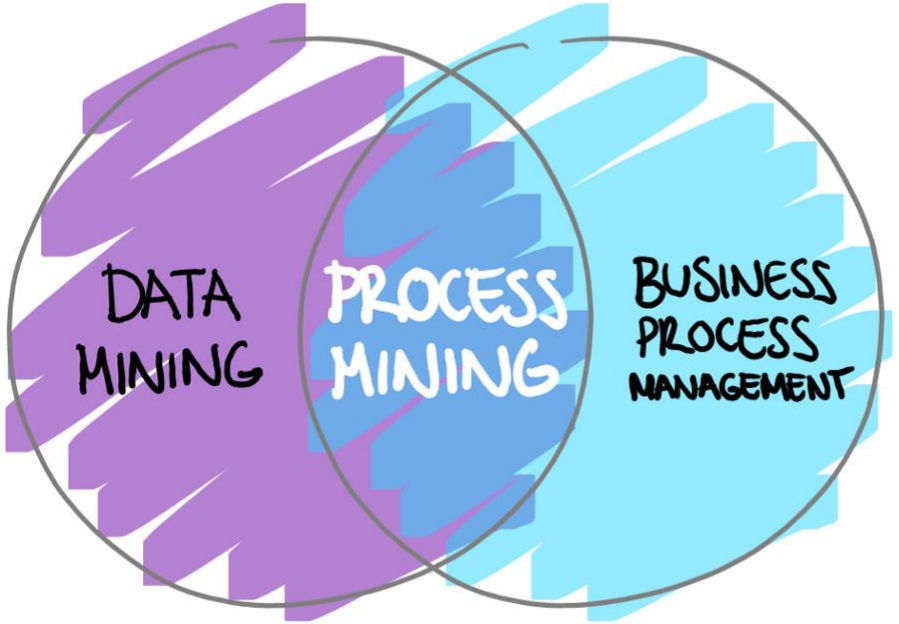

Any Employee
Purchasing
Warehouse
Accounts Payable
Create Purchase Requisition
Review Purchase Requisition
Approve?
No, notify department
Yes
Order Goods
copy of PO
Receive and Count Goods
Store Goods
Copy of PO and RR
Receive Vendor Invoice
Create Voucher Package and Update A/P
Pay Invoice and Update A/P

Process Mining

An approach to automate business process modeling and analysis.

Used for:

- Process discovery
- Locating inefficiencies
- Assessing compliance



Process Mining – Event Logs

Process mining platforms retrieve data from **event logs** to produce insights.

Three required data inputs:

- **Case ID**: a unique reference to identify each instance of a cycle flow. (**WHICH**)
- **Activity**: a description of the process that the instance has undergone (**WHAT**)
- **Timestamp**: a record of when the case went through an activity (**WHEN**)

Event Log Example

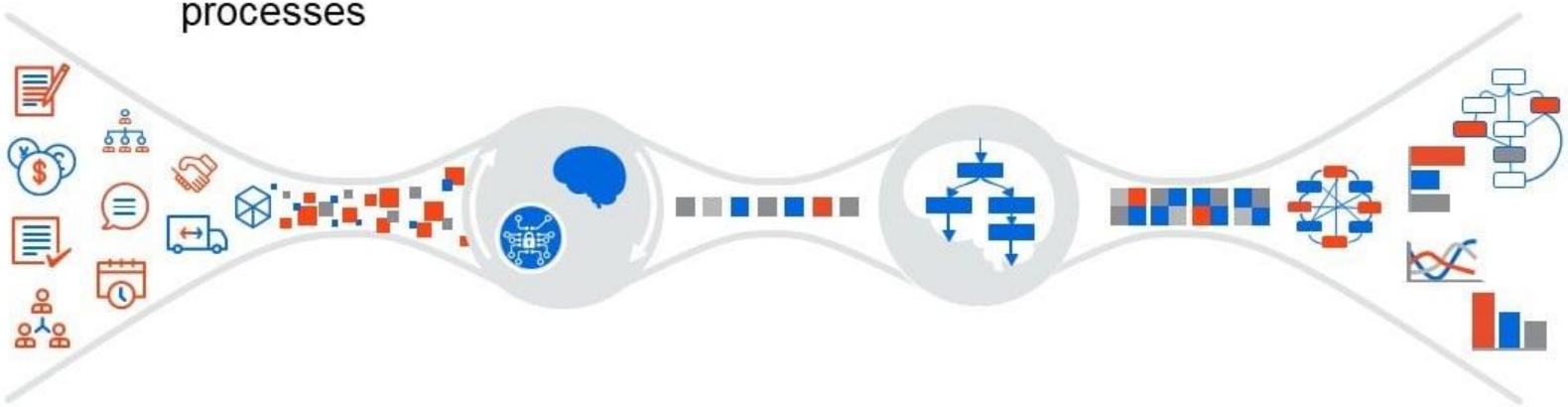
CaseID	Activity	Timestamp	Employee	Site
1001	Receive order	1/1/2022 8:00am	Tara (Sales)	CA
1001	Check credit	1/1/2022 8:03am	Liwei (Manager)	CA
1001	Pack order	1/1/2022 10:38am	Scott (Inventory)	CA
1002	Receive order	1/1/2022 10:45am	Tara (Sales)	CA
1002	Check credit	1/1/2022 10:59am	Liwei (Manager)	CA
1003	Receive order	1/1/2022 2:47pm	Carlos (Sales)	AZ
1001	Ship order	1/1/2022 2:58pm	Scott (Inventory)	CA
1003	Pack order	1/1/2022 3:12pm	Ana (Inventory)	AZ
1001	Send Invoice	1/2/2022 8:04am	Liwei (Manager)	CA
1004	Receive order	1/2/2022 10:12am	Carlos (Sales)	AZ
1004	Pack order	1/2/2022 12:42pm	Ana (Inventory)	AZ
1003	Ship order	1/2/2022 2:26pm	Ana (Inventory)	AZ

Start to Finish

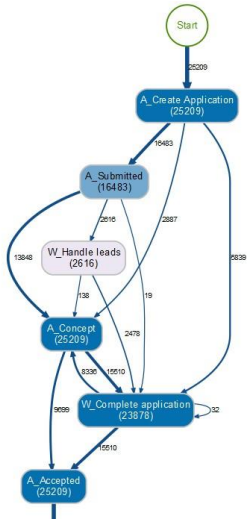
Reads
data related to your processes

Transforms
data to event log

Visualizes
the end-to-end process



CaseID	Activity	Timestamp	Employee	Site
1001	Receive order	1/1/2022 8:00am	Tara (Sales)	CA
1001	Check credit	1/1/2022 8:03am	Liwei (Manager)	CA
1001	Pack order	1/1/2022 10:38am	Scott (Inventory)	CA
1002	Receive order	1/1/2022 10:45am	Tara (Sales)	CA
1002	Check credit	1/1/2022 10:59am	Liwei (Manager)	CA
1003	Receive order	1/1/2022 2:47pm	Carlos (Sales)	AZ
1001	Ship order	1/1/2022 2:58pm	Scott (Inventory)	CA
1003	Pack order	1/1/2022 3:12pm	Ana (Inventory)	AZ
1001	Send Invoice	1/2/2022 8:04am	Liwei (Manager)	CA
1004	Receive order	1/2/2022 10:12am	Carlos (Sales)	AZ
1004	Pack order	1/2/2022 12:42pm	Ana (Inventory)	AZ
1003	Ship order	1/2/2022 2:26pm	Ana (Inventory)	AZ

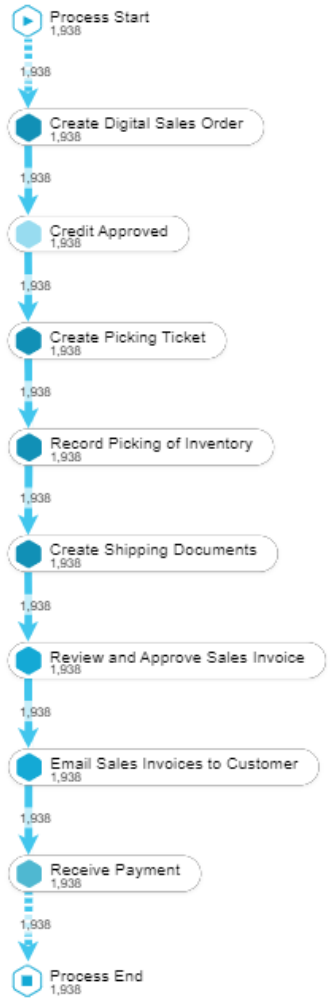


Variant Analysis

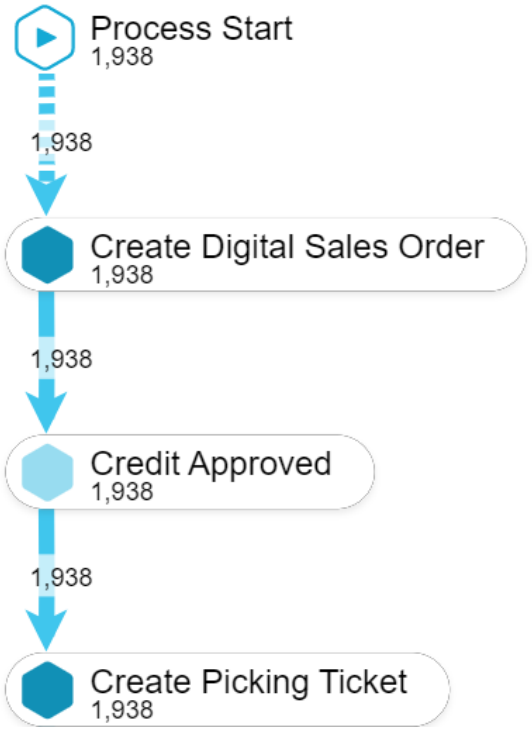
A **variant** is a unique sequence of activities taken by at least one case.

The number of times that an activity is performed is listed within each bubble. The number of times a case proceeds from one activity to the next is listed within each arrow.

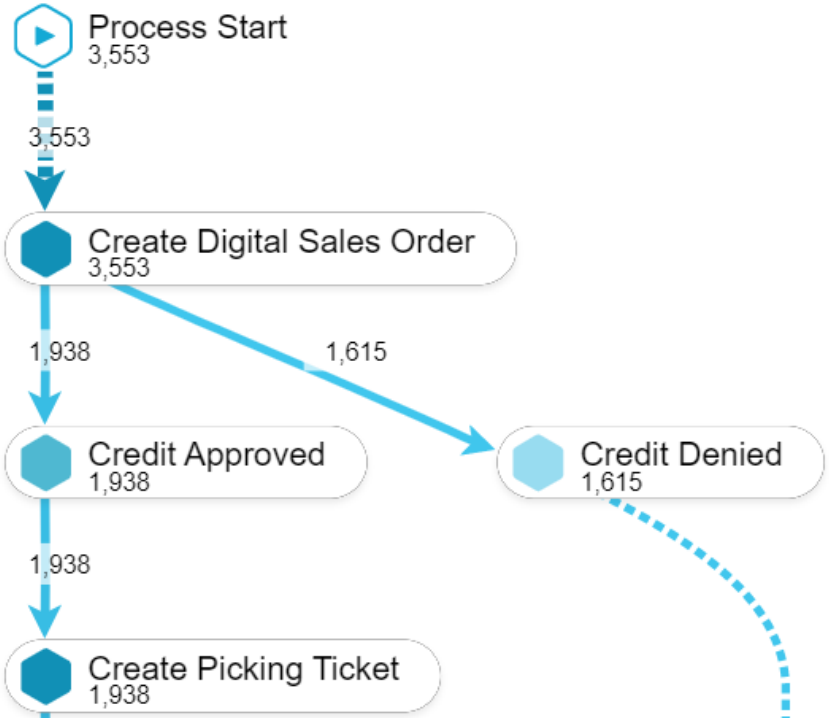
The most common variant is called the “happy path.”



Variant Analysis



One variant



Two variants

Limitations of Process Mining

1. The entire organization needs to sufficiently support providing the necessary data to create event logs.

2. Not all processes will be recorded (and hence cannot be analyzed).

3. Analysis of complicated processes (i.e., with many variants) can be unwieldy.

Day 9 – Statistics and Regression

Please log in to Canvas and download and extract the Stats.zip folder as you arrive

ACC 6300
Advanced Data Analytics

Mason Snow, PhD

Descriptive Statistics

Descriptive statistics summarize the distributional properties of a dataset:

Measures of Central Tendency:
Mean: The average value of the dataset.
Median: The middle value when the data points are arranged in order.
Mode: The most frequently occurring value in the dataset.

Measures of Dispersion:
Max/Min: The largest and smallest values.
Range: The difference between the maximum and minimum values.
Standard Deviation: Measures the dispersion or spread of data points around the mean.

Correlation

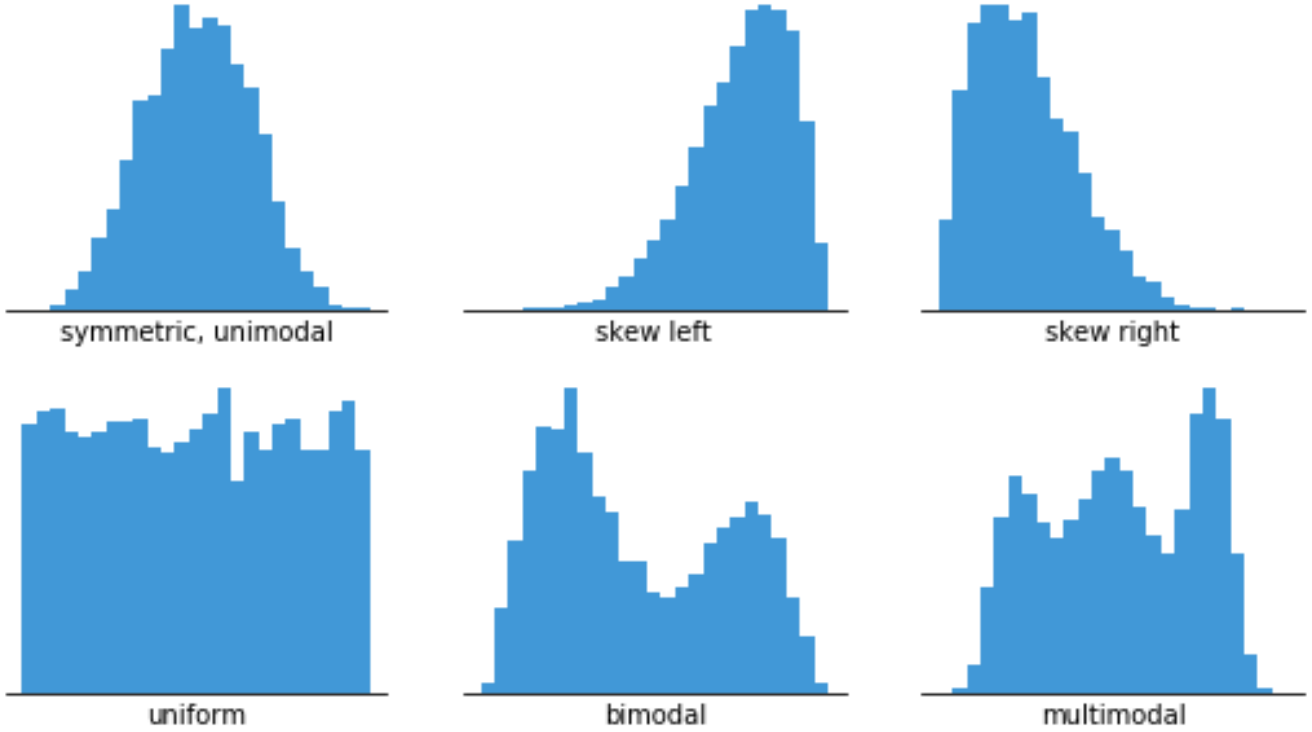
Correlation measures the strength and direction of a linear relationship between two variables.

Pearson Correlation Coefficient (r): Ranges from -1 to 1

Positive Correlation: As one variable increases, the other variable also increases.
Negative Correlation: As one variable increases, the other variable decreases.
No Correlation: No predictable linear relationship between variables.

Histograms

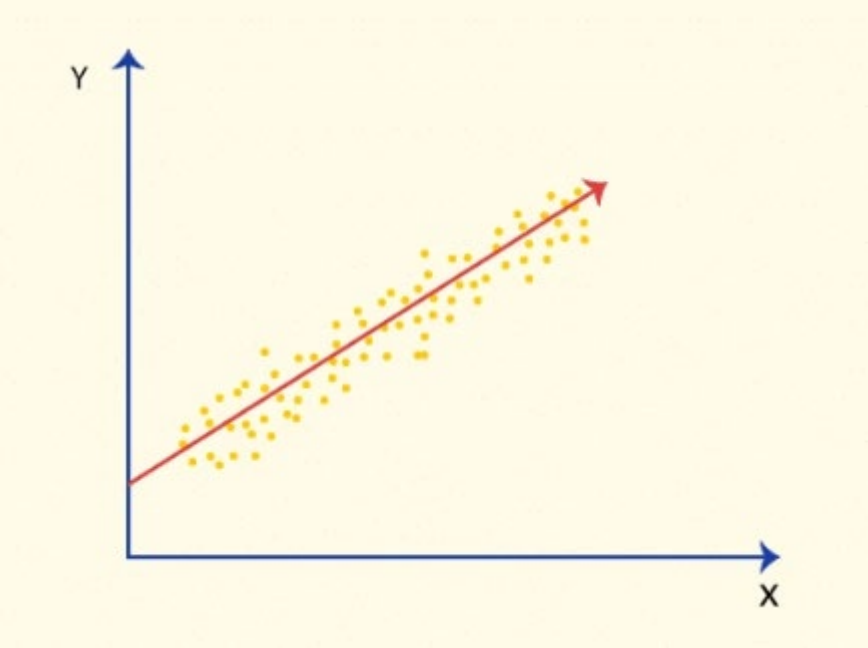
A histogram is a visual representation of the distribution of numerical data. It groups data into bins (intervals) and shows the frequency of data points within each bin.



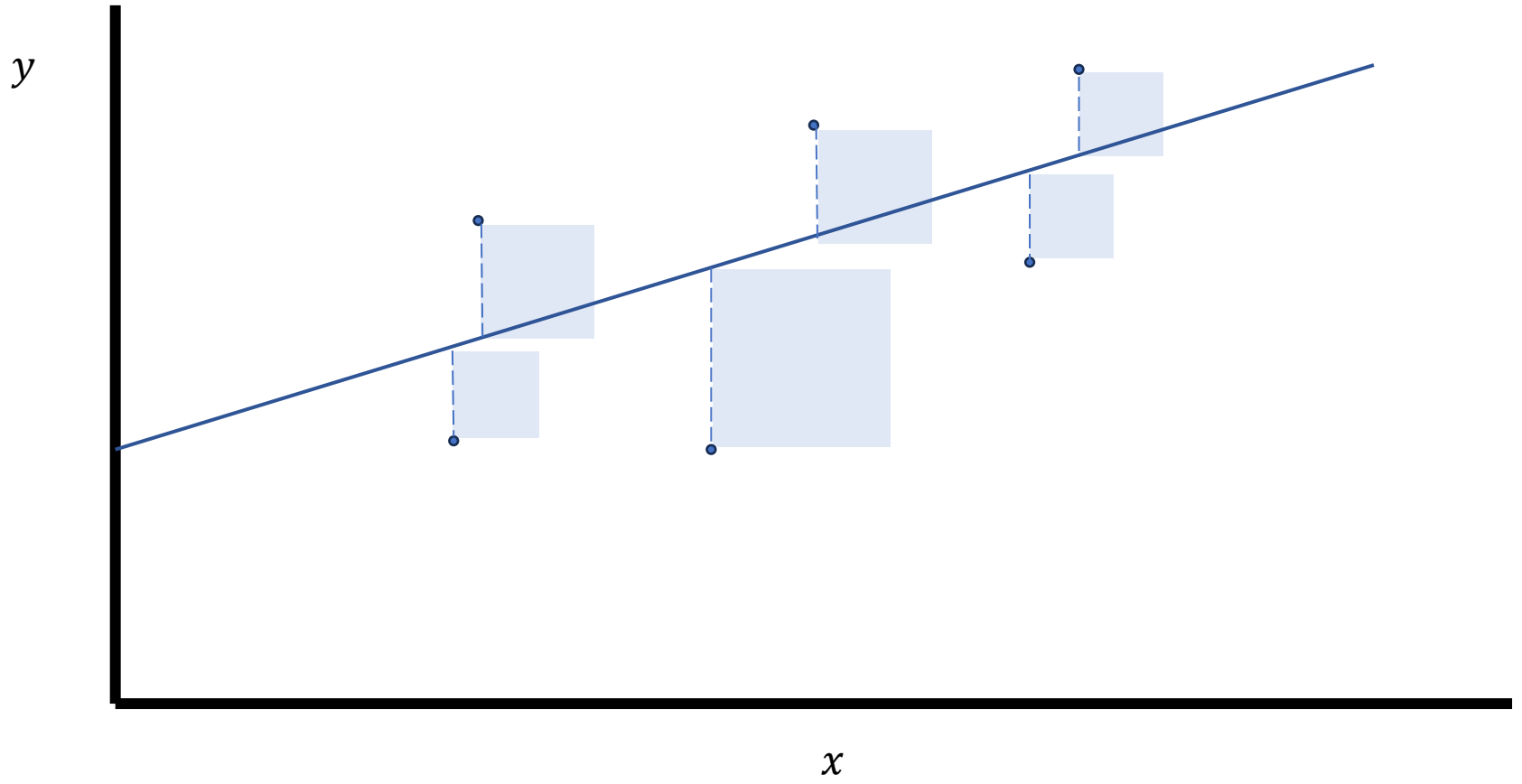
Linear Regression

A statistical technique that models the relationship between one or more independent variables and a dependent variable.

Useful for understanding associations, making predictions, and assessing potential causality.



Ordinary Least Squares



Linear Regression Model

$$y = \alpha + \beta x + \varepsilon$$

Linear Regression Model

$$y = \alpha + \beta x + \varepsilon$$

Dependent variable (*y*): The variable we are trying to predict or explain.

Independent variable (*x*): A variable we think may explain *y*.

Intercept (α)*:* The predicted value of *y* for when *x* equals zero.

Slope coefficient (β): Estimate of how *y* will change with a one-unit increase in *x*.

Error term (ε): The part of *y* that remains unexplained by *x*.

Multiple Linear Regression

$$y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_n x_n + \varepsilon$$

Multiple linear regression allows for more than one independent variable.

What is the benefit of including multiple X variables in one regression?

Each slope coefficient reflects the effect of the individual X_i on Y, holding all other independent variables **constant**. This is sometimes referred to as "controlling for" other X variables.

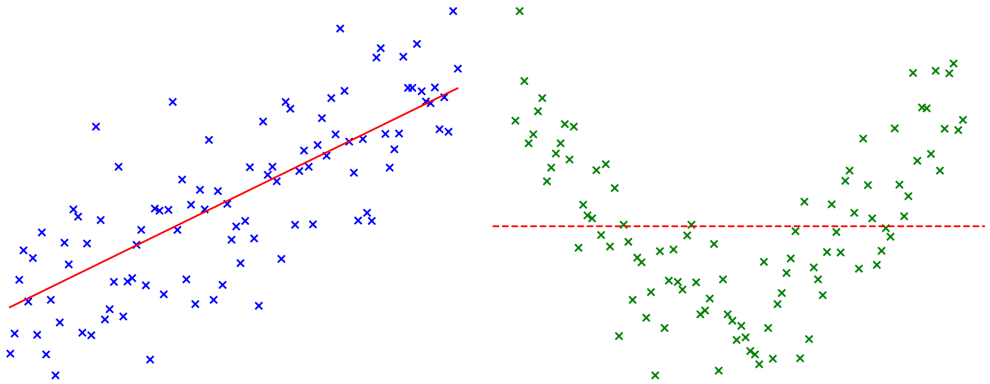
Multiple Linear Regression Example

$$y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

y: Monthly Sales ($)
x1: Local Advertising Spend
x2: Store Size
x3: Average Local Income

Multiple Linear Regression Assumptions

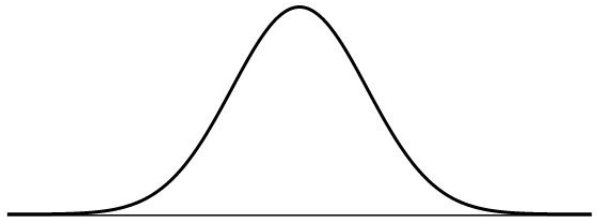
1. **Linearity**: The relationship between Y and X must be linear.



2. **Independence**: The Y value for one observation should not correlate with the next.

3. **No Multicollinearity**: The X variables should not be too highly correlated with each other (ρ>0.6)

4. **Normality**: The Y variable should be follow a normal distribution.



Backwards Elimination

One method to set up a prediction model is using ‘backwards elimination.’ This is a process of starting with a full model of independent variables and iteratively removing insignificant variables until all remaining are statistically significant.

1. Run a regression with a series of X variables that plausibly can help predict Y.
2. Identify which X variables were statistically significant predictors.
3. Run another regression with only the X variables from Step 2.
4. Repeat until the model contains only significant X variables.

Day 11 – Predictive Analytics

Please log in to Canvas and download and extract the Day11.zip folder as you arrive

ACC 6300
Advanced Data Analytics

Mason Snow, PhD

Reading Quiz #8

Let’s rattle off the “common regression mistakes” discussed in this chapter . . .

Today's topic:
Predictive Analytics

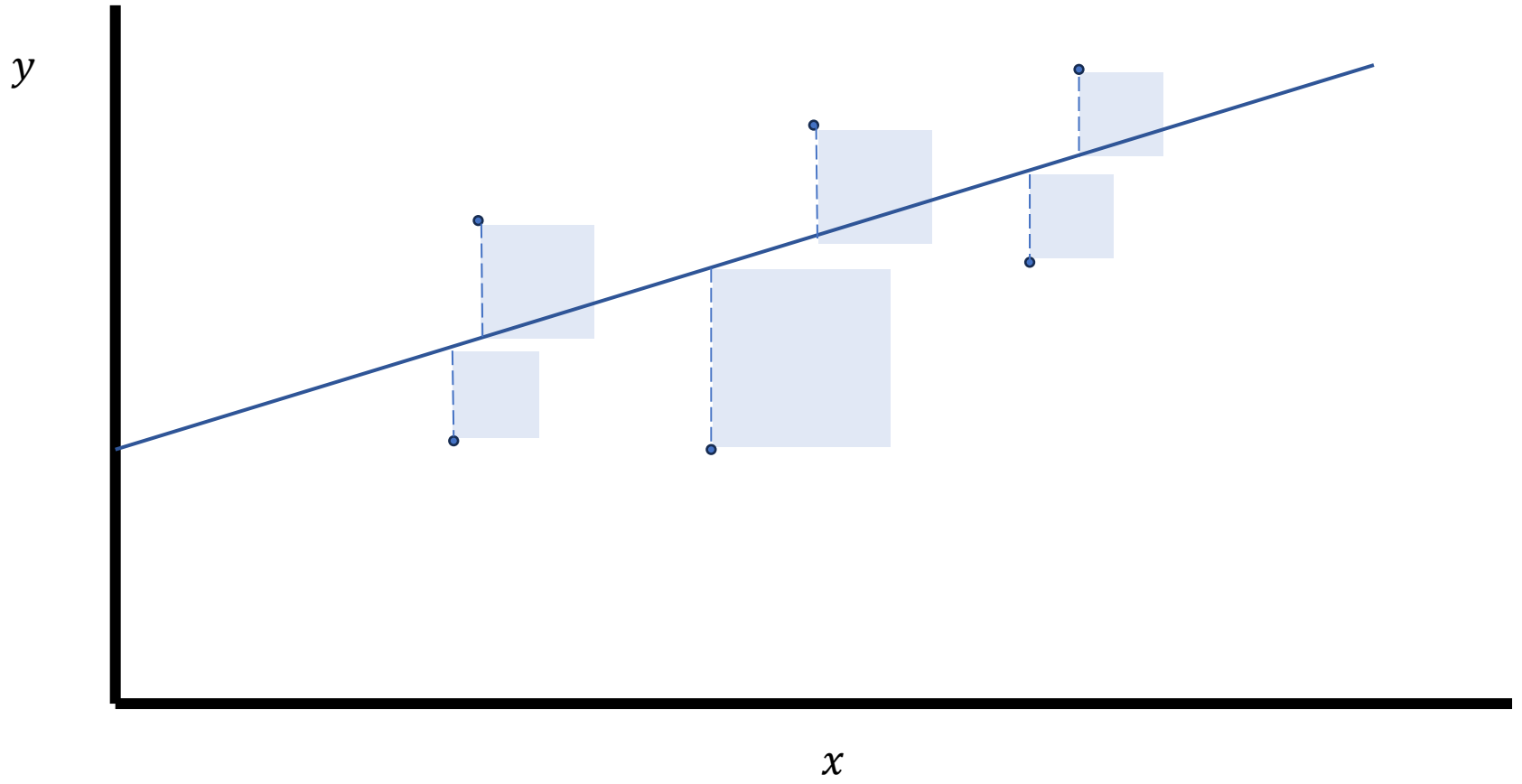
Let’s Review

Linear Regression

A statistical technique that models the relationship between one or more independent variables and a dependent variable.

Useful for understanding associations, making predictions, and assessing potential causality.

Ordinary Least Squares



Linear Regression Model

$$y = \alpha + \beta x + \varepsilon$$

Dependent variable (y): The variable we are trying to predict or explain.

Independent variable (x): A variable we think may explain y.

Intercept (α): The predicted value of y for when x equals zero.

Slope coefficient (β): Estimate of how y will change with a one-unit increase in x.

Error term (ε): The part of y that remains unexplained by x.

Multiple Linear Regression

$$y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_n x_n + \varepsilon$$

Multiple linear regression allows for more than one independent variable.

What is the benefit of including multiple X variables in one regression?

Each slope coefficient reflects the effect of the individual X_i on Y, holding all other independent variables **constant**. This is sometimes referred to as "controlling for" other X variables.

Interpreting Output in Excel

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.774478657
R Square	0.599817191
Adjusted R Square	0.551795253
Standard Error	19.70540981
Observations	100

The percent of the variation in Y explained by variation in X.

Number of rows in the dataset

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	10639.60018	3546.533393	12.84205	7.98429E-05
Residual	35	9660.142681	276.0040766		
Total	38	17749.74359			

Don’t worry about the ANOVA section

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-7.681253169	12.79255739	-0.600447036	0.553613	-34.02801831	18.66551197
X1	0.693146954	0.343142056	2.032486533	0.045688	0.221525846	1.164768061
X2	-0.193986362	0.225856579	-0.858891792	0.398561	-0.659146693	0.27117397
X3	0.302458404	0.101901463	2.968145855	0.006518	0.092588413	0.512328395

The explanatory variables included in the regression

The coefficient estimates for each of the variables (and the intercept)

P-values for each estimate (p<0.05 can be considered statistically significant)

1 | Cost Behavior

Managerial Accounting

A few conceptual questions . . .

Why do manufacturers “apply” overhead costs to jobs using a pre-determined rate instead of simply recording actual costs?

What’s the difference between fixed vs variable costs, and why should we care?

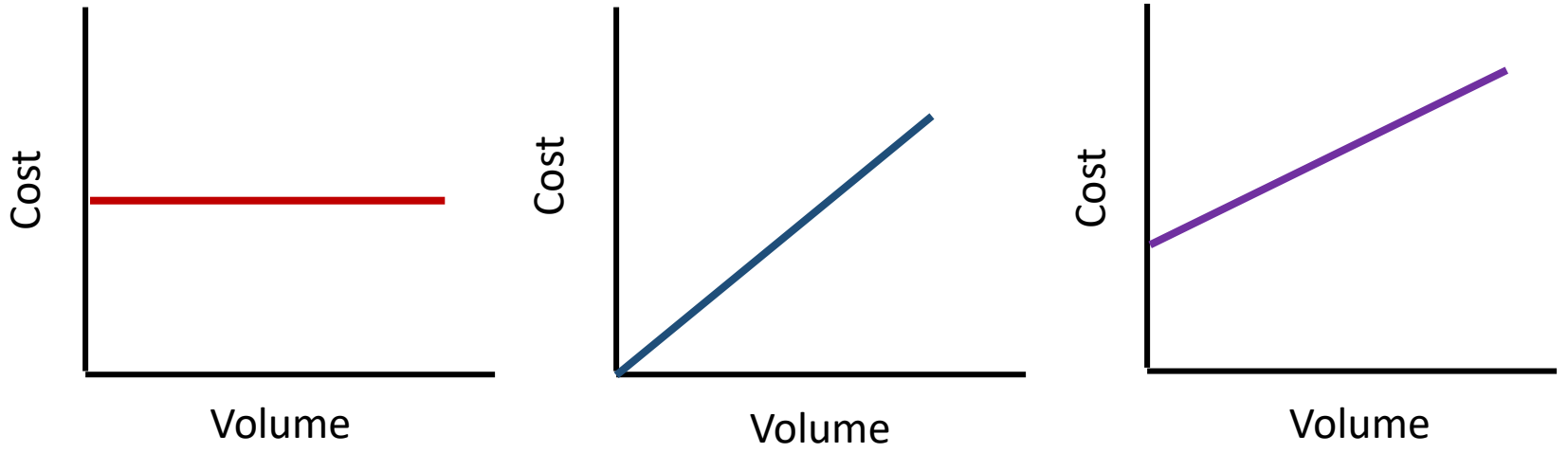
Is the distinction between fixed and variable costs always obvious?

Cost Behavior

Fixed costs: do not vary with production/sales volume.

Variable costs: vary entirely based on production/sales volume.

Mixed costs: vary somewhat with production volume but also have a fixed cost component.



2 | Earnings Properties

Financial Statement Analysis

Earnings Quality

Earnings Persistence

A measure of how well current earnings predict future earnings.

Higher earnings persistence indicates stable and predictable earnings over time.

Investors prefer firms with high earnings persistence.

Earnings Persistence

$$NetIncome_{t+1} = \alpha + \beta NetIncome_t + \varepsilon$$

β is typically a number between 0 and 1.

The higher the β, the more persistent the earnings.

Differential Persistence of Accruals vs Cash

$$NetIncome_{t+1} = \alpha + \beta_1 CFO_t + \beta_2 Accruals_t + \varepsilon$$

Discretionary Accruals

The Jones (1991) Model

TotalAccruals is the difference between Net Income and Cash Flow from Operations.

$$\frac{TotalAccruals_t}{Assets_{t-1}} = \beta_0 + \beta_1 \frac{1}{Assets_{t-1}} + \beta_2 \frac{\Delta Sales_t}{Assets_{t-1}} + \beta_3 \frac{PP\&E_t}{Assets_{t-1}} + \varepsilon$$

The explanatory (X) variables used in this model indicate factors as to why accruals will vary ***WITHOUT*** the use of accounting discretion (i.e., changes in sales, depreciation expense).

Discretionary Accruals

The Jones (1991) Model

$$\frac{TotalAccruals_t}{Assets_{t-1}} = \beta_0 + \beta_1 \frac{1}{Assets_{t-1}} + \beta_2 \frac{\Delta Sales_t}{Assets_{t-1}} + \beta_3 \frac{PP\&E_t}{Assets_{t-1}} + \boldsymbol{\varepsilon}$$

The regression output we care about is the estimated residual!
This is a measure of 'unexplained' or "discretionary" accruals.

3 | Prediction & Forecasting

Audit, Managerial, and Financial

Altman Z-Score

A model invented in 1968 by Edward Altman to predict bankruptcy risk.

Designed to assess the likelihood of a firm going bankrupt within the next two years.

While others have improved upon the original model, it is still a widely used model.

Altman Z-Score

Equation for Altman's Z-Score Model (1968):

$$Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 1X_5$$

X_1 = Working Capital / Total Assets

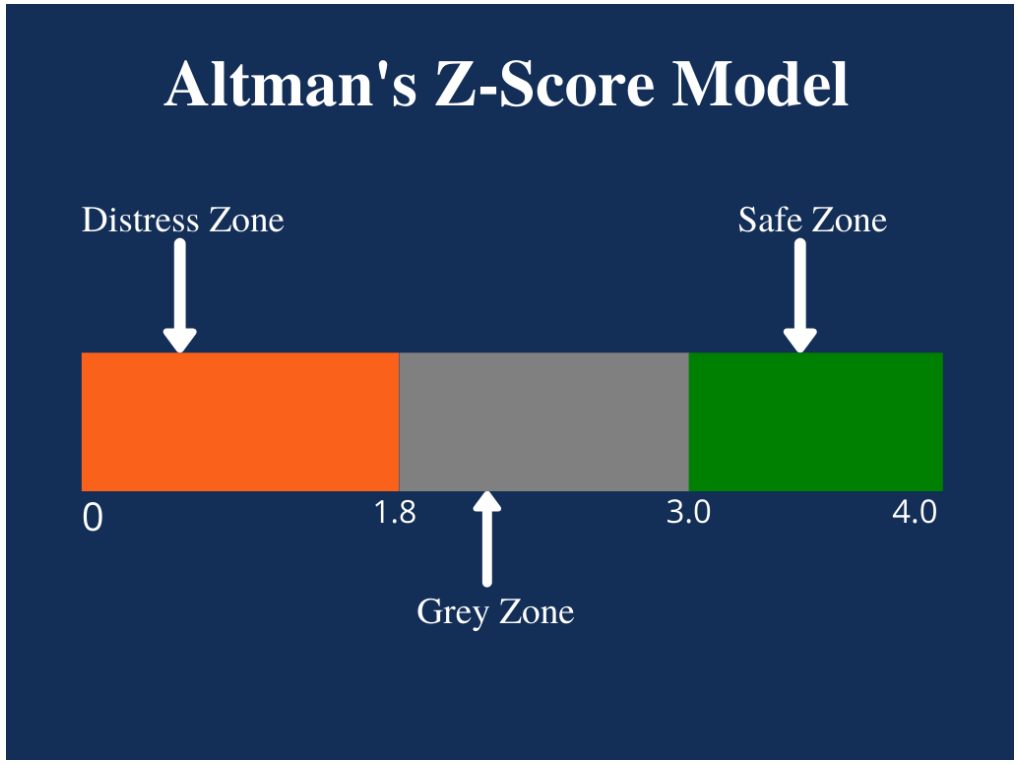
X_2 = Retained Earnings / Total Assets

X_3 = Earnings Before Interest & Tax (EBIT) / Total Assets

X_4 = Market Capitalisation / Total Liabilities

X_5 = Sales / Total Assets

Altman Z-Score



Z > 2.99 → Safe Zone

- The company is financially stable with a low probability of bankruptcy.
- Typically applies to well-established firms with solid financials.

1.81 ≤ Z ≤ 2.99 → Gray Zone

- The company is in a financial risk zone where distress is possible.
- Requires closer analysis of financial trends and industry conditions.

Z < 1.81 → Distress Zone

- High likelihood of financial distress or bankruptcy within two years.
- Often seen in struggling firms or industries facing downturns.

Monte Carlo Simulations

A computational technique that uses repeated random sampling to estimate outcomes.

Helps to model uncertainty across multiple dimensions.

Process:

1. Define the problem and identify key variables
2. Assign probability distributions to the variables
3. Run simulations
4. Analyze the results

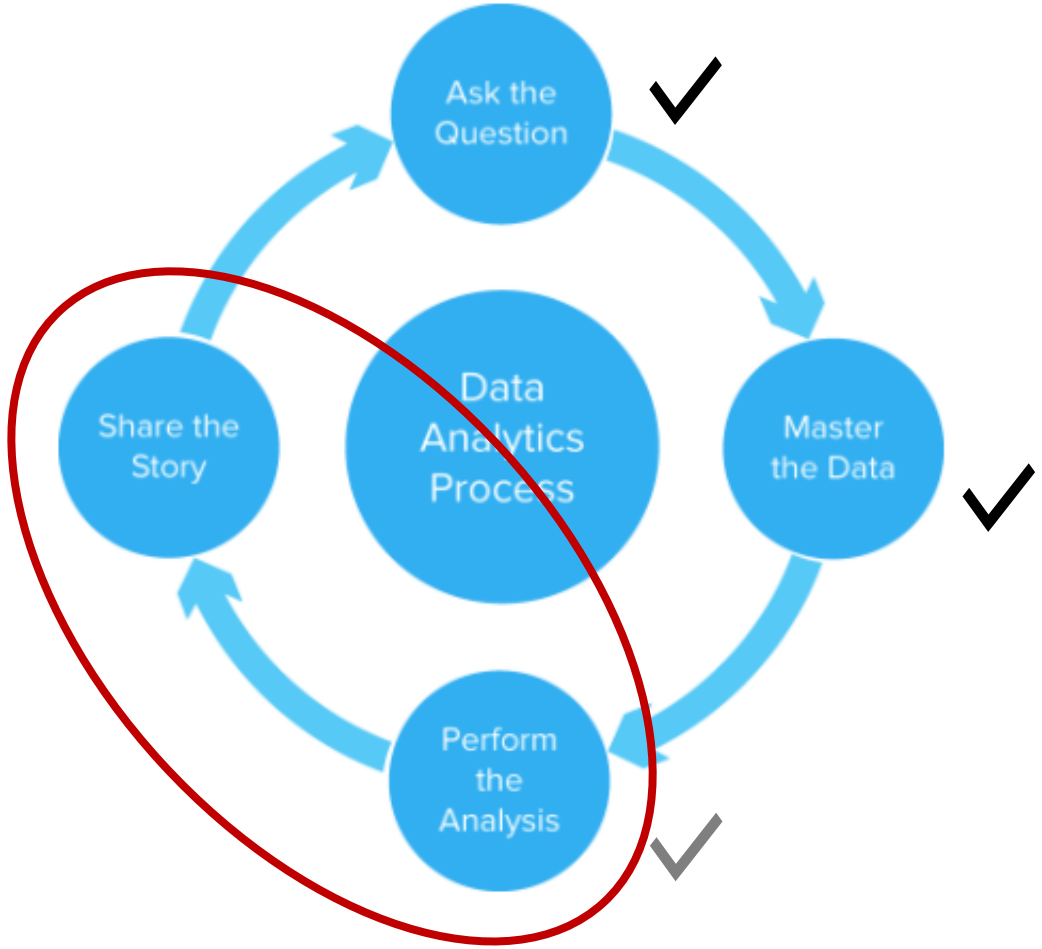
Day 13 – Exploratory Data Visualization

Please log in to Canvas and download and extract the Day13.zip folder as you arrive

ACC 6300
Advanced Data Analytics

Mason Snow, PhD

Touching base with the AMPS model



Data visualization objectives

Exploratory:

- Allows the user to *explore* data as a form of analysis
- Can be conducted for a problem that has not been clearly defined

Exploratory visualization is what we do to understand the data, to develop and assess a hypothesis, or discover insights from the data.

Explanatory:

- Explains to an audience what it needs to know
- Effectively communicates critical takeaways.

For explanatory data visualization, the insight is already known to the analyst. The objective is to communicate the information to others.

Humans cannot process data instantly

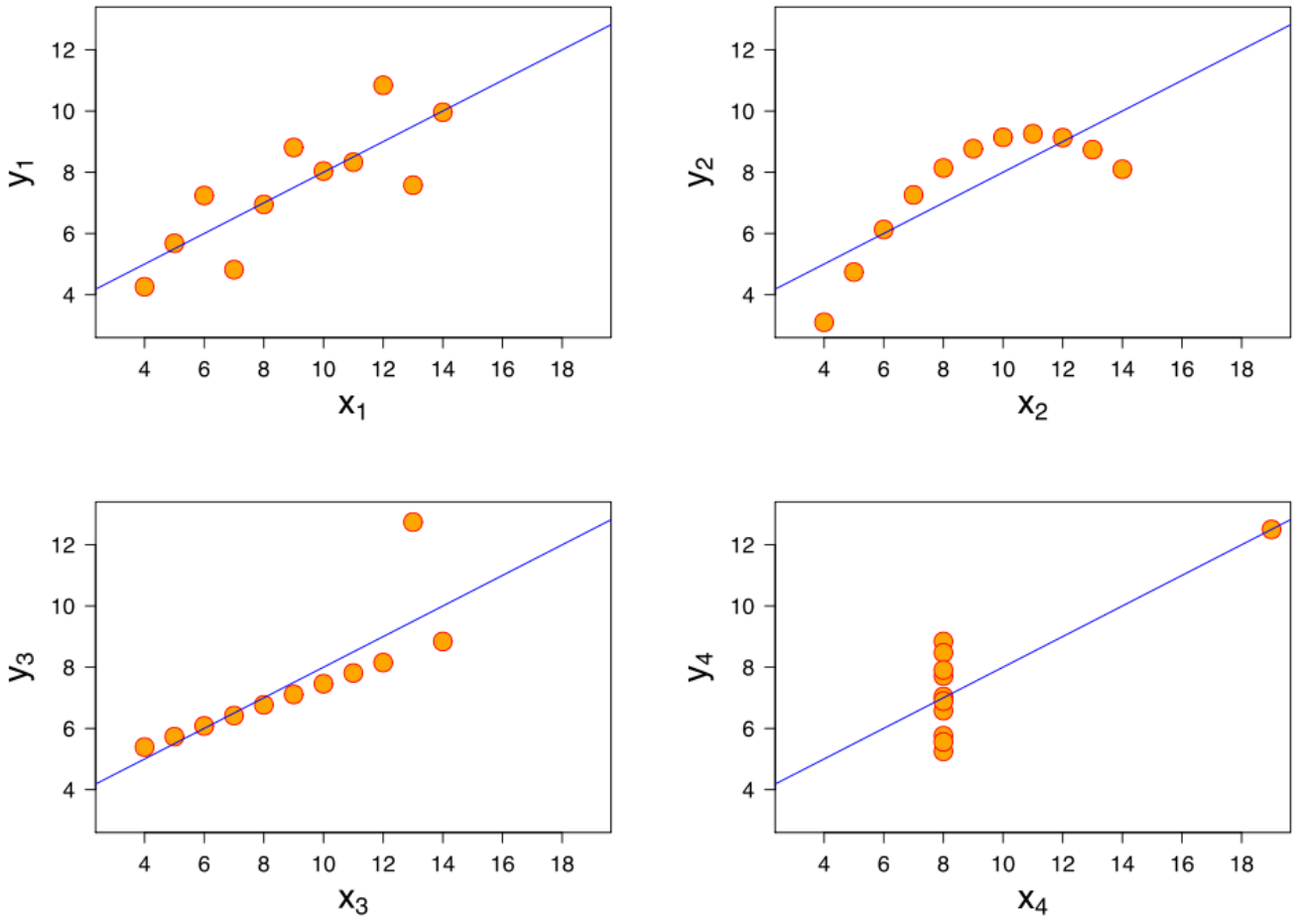
Set 1		Set 2		Set 3		Set 4	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

What patterns (if any) do you observe between X and Y?

Descriptive stats may mislead

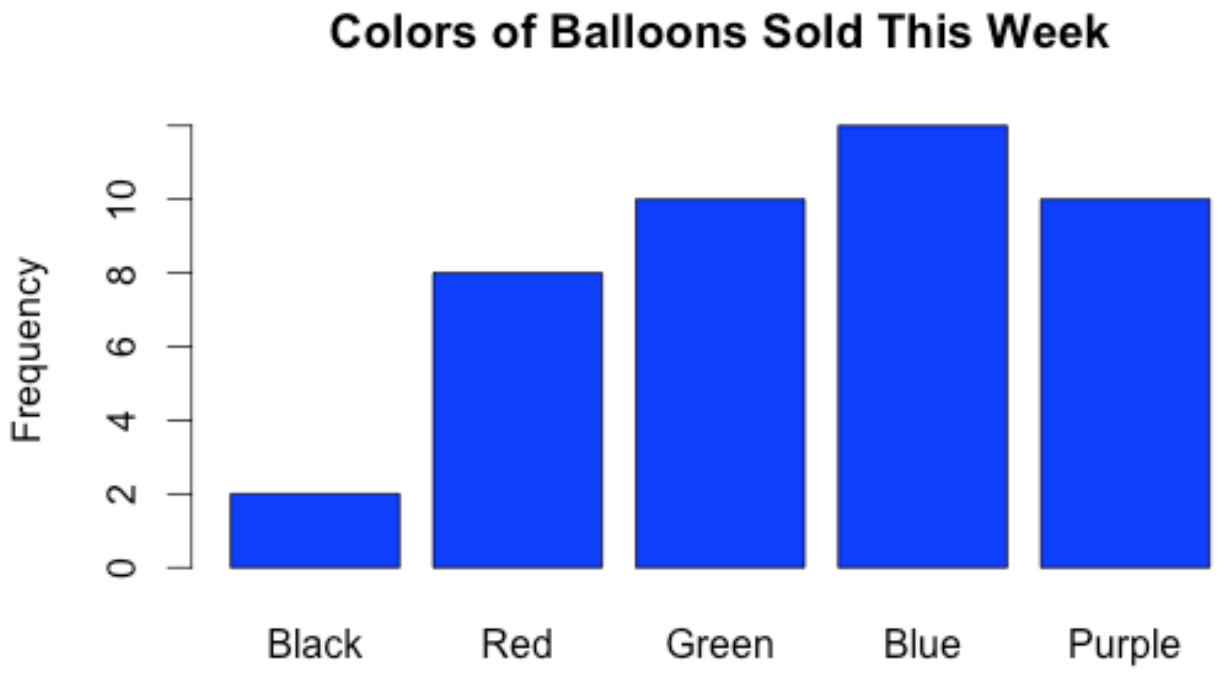
	Set 1	Set 2	Set 3	Set 4
Mean (x)	9.00	9.00	9.00	9.00
Mean (y)	7.50	7.50	7.50	7.50
Variance (x)	11.00	11.00	11.00	11.00
Variance (y)	4.13	4.13	4.12	4.12
Correlation	0.82	0.82	0.82	0.82
Intercept	3.00	3.00	3.00	3.00
Slope	0.50	0.50	0.50	0.50

Humans can process pictures quickly



Categorical Data

Categorical data presents an analysis of different groups. Its purpose is to facilitate comparison across categories.

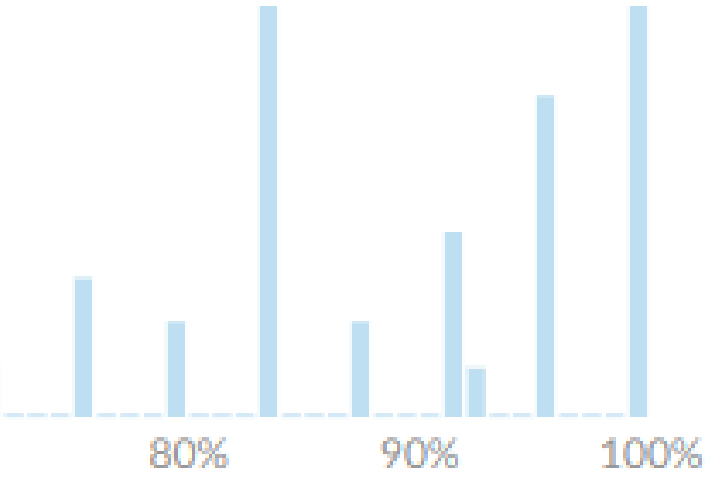


Bar chart

Color
Blue
Purple
Green
Red
Purple
Purple
Blue
Purple
Blue
Blue
Purple
Red
Black
Blue
...

Univariate Data

Univariate data can be used to map summary statistics (e.g., mean, variance, and skew) for a single variable. Its purpose is to better understand the distribution of a variable.

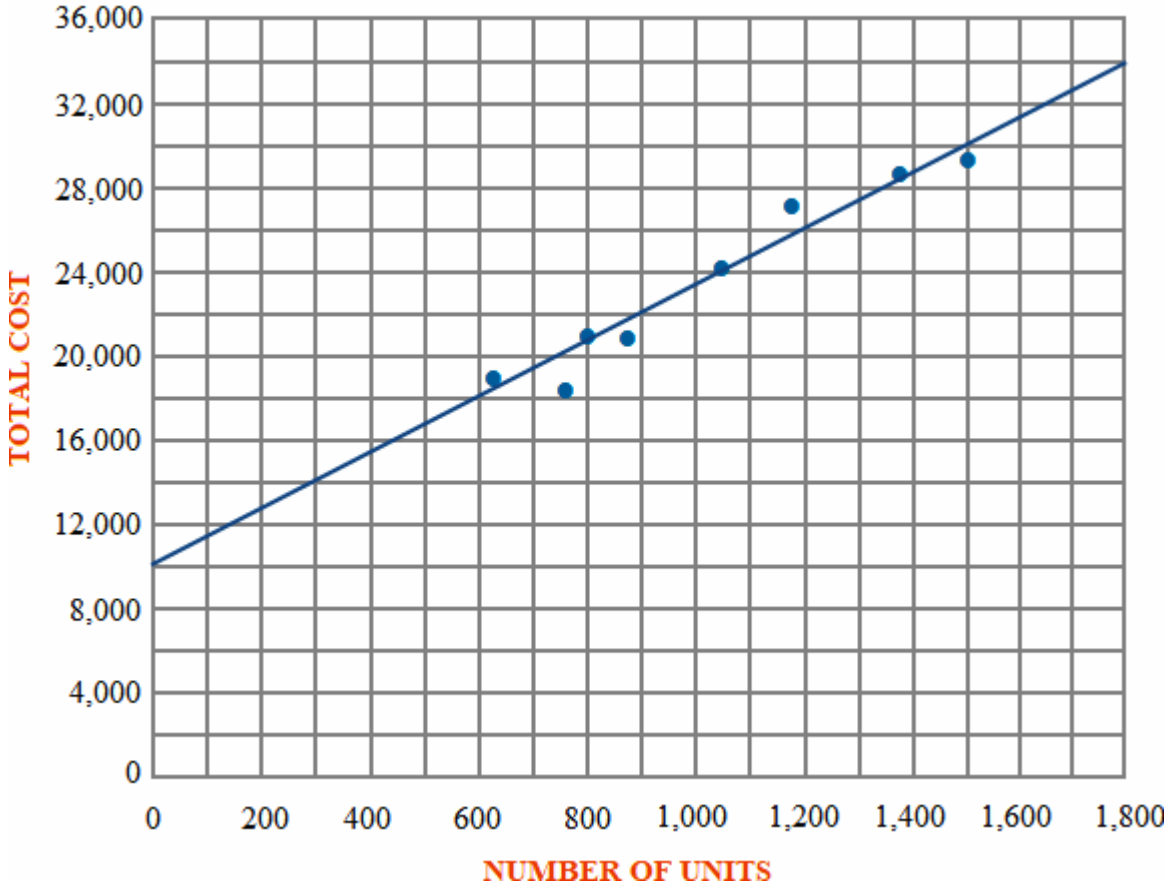


Histogram

Multivariate Data

Multivariate data analyses work with two (or more) variables. Its purpose is to examine relationships and correlations between these variables.

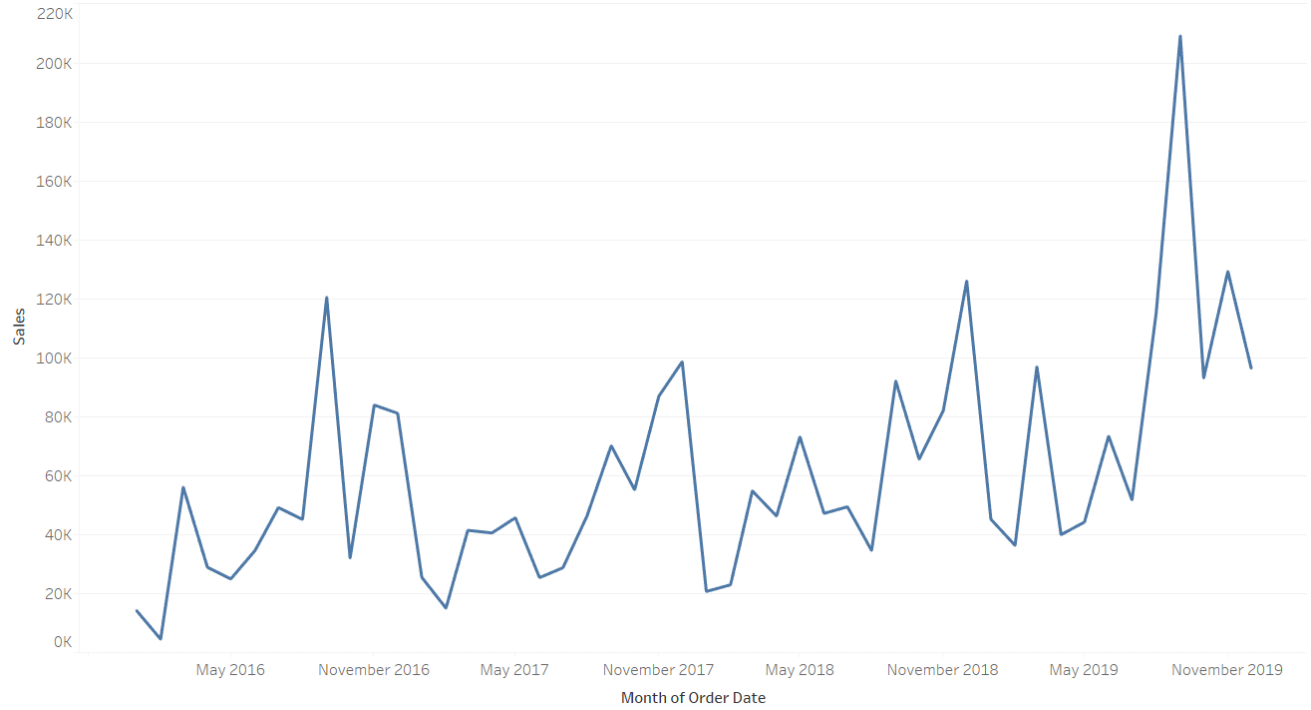
For visualization, relationships are illustrated between two variables at a time.



Scatterplot

Time Series Data

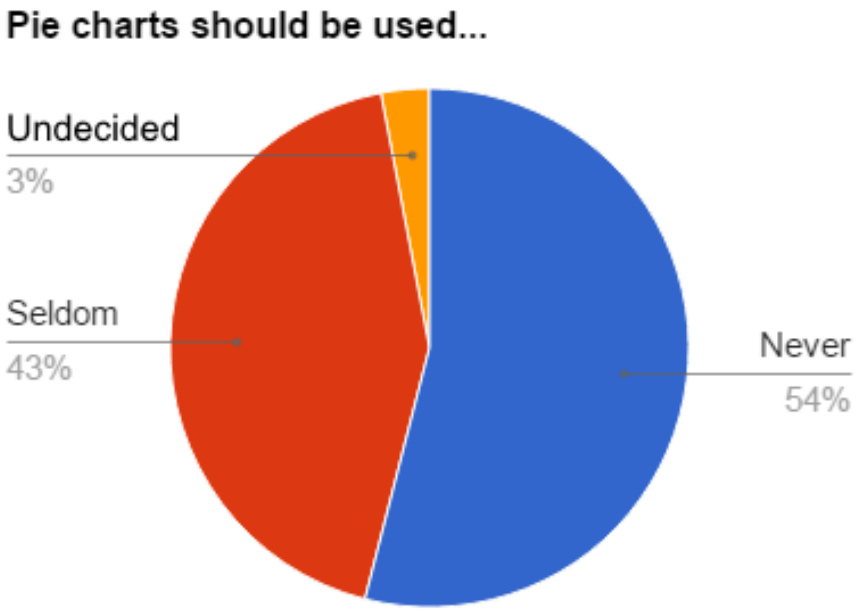
Data is considered a '**time series**' if it captures differences observations within a unit over time. Its purpose is to assess trends.



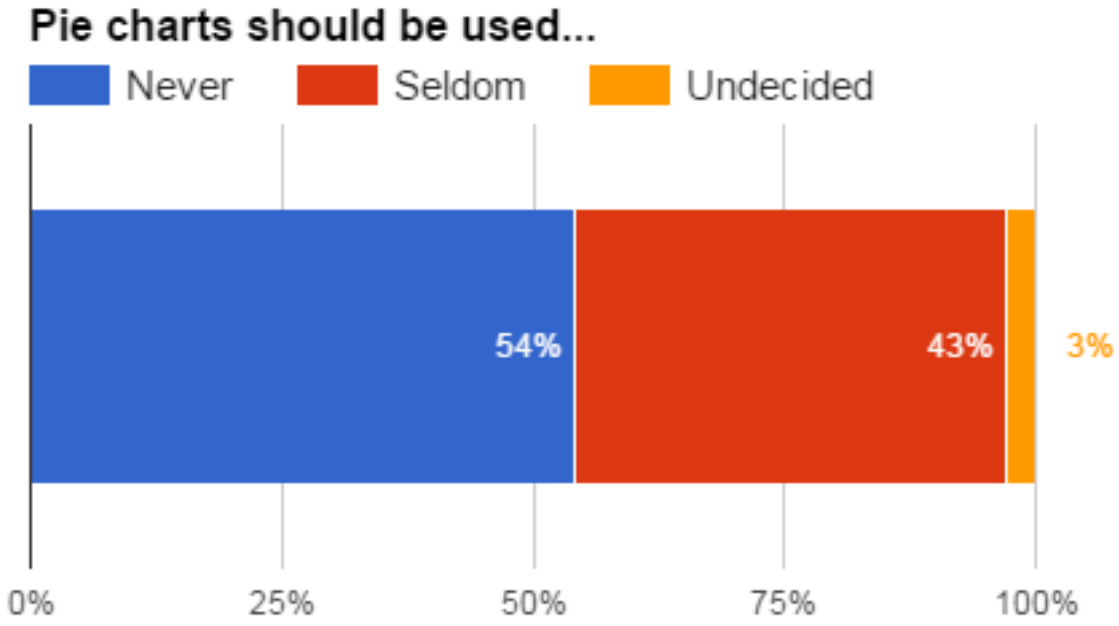
Line Charts

Proportional Data

Proportional data tracks response variables that express percentages or fractions. Its purpose is to evaluate parts to a whole.



Pie Charts



100% Stacked Bar Charts

Intro to Power BI

Why Power BI?

1. Free now, free later
2. Synergy with Power Query
3. Currently most relevant software

Create Visuals to Answer the Following:

1. Create a card to summarize total sales during 2023.

2. Create a table that shows sum of *SaleAmount* by *Region* and *ProductCategory*.

3. Compare average *ProfitMargin%* by *CustomerSegment*. Which is highest?

4. How are *total Sales* of the electronics *ProductCategory* trending during 2024?

5. Compare the proportion of *UnitsSold* by *CustomerSegment* for the West and North *Regions*. Which leverages online sales to a greater extent?

Basic DAX functions and syntax

Aggregation functions
SUM, AVERAGE, MAX, MIN, COUNTROWS

Relationships
RELATED – to reference fields in foreign tables
RELATEDTABLE – to reference foreign tables themselves

Conditional Logic
CALCULATE – performs an operation for a subset of data
IF – performs singular logical test, returns a value if true and false
SWITCH – checks multiple conditions and returns corresponding values

Variables
VAR – sets a temporary variable to store intermediate values with longer DAX formulas
RETURN – actually returns values based on calculations that can use previously defined variables

Revisiting Power Query

append

- 2023SalesInvoices
- 2023CashReceipts
- 2023AR_Adj
- Customers
- Salespeople

- 2024SalesInvoices
- 2024CashReceipts
- 2024AR_Adj
- Territories

RPA and Midterm Exam Review

Please log in to Canvas and download and extract the MidtermReview.zip folder as you arrive

ACC 6300
Advanced Data Analytics
Mason Snow, PhD

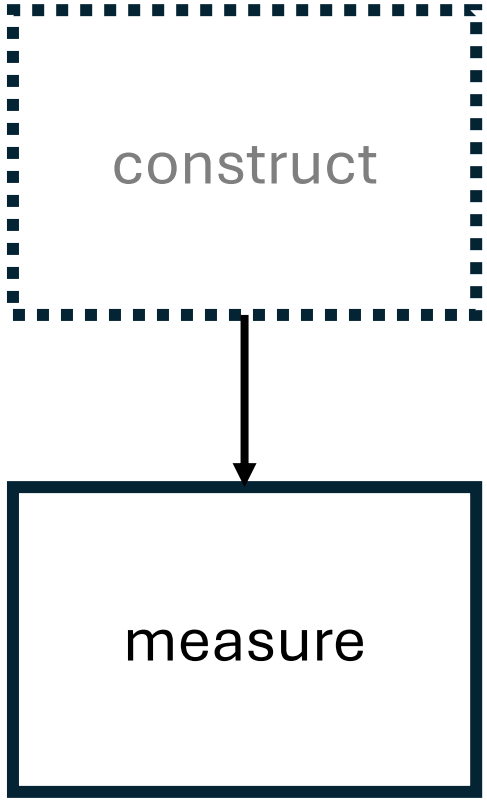
1 | Analytics Concepts

Multiple Choice Questions

AMPS Model Revisited

Accounting = measurement

Accountants work with economic constructs and distill them into measures. These measures should be comparable, consistent, decision-useful, and perhaps most importantly – VALID.



Constructs are what “count” and measures are what get counted. As data analysts, we only ever work with measures.

Internal Validity: does X cause Y?

External Validity: would an observed effect generalize to other settings?

Construct Validity: does the measure sufficiently capture the construct?

Statistical Conclusion Validity: have proper analytical procedures been followed?

What are three prerequisites for causal inference?

Temporal precedence: (x must occur before y)

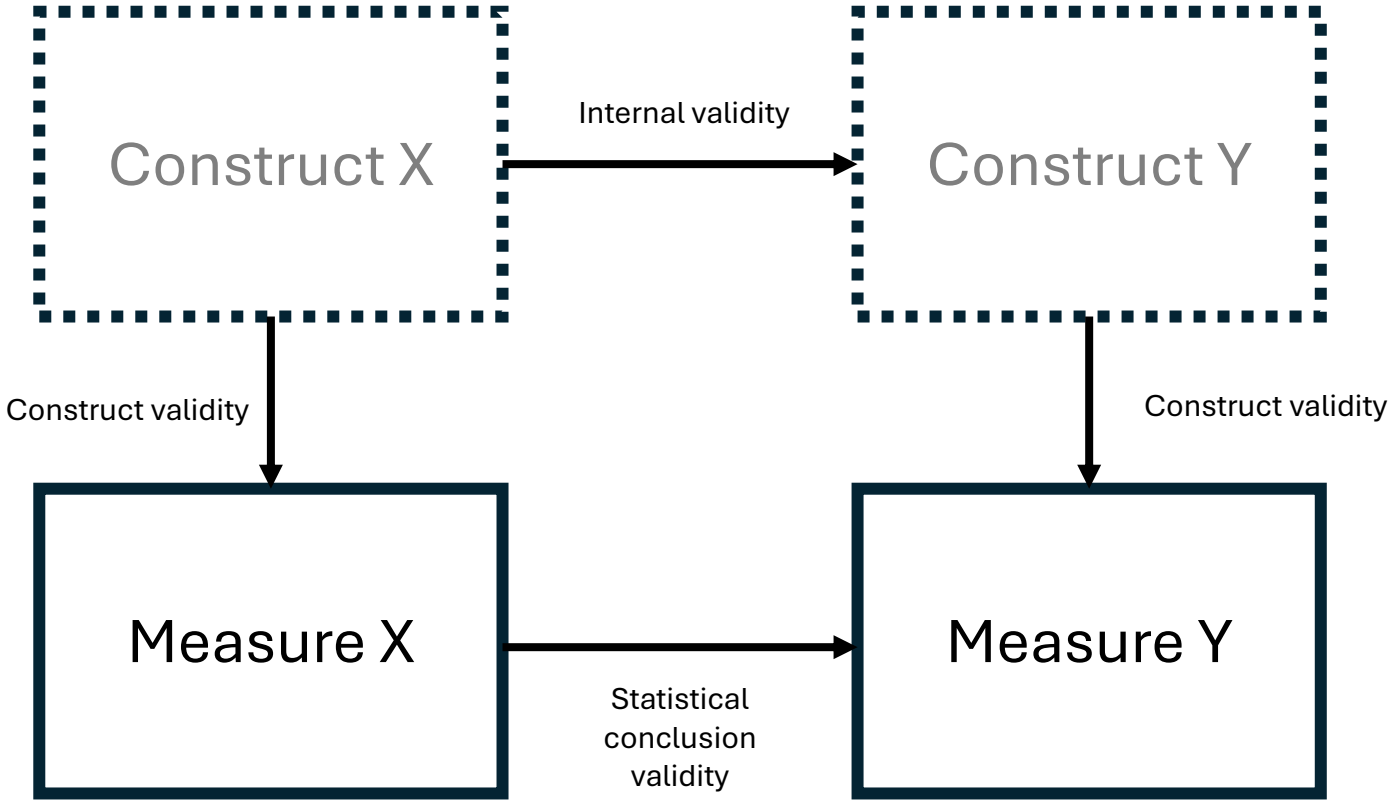
Significant correlation: (x must be related to y)

Alternative explanations must be eliminated ← the difficult part

Five Groups of Alternative Explanations

- **Correlated Omitted Variables**: Assuming that X causes Y when really Z causes both X and Y.
- **Reverse Causality**: Assuming that X causes Y when really Y causes X.
- **Selection Bias**: when the sample of individuals analyzed systematically differs from the population.
- **Measurement Error**: when values of measured observations randomly/systematically differ from true values.
- **Spurious Correlation**: when the correlation between X and Y is really just due to chance.

Libby Box Framework



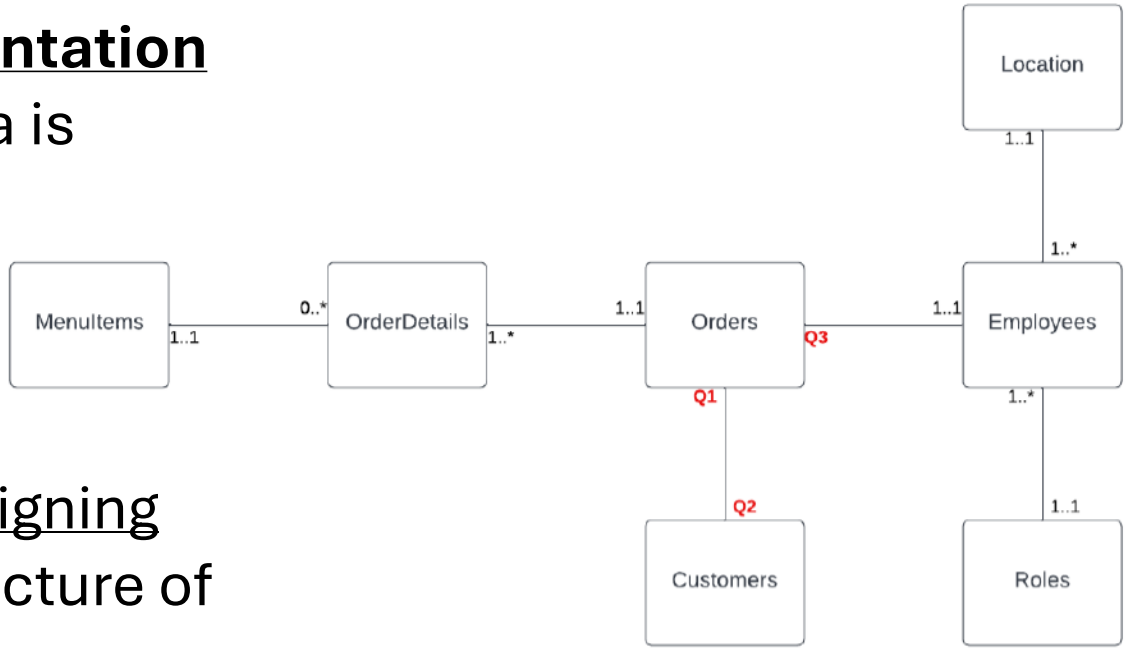
Weakness in any of these links can cast doubt on a causal relation between X and Y.

What is Data Modeling?

The process of creating a **visual representation** of information systems to show how data is stored and connected within a system.

Why?

Data models serve as a blueprint for designing and communicating the information structure of a system.



Relational Data Models

Relational databases dominate the landscape of database management systems.

Class	Attributes	Associations
• A collection of things about which an organization wants to collect and store data	• The specific facts or dimensions of a class for which we will collect and store data	• A formally stated or acknowledged relationship between two classes.

Classes and Attributes

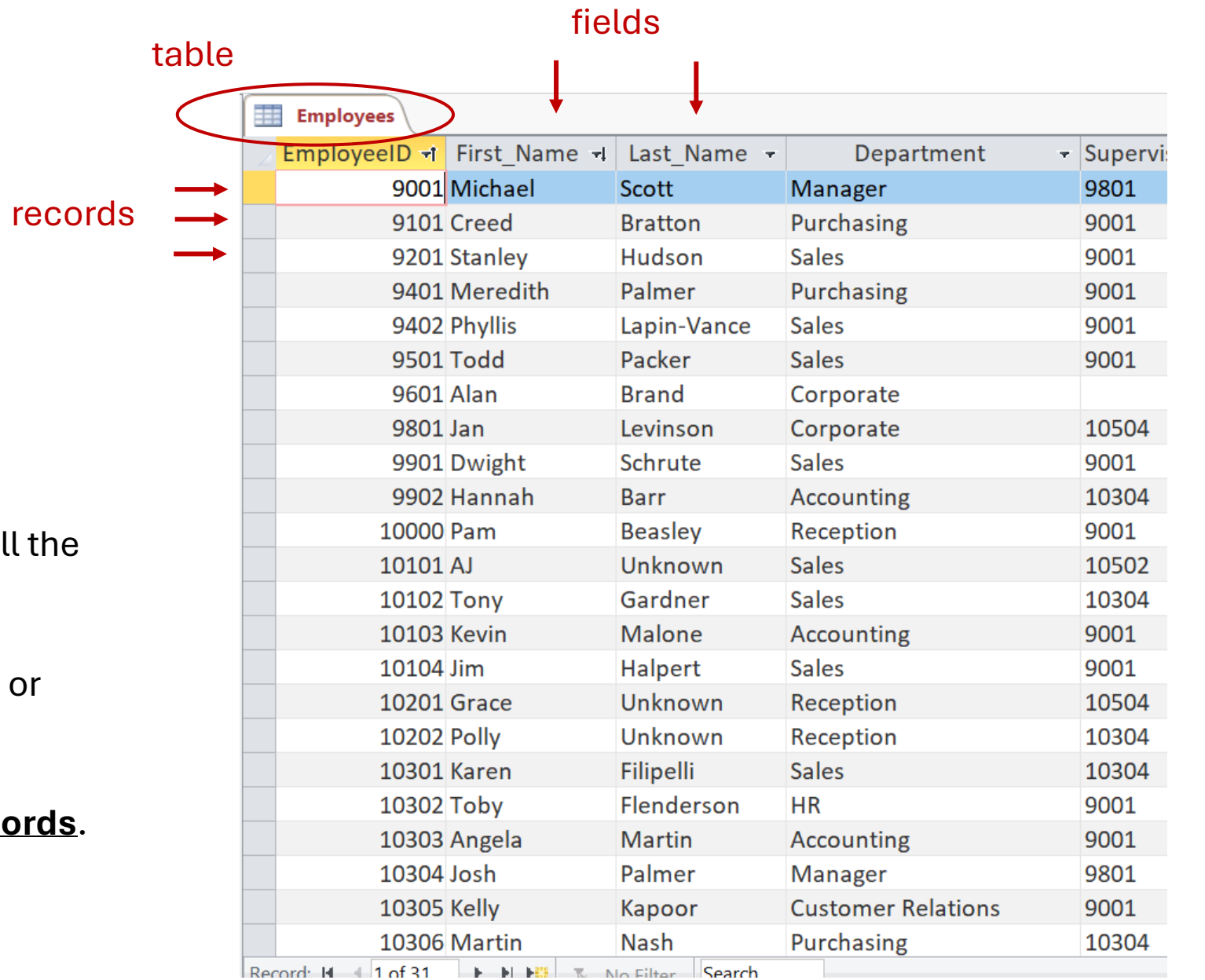
Employees

- EmployeeID
- FirstName
- LastName
- Department
- SupervisorID
- Etc.

The **class** refers to the entire **table** (all the columns and rows)

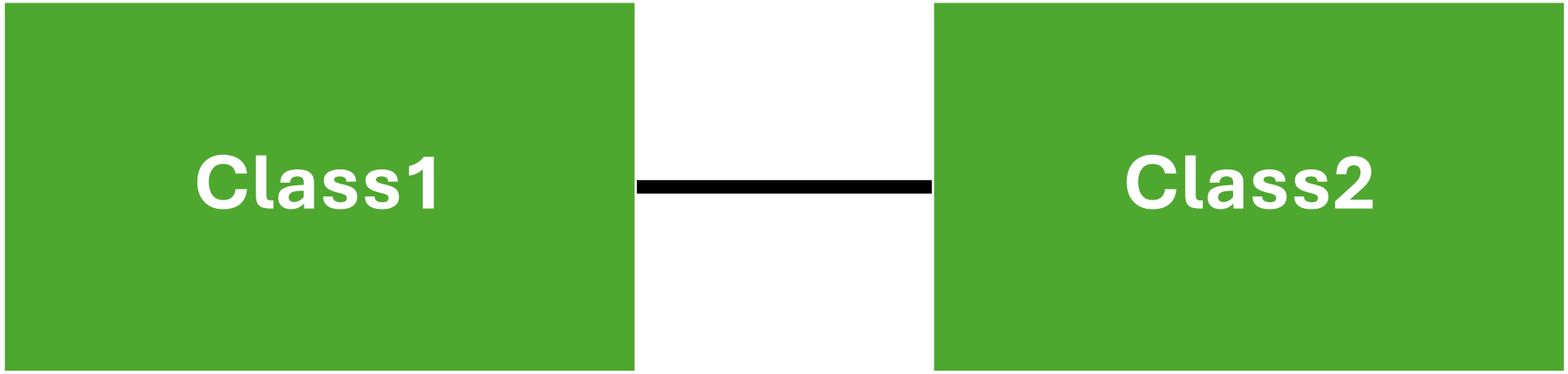
Attributes are reflected as the **fields** or columns of the table.

Rows of a table are referred to as **records**.



Associations

A formal designation of a relationship
between classes



Primary Keys

An attribute that uniquely identifies every instance in a class (i.e., a row of a table).

- Each record must have a primary key
- Values for a table's primary key must NOT repeat.

Natural primary keys are derived from existing, real-world data:

- Social Security Numbers (SSNs)
- Phone Numbers
- Vehicle Identification Numbers (VINs)

Often, the best primary keys are assigned, sequential numbers:

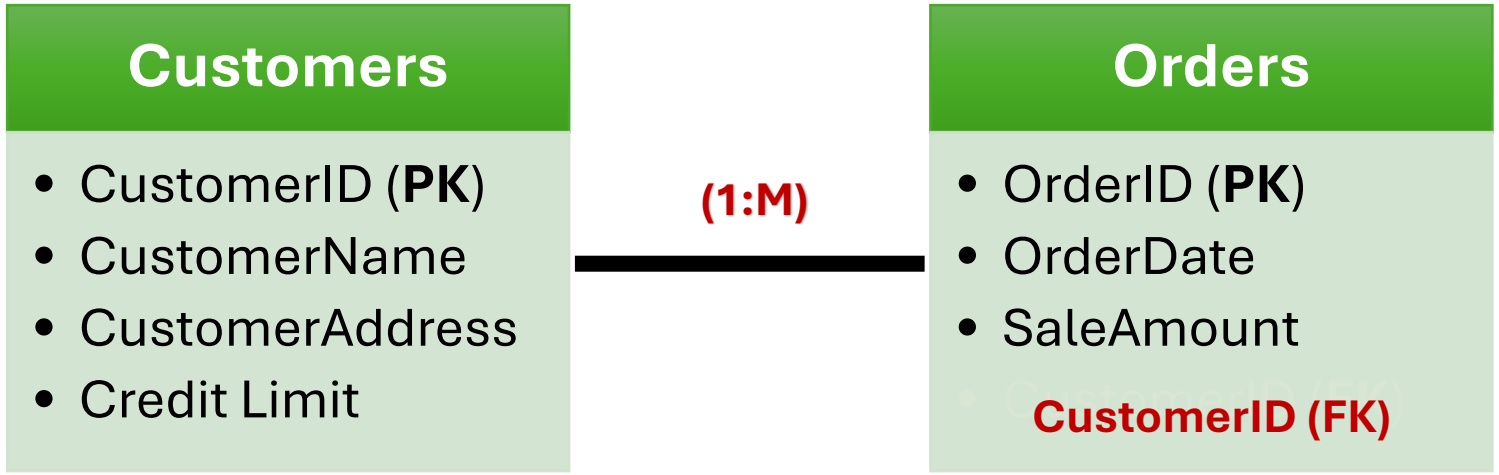
- E.g.) CUST-10023

Foreign Keys

A field in one table that is the primary key of another table in the database.

This is what links tables together!

The table that is the “many” side of a one-to-many relationship gets the foreign key.



Query Template (multiple tables)

```
SELECT   [table1.field], [table2.field]
FROM     [table1] {JOIN} [table2]
ON       [table1.field] = [table2.field]
WHERE    [field] {meets criteria}
GROUP BY [field]
ORDER BY [field];
```

Dynamic Arrays

- Introduced with Excel 2021
- A dynamic array is a feature in Excel that allows a single formula to return multiple values in multiple cells.
  - These typically “spill” into adjacent cells.

- Why? They automatically expand/contract to fit results, which often eliminates the need for manual tasks (copying formulas, performing sorts, etc.)

Dynamic Array Functions

- **UNIQUE**
  Returns a list of unique values from a range, eliminating duplicates.
- **SORT**
  Sorts data in ascending or descending order automatically.
- **SEQUENCE**
  Generates a sequence of numbers, with an optional starting value and step increment.

Dynamic Array Functions

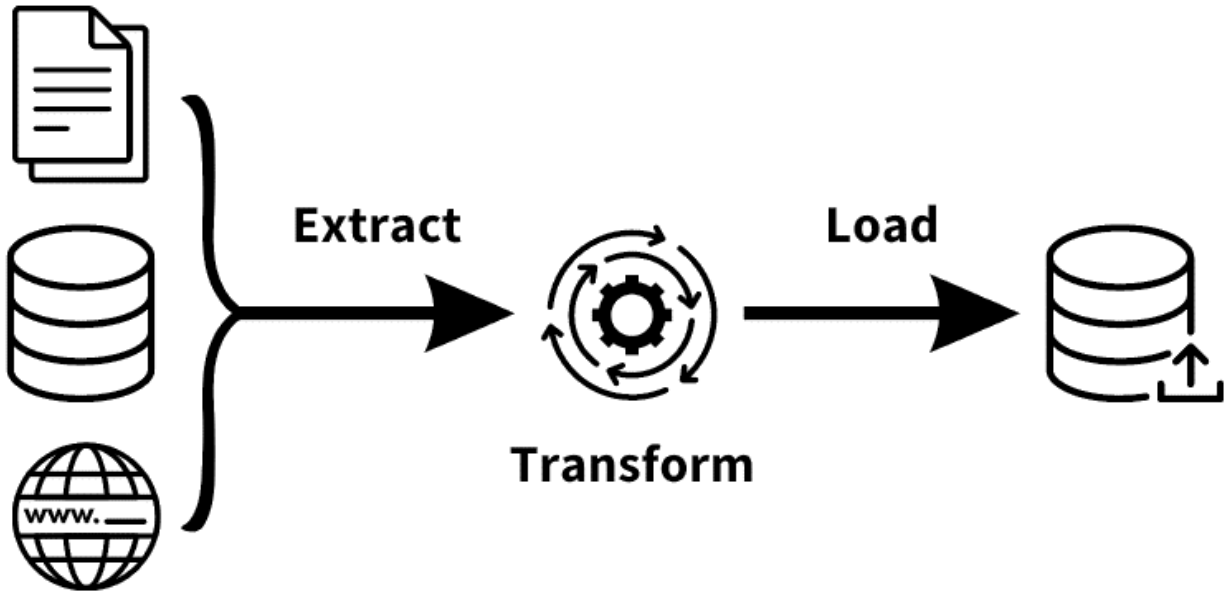
- **XLOOKUP**
  Searches for a value in a range and returns the corresponding value from another range, with additional options for matching and handling missing data.
- **TEXTSPLIT**
  Splits text into multiple cells based on a delimiter, with an option to ignore empty results.
- **FILTER**
  Filters data based on a specified condition and returns the matching results.

ETL Process – Big Picture

Consists of all the activities needed to prepare the data for analysis. This could include:

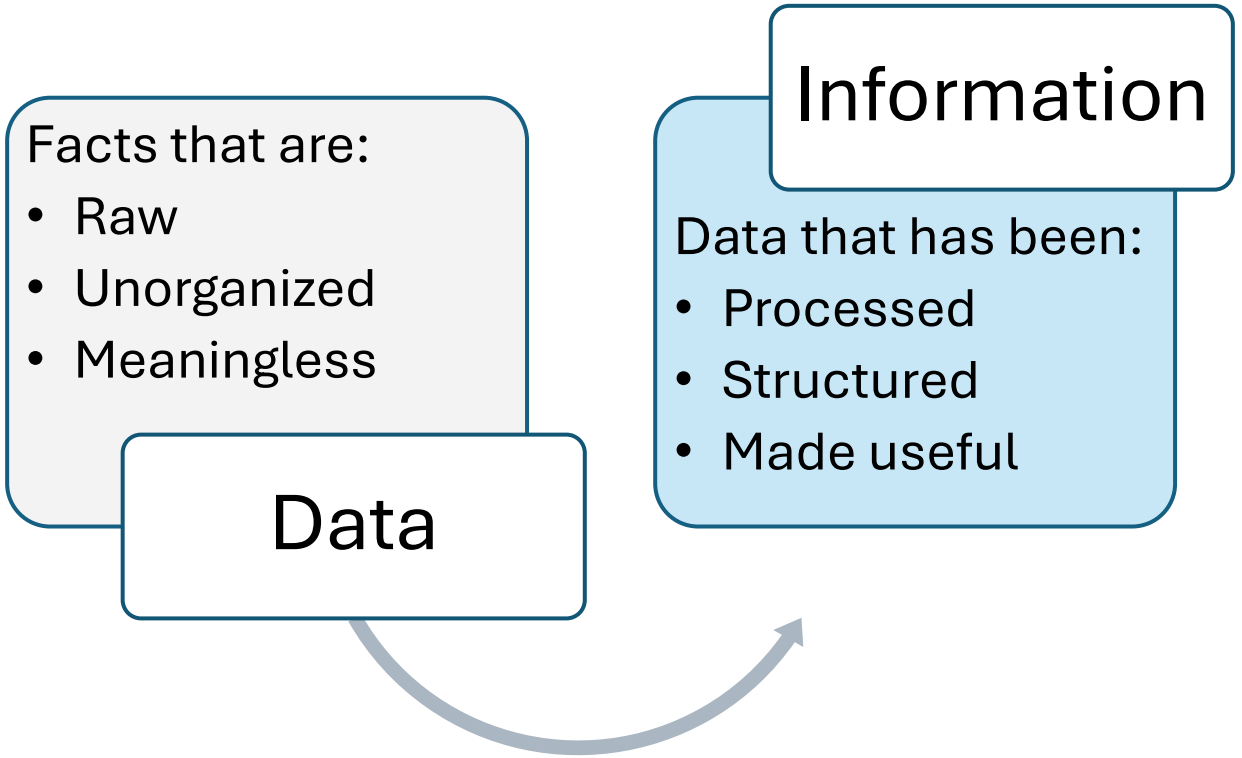
collection, cleaning, combining, structuring, transformation, profiling, formatting, etc.

Often the most time-consuming part of the whole data analytics process.



Often referred to as the “ETL” process.

Data vs Information



High Quality Information

Characteristic	Description
Accurate	Correct; free from error; accurately represents underlying facts
Available	Accessible to users when needed
Complete	Does not omit needed data; sufficient in depth and breadth
Consistent	Presented in the same format every time
Current	Includes data that is up-to-date to the present
Objective	Unbiased; impartial
Relevant	Appropriately pertains to the requested situation
Timely	Provided in time for users to make decisions
Understandable	Easily comprehended and communicated
Verifiable	Can be checked/confirmed by others; supported with documentation.

Data Structuring

Transposing

- Switching the rows and columns of the data

Pivoting

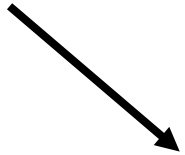
- Reshaping the data from tall to wide format. The unique values of one column are converted to columns, values are specified for aggregation.

Unpivoting

- Reshaping the data from wide to tall format. Takes multiple related columns and transforming them into a single column of values.

Transposing

County	Utah	Salt Lake	Davis	Summit
Jan	141	126	114	148
Feb	111	61	103	131
Mar	144	110	153	145
Apr	112	172	123	117
May	95	126	137	73
Jun	105	109	106	150



Changing the orientation of the dataset, switching columns and rows

County	Jan	Feb	Mar	Apr	May	Jun
Utah	141	111	144	112	95	105
Salt Lake	126	61	110	172	126	109
Davis	114	103	153	123	137	106
Summit	148	131	145	117	73	150

Pivoting

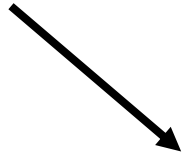
Date	Product	Sales
1/1/2024	A	$100
1/1/2024	A	$120
1/1/2024	B	$150
1/2/2024	A	$200
1/2/2024	B	$100
1/2/2024	B	$130

Converting from tall to wide format, aggregating repeat values for a designated column.

Date	A	B
1/1/2024	$220	$150
1/2/2024	$200	$230

Unpivoting

Country Name	2016	2017	2018
Argentina	0.7	0.6	1
Canada	0.2	0.2	0.2
Peru	4.6	4.5	3.6



Converting from wide to tall format, taking related columns and collapsing them into a single column

Country Name	Attribute	Value
Argentina	2016	0.7
Argentina	2017	0.6
Argentina	2018	1
Canada	2016	0.2
Canada	2017	0.2
Canada	2018	0.2
Peru	2016	4.6
Peru	2017	4.5
Peru	2018	3.6

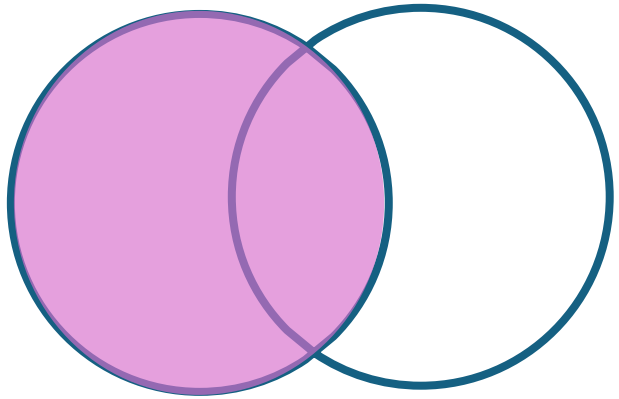
Appending

Location	OrderID	OrderDate	SaleAmount
Lindon	PO-2408	8/1/2026	$ 1,944.96
Lindon	PO-2412	8/1/2026	$ 2,064.61
Lindon	PO-2413	8/3/2026	$ 1,872.19
Lindon	PO-2414	8/4/2026	$ 1,399.68

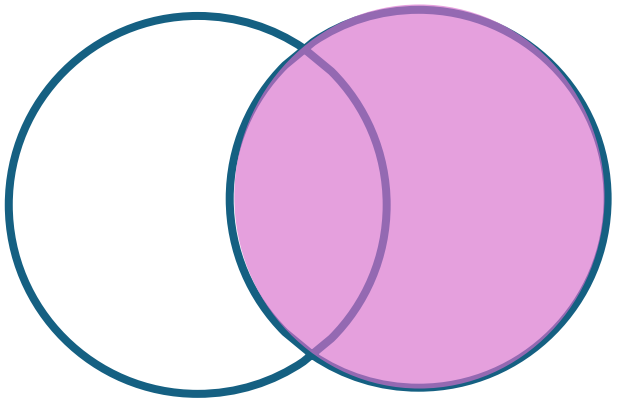
Location	OrderID	OrderDate	SaleAmount
Orem	PO-2409	8/1/2026	$ 1,739.61
Orem	PO-2410	8/2/2026	$ 2,038.60
Orem	PO-2411	8/3/2026	$ 1,659.36
Orem	PO-2415	8/4/2026	$ 1,814.87

Location	OrderID	OrderDate	SaleAmount
Lindon	PO-2408	8/1/2026	$ 1,944.96
Lindon	PO-2412	8/1/2026	$ 2,064.61
Lindon	PO-2413	8/3/2026	$ 1,872.19
Lindon	PO-2414	8/4/2026	$ 1,399.68
Orem	PO-2409	8/1/2026	$ 1,739.61
Orem	PO-2410	8/2/2026	$ 2,038.60
Orem	PO-2411	8/3/2026	$ 1,659.36
Orem	PO-2415	8/4/2026	$ 1,814.87

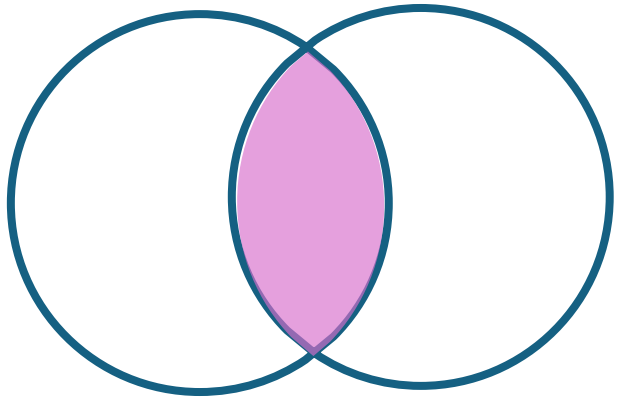
Four Major Types of Joins



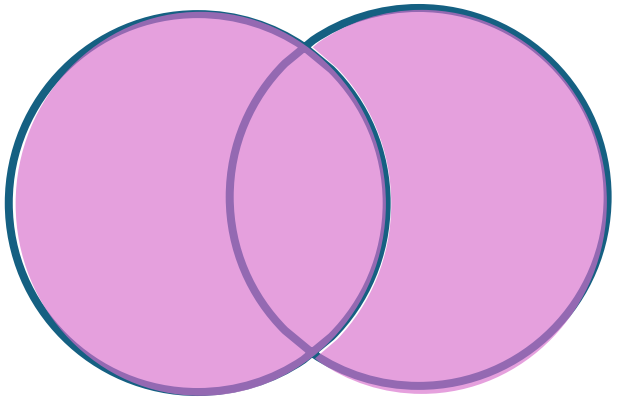
Left Join
Retains all records from the first table, and only matches from the second.



Right Join
Retains all records from the second table, and only matches from the first.

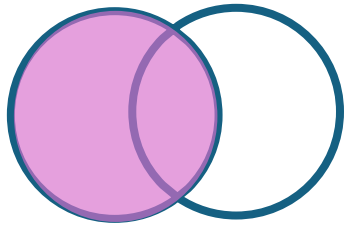


Inner Join
Retains only matches between the two tables.



Full Outer Join
Retains all records of both tables, with matching records being joined where possible.

Left Join

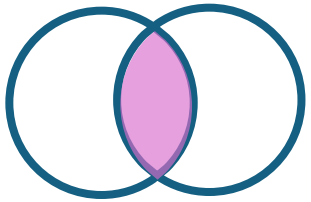


LocationID	OrderID	OrderDate	SaleAmount
1	PO-2408	8/1/2026	$ 1,944.96
3	PO-2412	8/1/2026	$ 2,064.61
1	PO-2413	8/3/2026	$ 1,872.19
2	PO-2414	8/4/2026	$ 1,399.68
3	PO-2409	8/1/2026	$ 1,739.61
1	PO-2410	8/2/2026	$ 2,038.60
2	PO-2411	8/3/2026	$ 1,659.36
2	PO-2415	8/4/2026	$ 1,814.87

LocationID	LocationName
1	Lindon
2	Orem
4	Provo

LocationID	OrderID	OrderDate	SaleAmount	LocationName
1	PO-2408	8/1/2026	$ 1,944.96	Lindon
3	PO-2412	8/1/2026	$ 2,064.61	
1	PO-2413	8/3/2026	$ 1,872.19	Lindon
2	PO-2414	8/4/2026	$ 1,399.68	Orem
3	PO-2409	8/1/2026	$ 1,739.61	
1	PO-2410	8/2/2026	$ 2,038.60	Lindon
2	PO-2411	8/3/2026	$ 1,659.36	Orem
2	PO-2415	8/4/2026	$ 1,814.87	Orem

Inner Join

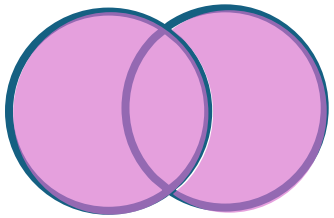


LocationID	OrderID	OrderDate	SaleAmount
1	PO-2408	8/1/2026	$ 1,944.96
3	PO-2412	8/1/2026	$ 2,064.61
1	PO-2413	8/3/2026	$ 1,872.19
2	PO-2414	8/4/2026	$ 1,399.68
3	PO-2409	8/1/2026	$ 1,739.61
1	PO-2410	8/2/2026	$ 2,038.60
2	PO-2411	8/3/2026	$ 1,659.36
2	PO-2415	8/4/2026	$ 1,814.87

LocationID	LocationName
1	Lindon
2	Orem
4	Provo

LocationID	OrderID	OrderDate	SaleAmount	LocationName
1	PO-2408	8/1/2026	$ 1,944.96	Lindon
1	PO-2413	8/3/2026	$ 1,872.19	Lindon
2	PO-2414	8/4/2026	$ 1,399.68	Orem
1	PO-2410	8/2/2026	$ 2,038.60	Lindon
2	PO-2411	8/3/2026	$ 1,659.36	Orem
2	PO-2415	8/4/2026	$ 1,814.87	Orem

Full Outer Join



LocationID	OrderID	OrderDate	SaleAmount
1	PO-2408	8/1/2026	$ 1,944.96
3	PO-2412	8/1/2026	$ 2,064.61
1	PO-2413	8/3/2026	$ 1,872.19
2	PO-2414	8/4/2026	$ 1,399.68
3	PO-2409	8/1/2026	$ 1,739.61
1	PO-2410	8/2/2026	$ 2,038.60
2	PO-2411	8/3/2026	$ 1,659.36
2	PO-2415	8/4/2026	$ 1,814.87

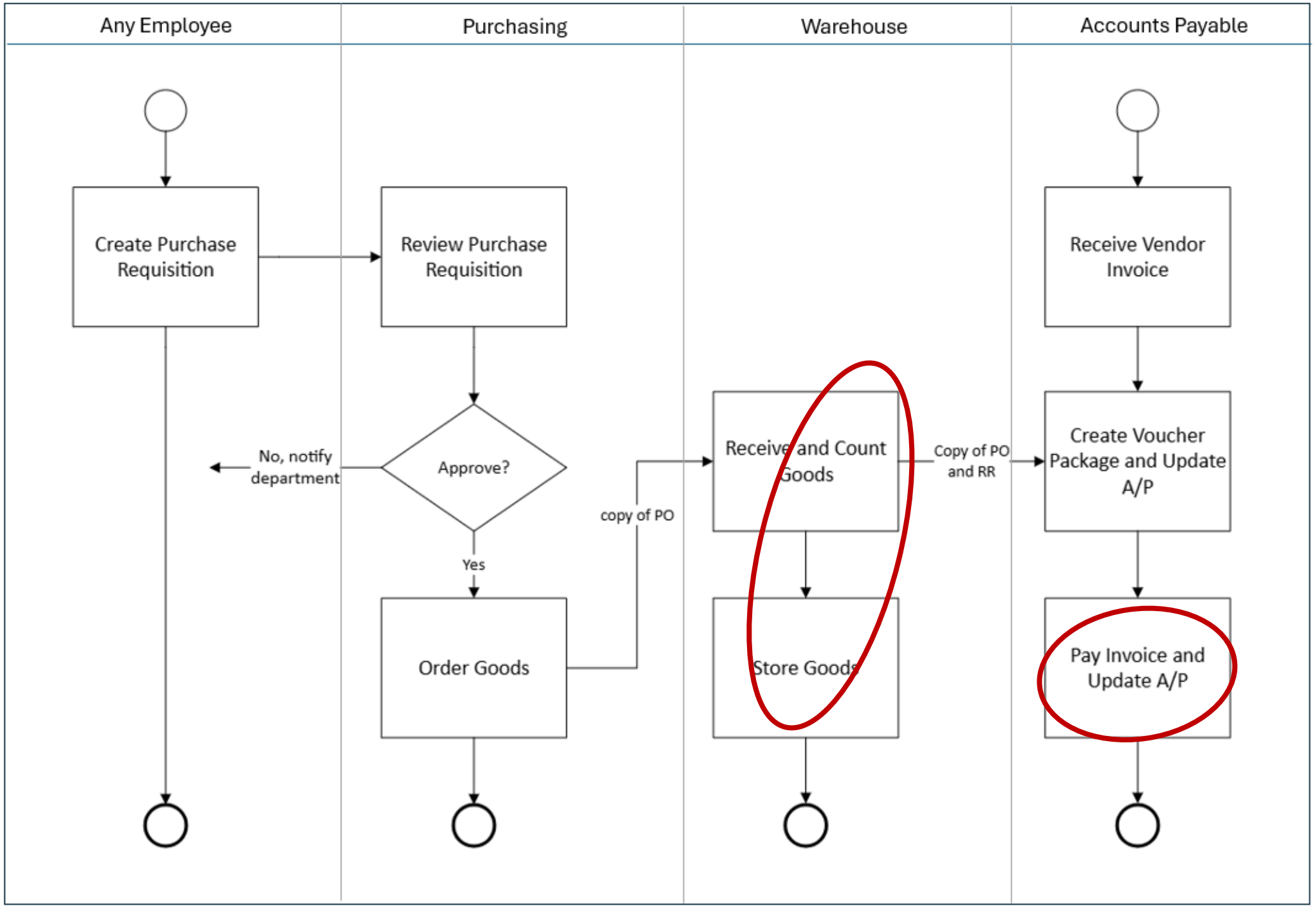
LocationID	LocationName
1	Lindon
2	Orem
4	Provo

LocationID	OrderID	OrderDate	SaleAmount	LocationName
1	PO-2408	8/1/2026	$ 1,944.96	Lindon
3	PO-2412	8/1/2026	$ 2,064.61	
1	PO-2413	8/3/2026	$ 1,872.19	Lindon
2	PO-2414	8/4/2026	$ 1,399.68	Orem
3	PO-2409	8/1/2026	$ 1,739.61	
1	PO-2410	8/2/2026	$ 2,038.60	Lindon
2	PO-2411	8/3/2026	$ 1,659.36	Orem
2	PO-2415	8/4/2026	$ 1,814.87	Orem
4				Provo

Segregation of Duties

Requiring certain tasks be performed by separate individuals. Within a business process, individuals should not perform duties across more than one of the following areas:

Custody (of assets)
Authorization (of procedures)
Recording (of information)


Any Employee
Purchasing
Warehouse
Accounts Payable
Create Purchase Requisition
Review Purchase Requisition
Approve?
No, notify department
Yes
Order Goods
copy of PO
Receive and Count Goods
Store Goods
Copy of PO and RR
Receive Vendor Invoice
Create Voucher Package and Update A/P
Pay Invoice and Update A/P

Process Mining – Event Logs

Process mining platforms retrieve data from **event logs** to produce insights.

Three required data inputs:

- **Case ID**: a unique reference to identify each instance of a cycle flow. (**WHICH**)
- **Activity**: a description of the process that the instance has undergone (**WHAT**)
- **Timestamp**: a record of when the case went through an activity (**WHEN**)

Event Log Example

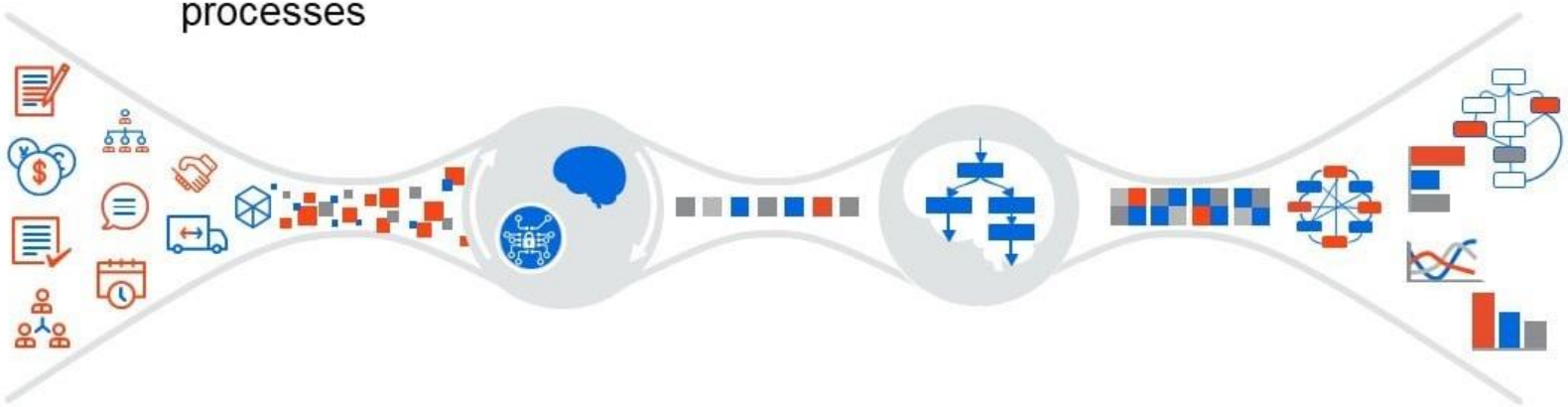
CaseID	Activity	Timestamp	Employee	Site
1001	Receive order	1/1/2022 8:00am	Tara (Sales)	CA
1001	Check credit	1/1/2022 8:03am	Liwei (Manager)	CA
1001	Pack order	1/1/2022 10:38am	Scott (Inventory)	CA
1002	Receive order	1/1/2022 10:45am	Tara (Sales)	CA
1002	Check credit	1/1/2022 10:59am	Liwei (Manager)	CA
1003	Receive order	1/1/2022 2:47pm	Carlos (Sales)	AZ
1001	Ship order	1/1/2022 2:58pm	Scott (Inventory)	CA
1003	Pack order	1/1/2022 3:12pm	Ana (Inventory)	AZ
1001	Send Invoice	1/2/2022 8:04am	Liwei (Manager)	CA
1004	Receive order	1/2/2022 10:12am	Carlos (Sales)	AZ
1004	Pack order	1/2/2022 12:42pm	Ana (Inventory)	AZ
1003	Ship order	1/2/2022 2:26pm	Ana (Inventory)	AZ

Start to Finish

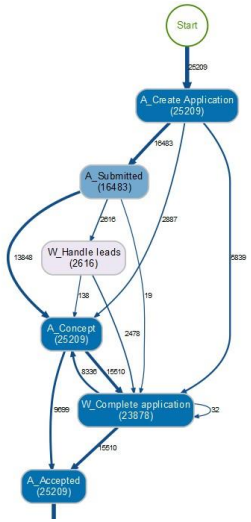
Reads
data related to your processes

Transforms
data to event log

Visualizes
the end-to-end process



CaseID	Activity	Timestamp	Employee	Site
1001	Receive order	1/1/2022 8:00am	Tara (Sales)	CA
1001	Check credit	1/1/2022 8:03am	Liwei (Manager)	CA
1001	Pack order	1/1/2022 10:38am	Scott (Inventory)	CA
1002	Receive order	1/1/2022 10:45am	Tara (Sales)	CA
1002	Check credit	1/1/2022 10:59am	Liwei (Manager)	CA
1003	Receive order	1/1/2022 2:47pm	Carlos (Sales)	AZ
1001	Ship order	1/1/2022 2:58pm	Scott (Inventory)	CA
1003	Pack order	1/1/2022 3:12pm	Ana (Inventory)	AZ
1001	Send Invoice	1/2/2022 8:04am	Liwei (Manager)	CA
1004	Receive order	1/2/2022 10:12am	Carlos (Sales)	AZ
1004	Pack order	1/2/2022 12:42pm	Ana (Inventory)	AZ
1003	Ship order	1/2/2022 2:26pm	Ana (Inventory)	AZ

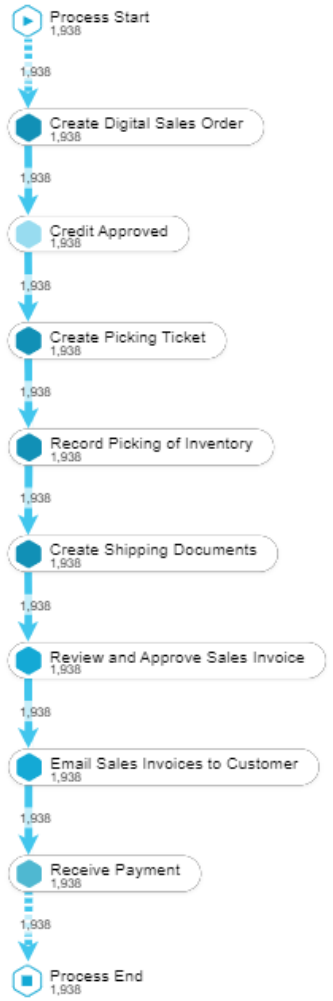


Variant Analysis

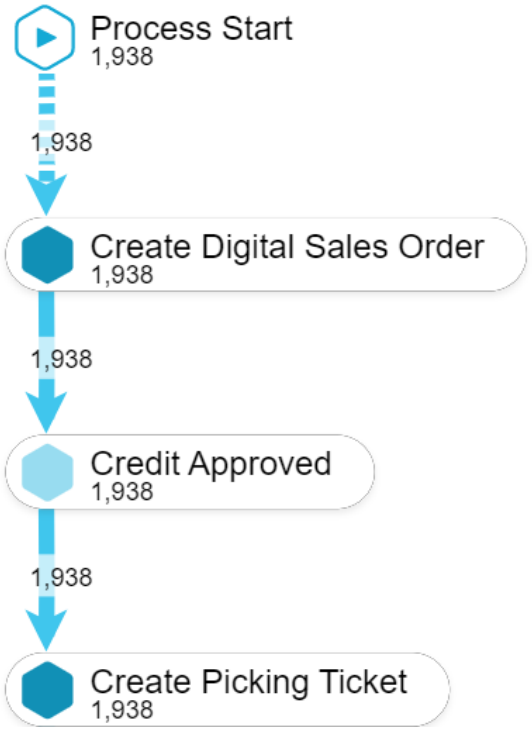
A **variant** is a unique sequence of activities taken by at least one case.

The number of times that an activity is performed is listed within each bubble. The number of times a case proceeds from one activity to the next is listed within each arrow.

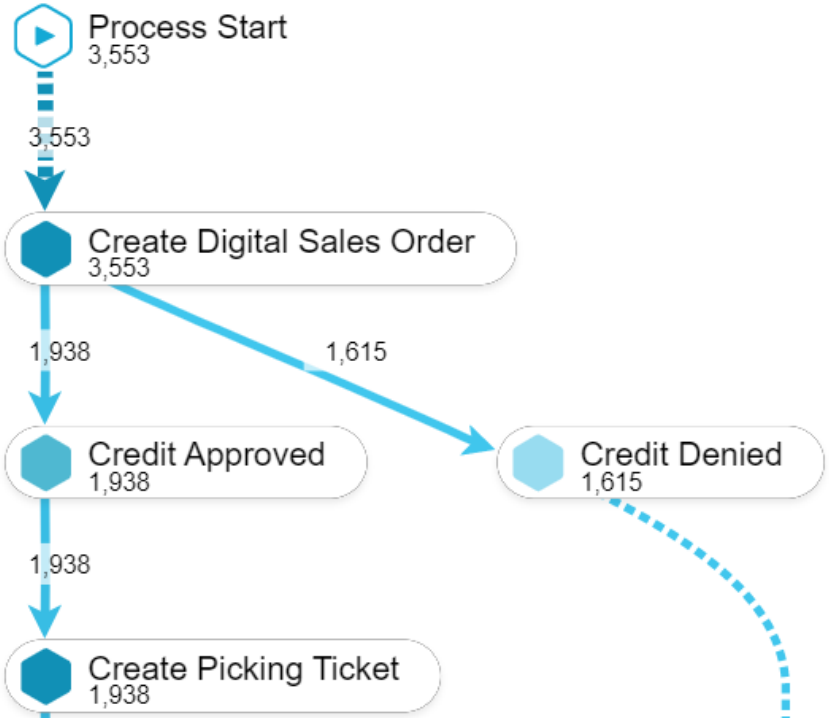
The most common variant is called the “happy path.”



Variant Analysis



One variant



Two variants

Limitations of Process Mining

1. The entire organization needs to sufficiently support providing the necessary data to create event logs.

2. Not all processes will be recorded (and hence cannot be analyzed).

3. Analysis of complicated processes (i.e., with many variants) can be unwieldy.

2 | Class Discussion

Statistical Concepts

Discussion Questions from Readings

- What are the four types of **validity**? How do these relate to the predictive validity framework (Libby boxes)?
- What are common **alternative explanations** for why a significant correlation does not imply causation?
- What type of **biases** yield the garbage-in, garbage-out problem?
- What is probability? What **probability-related errors** do humans often commit?
- What are p-values? What are common **misunderstandings relating to p-values**?

Sample Question

An auditor is evaluating whether a going concern opinion is warranted for a client experiencing financial distress. To support their decision, the auditor conducts an analysis of peer companies that faced financial distress in the past and recovered. The auditor notes that the client's financial ratios appear weaker in comparison to these peer firms. Based on this analysis, the auditor concludes that the client is at a higher risk of failure and may require a going concern opinion.

Identify and explain any potential bias in the auditor's analysis. Discuss how this could lead to a flawed conclusion and suggest a way to improve the analysis.

Sample Responses

The auditor's analysis is affected by **survivorship bias** because it only includes companies that recovered from financial distress, ignoring those that failed. This skews the comparison and may lead the auditor to overestimate the client's risk of failure simply because its financial ratios appear weaker than those of surviving firms. To improve the analysis, the auditor should **also include companies that did not survive** financial distress.

Full credit 15/15

Sample Responses

The auditor's analysis is affected by **selection bias** because they only selected a subset of companies. They could improve the analysis by looking at a **broader sample** of companies.

Not sufficiently specific, reasoning is underdeveloped

10/15

Sample Responses

The auditor's analysis is affected by **selection bias** because they only selected a subset of companies.

Not sufficiently specific, reasoning is underdeveloped, answer is incomplete

5/15

Sample Responses

The auditor’s analysis might be off base because financial ratios are not a good indicator of a company’s future success. The client may have other strengths, such as strong management or industry growth.

Not relevant to the ‘bias’ issue, answer is incomplete

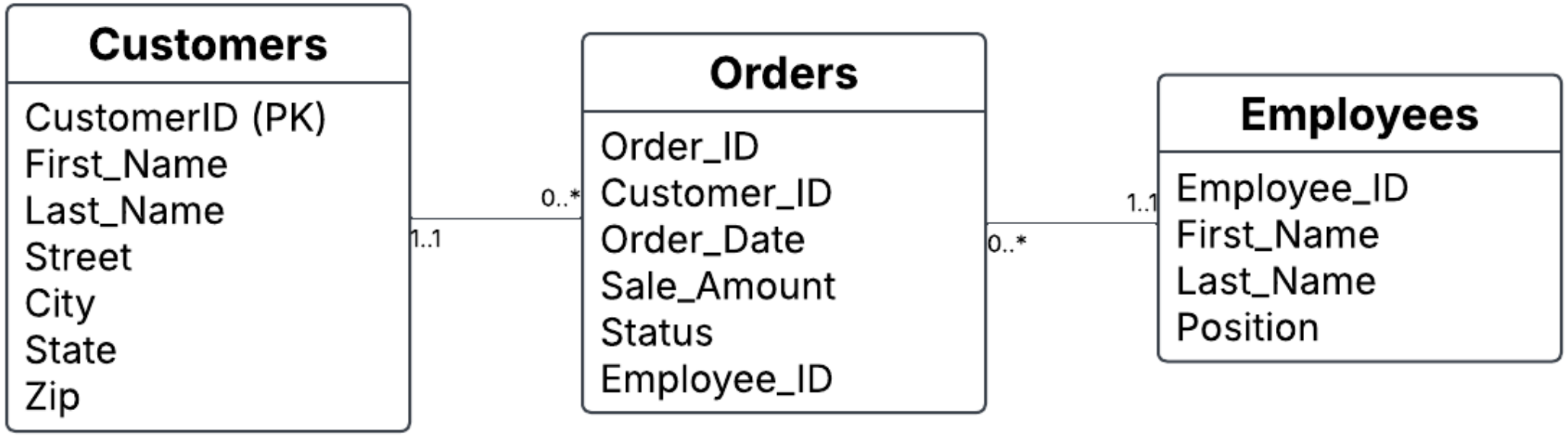
0/15

3 | Analytics Skills

Practice Questions

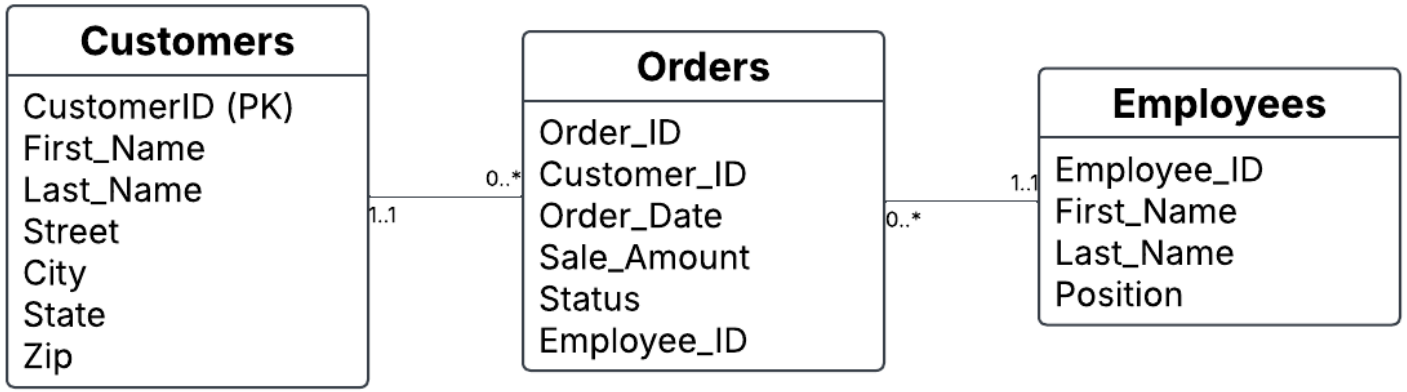
Practice 1 - SQL

The exam will ask for discrete answers to questions. After each question, you will be prompted to optionally provide your SQL code for potential partial credit.



Practice 1 - SQL

The exam will ask for discrete answers to questions. After each question, you will be prompted to optionally provide your SQL code for potential partial credit.



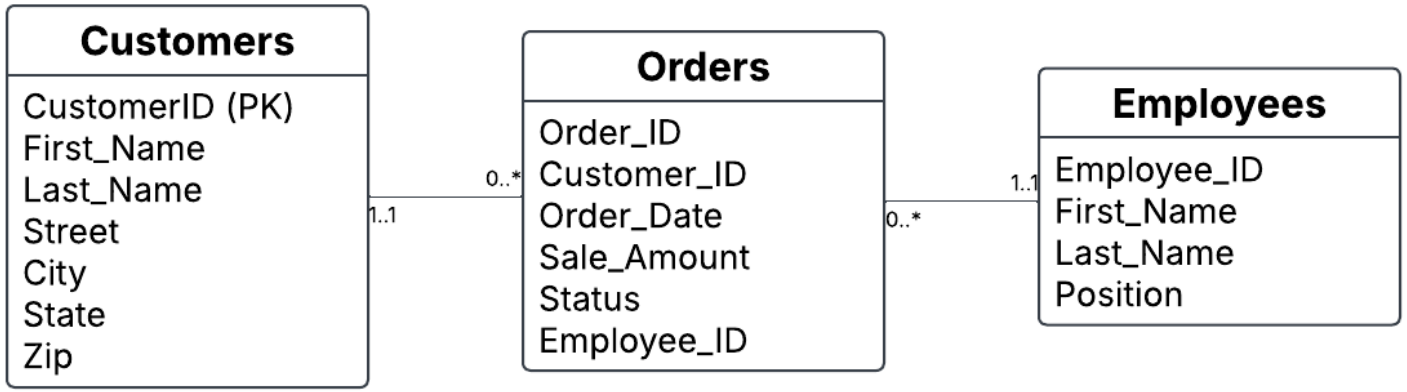
Generate a comprehensive list of information for customers located in the state of Utah (UT), sorted in alphabetical order by last name.

Q1: Which customer appears first on this list?

	Customer_ID	First_Name	Last_Name	Street	City	State	Zip
1	C0025	Brandon	Christian	22031 Lee Bypass	Barretthaven	UT	77872

Practice 1 - SQL

The exam will ask for discrete answers to questions. After each question, you will be prompted to optionally provide your SQL code for potential partial credit.



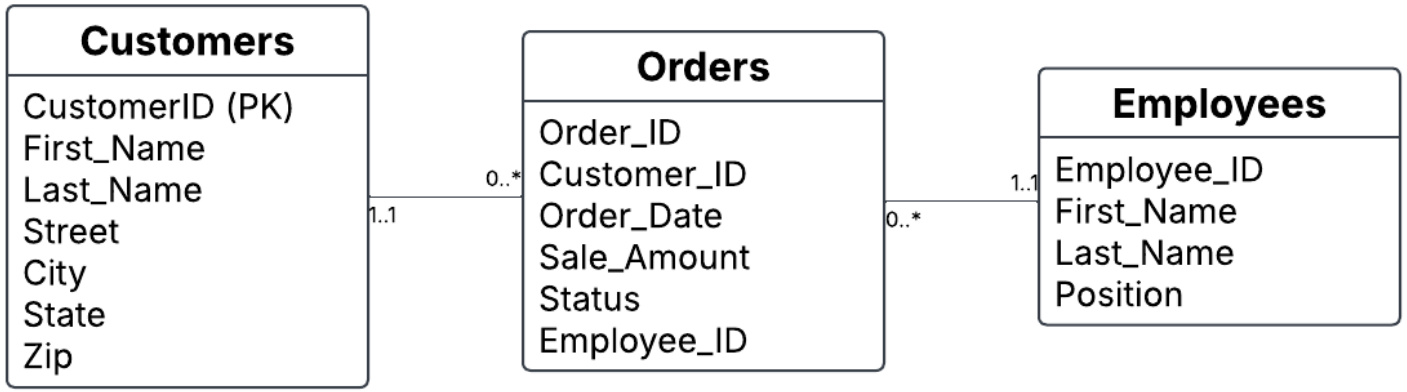
Generate a list of the count of employees by Position.

Q2: How many Warehouse Staff are there?

	Position	Employee_Count
1	Customer Support	20
2	Manager	30
3	Sales Associate	100
4	Warehouse Staff	50

Practice 1 - SQL

The exam will ask for discrete answers to questions. After each question, you will be prompted to optionally provide your SQL code for potential partial credit.



Generate a list of the total sales (the sum of Sale_Amount) by State

Q3: What are total sales in Texas (TX)?

	State	Total_Sales
1	TX	59949.43

Practice 2 – Relational Data Analysis

You can use either Power Query, Power Pivot, or a combination of the two to answer these questions.

You need to know how to:

1. Connect to data
2. Establish relationships
3. Calculate fields across tables

Practice 2 – Relational Data Analysis

Q4: What were <u>total</u> sales in Massachusetts (MA)?

Row Labels	Sum of SaleAmount
MA	$ 485,000.00

Q5: Which Region had the highest <u>total</u> sales?

Row Labels	Sum of SaleAmount
East	$ 1,421,380.00
South	$ 1,189,810.00
West	$ 1,052,240.00
Midwest	$ 820,100.00

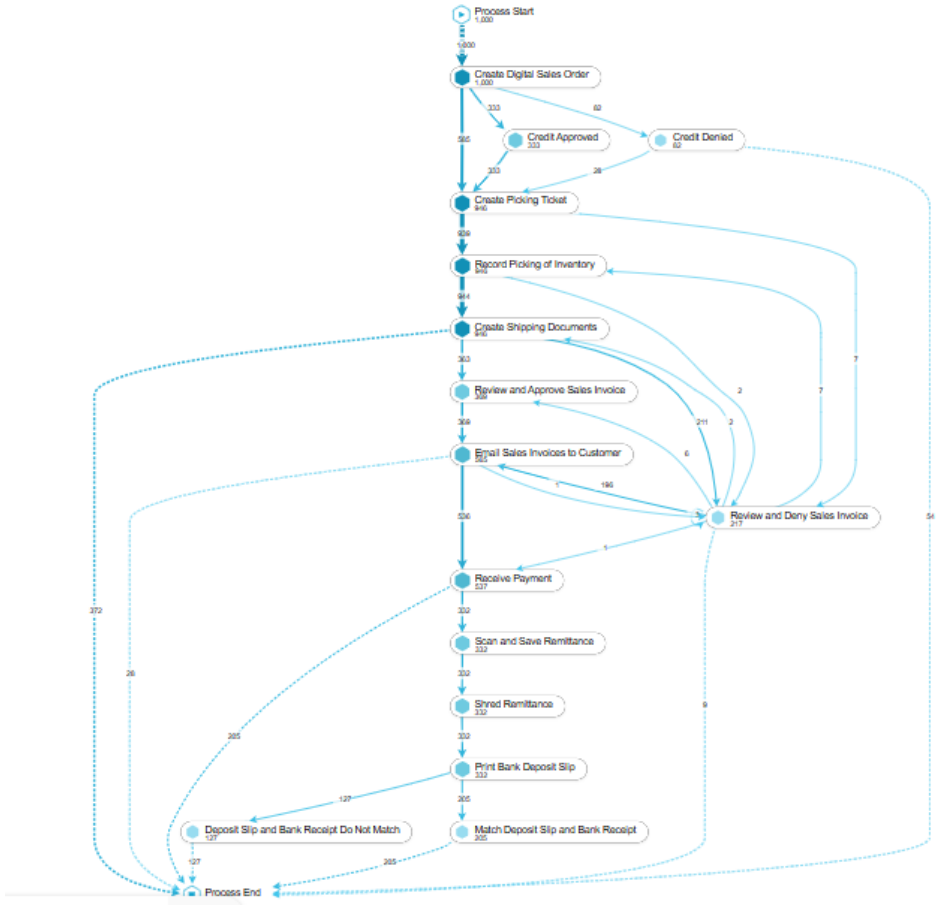
Q6: Which customer (*BusinessName*) is associated with the greatest <u>number</u> of sales?

Row Labels	Count of SalesID
Stanley Inc	7

Practice 3 – Profile, Transform, Automate

This practice is harder than what you will see on the exam!

Practice 4 – Process Mining



Q10: How many cases directly flow from 'Credit Denied' to 'Create Picking Ticket?'

Q11: Identify a violation of segregation of duties at the company. Justify your answer.

Review Slides – Final Exam

ACC 6300
Advanced Data Analytics

Mason Snow, PhD

Final Exam Structure and Policies

- Exam window: **Monday, December 8 - Wednesday, December 10**
- You will have **100 minutes** to answer **12 questions**
  - Multiple choice, numerical answer, free response
- Covered topics will only include material since the midterm exam (i.e., regression through data visualization).

Final Exam Structure and Policies

- Exam is open-note, open-internet, HOWEVER
- The only permitted AI tool during the exam **is Copilot**
- You may not share ANY information about the exam until after the exam window has closed.

- On the exam, you will acknowledge your understanding and commitment to abide by these policies.

Essential Skills

- Perform regressions and interpret the output
  - Interpret regression intercept, slope coefficients, p-values
  - Use backwards elimination as a regression technique
  - Construct an equation to predict Y outcomes as a function of X inputs
- Apply predictive techniques within accounting contexts such as:
  - Evaluating earnings properties (i.e., persistence, discretionary accruals)
  - Forecasting using prediction models (e.g., Altman Z) or Monte Carlo simulations
  - Analyzing cost behavior with one or more cost drivers
- Construct visualizations (including matrices) in Power BI to answer questions
  - Use appropriate visuals, apply filters, handle date hierarchies
- Identify and explain principles of highly effective data visualization

Linear Regression Model

$$y = \alpha + \beta x + \varepsilon$$

Dependent variable (*y*): The variable we are trying to predict or explain.

Independent variable (*x*): A variable we think may explain *y*.

Intercept (α)*:* The predicted value of *y* for when *x* equals zero.

Slope coefficient (β): Estimate of how *y* will change with a one-unit increase in *x*.

Error term (ε): The part of *y* that remains unexplained by *x*.

Multiple Linear Regression

$$y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_n x_n + \varepsilon$$

Multiple linear regression allows for more than one independent variable.

What is the benefit of including multiple X variables in one regression?

Each slope coefficient reflects the effect of the individual X_i on Y, holding all other independent variables **constant**. This is sometimes referred to as "controlling for" other X variables.

Interpreting Output in Excel

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.774478657
R Square	0.599817191
Adjusted R Square	0.551795253
Standard Error	19.70540981
Observations	100

The percent of the variation in Y explained by variation in X.

Number of rows in the dataset

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	9989.608058	3329.869353	12.05236	7.98429E-05
Residual	35	9660.142681	276.0040766		
Total	38	17749.74359			

Don't worry about the ANOVA section

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-7.681253169	12.79255739	-0.600447036	0.553613	-34.02801831	18.66551197
X1	0.693146954	0.343142056	2.032486533	0.045688	0.221525846	1.164768061
X2	-0.193986362	0.225856579	-0.858891792	0.398561	-0.659146693	0.27117397
X3	0.302458404	0.101901463	2.968145855	0.006518	0.092588413	0.512328395

The explanatory variables included in the regression

The coefficient estimates for each of the variables (and the intercept)

P-values for each estimate ($p<0.05$ can be considered statistically significant)

Backwards Elimination

One method to set up a prediction model is using ‘backwards elimination.’ This is a process of starting with a full model of independent variables and iteratively removing insignificant variables until all remaining are statistically significant.

1. Run a regression with a series of X variables that plausibly can help predict Y.
2. Identify which X variables were statistically significant predictors.
3. Run another regression with only the X variables from Step 2.
4. Repeat until the model contains only significant X variables.

Earnings Persistence

A measure of how well current earnings predict future earnings.

Higher earnings persistence indicates stable and predictable earnings over time.

Investors prefer firms with high earnings persistence.

Earnings Persistence

$$NetIncome_{t+1} = \beta_0 + \beta_1 NetIncome_t + \varepsilon$$

β_1, the earnings persistence coefficient, is typically a number between 0 and 1.

The higher the β, the more persistent the earnings.

Discretionary Accruals

The Jones (1991) Model

TotalAccruals is the difference between Net Income and Cash Flow from Operations.

$$\frac{TotalAccruals_t}{Assets_{t-1}} = \beta_0 + \beta_1 \frac{1}{Assets_{t-1}} + \beta_2 \frac{\Delta Sales_t}{Assets_{t-1}} + \beta_3 \frac{PP\&E_t}{Assets_{t-1}} + \varepsilon$$

The explanatory (X) variables used in this model indicate factors as to why accruals will vary ***WITHOUT*** the use of accounting discretion (i.e., changes in sales, depreciation expense).

Discretionary Accruals

The Jones (1991) Model

$$\frac{TotalAccruals_t}{Assets_{t-1}} = \beta_0 + \beta_1 \frac{1}{Assets_{t-1}} + \beta_2 \frac{\Delta Sales_t}{Assets_{t-1}} + \beta_3 \frac{PP\&E_t}{Assets_{t-1}} + \boldsymbol{\varepsilon}$$

The regression output we care about is the estimated residual!
This is a measure of 'unexplained' or "discretionary" accruals.

Altman Z-Score

A model invented in 1968 by Edward Altman to predict bankruptcy risk.

Designed to assess the likelihood of a firm going bankrupt within the next two years.

While others have improved upon the original model, it is still a widely used model.

Altman Z-Score

Equation for Altman's Z-Score Model (1968):

$$Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 1X_5$$

X_1 = Working Capital / Total Assets

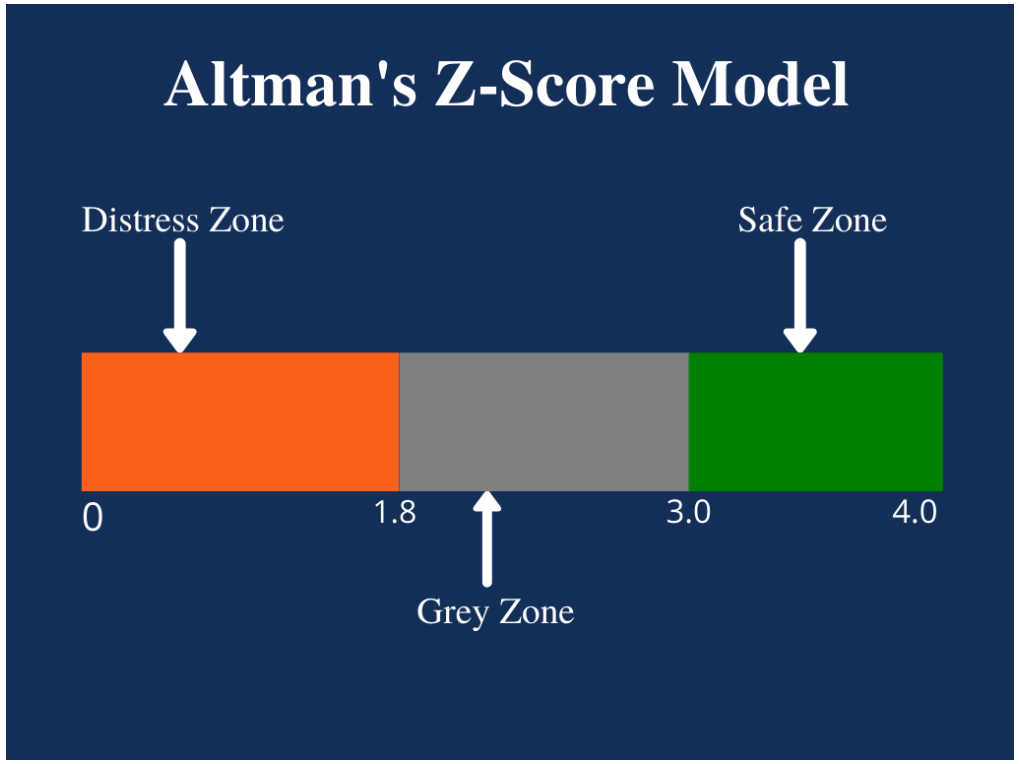
X_2 = Retained Earnings / Total Assets

X_3 = Earnings Before Interest & Tax (EBIT) / Total Assets

X_4 = Market Capitalisation / Total Liabilities

X_5 = Sales / Total Assets

Altman Z-Score



Z > 2.99 → Safe Zone

- The company is financially stable with a low probability of bankruptcy.
- Typically applies to well-established firms with solid financials.

1.81 ≤ Z ≤ 2.99 → Gray Zone

- The company is in a financial risk zone where distress is possible.
- Requires closer analysis of financial trends and industry conditions.

Z < 1.81 → Distress Zone

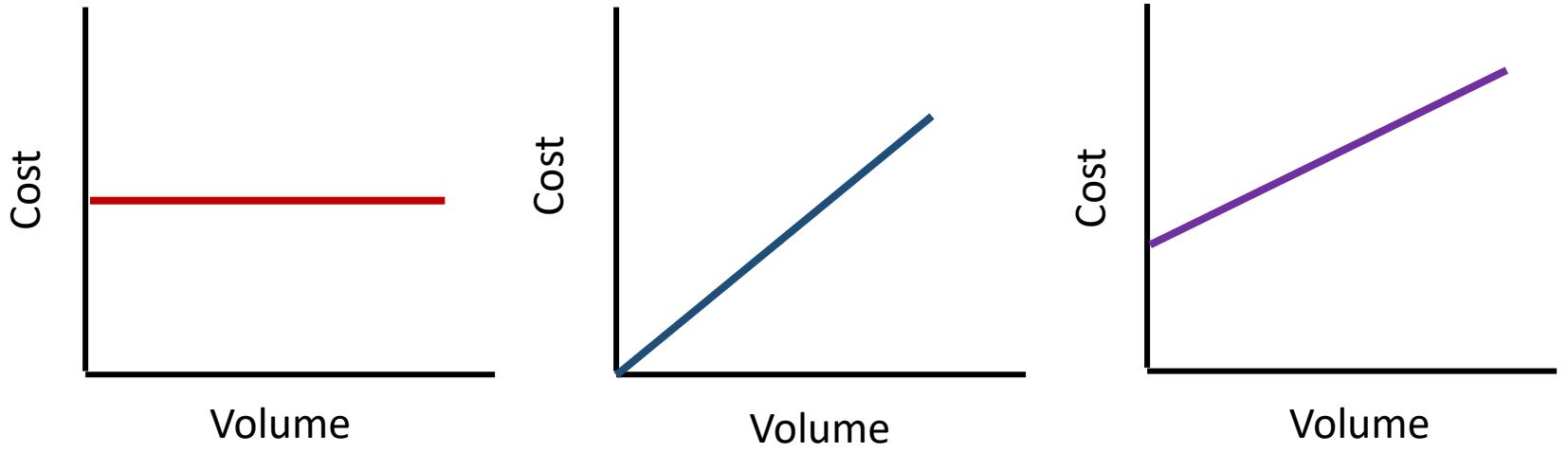
- High likelihood of financial distress or bankruptcy within two years.
- Often seen in struggling firms or industries facing downturns.

Cost Behavior

Fixed costs: do not vary with production/sales volume.

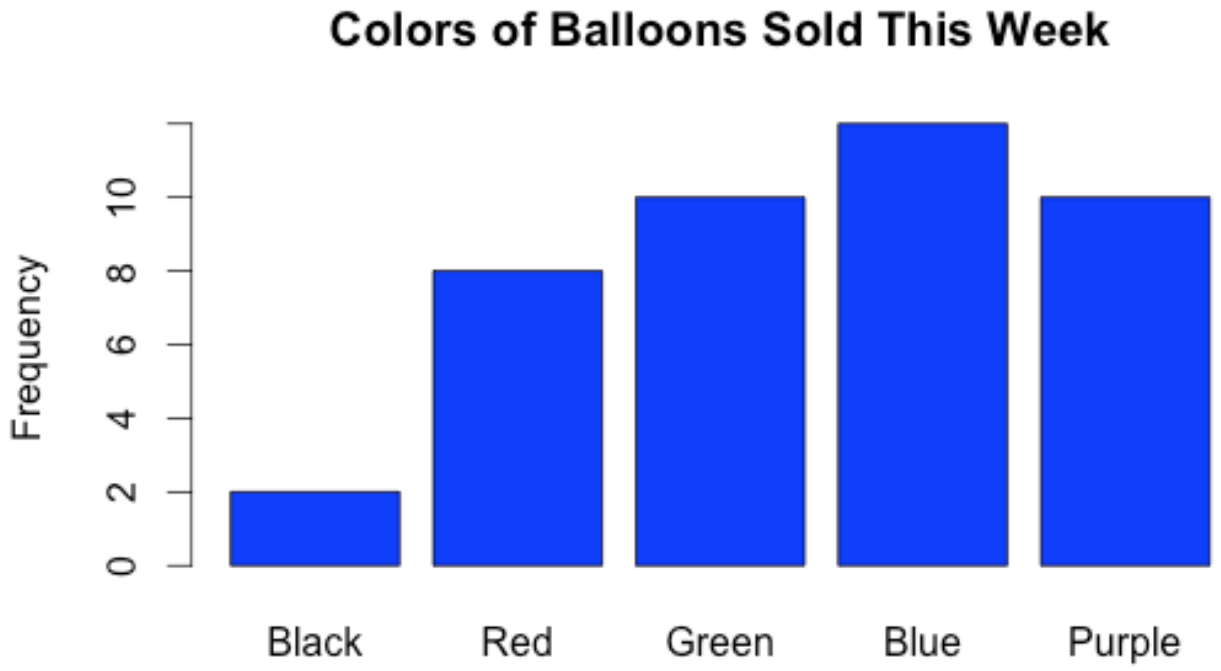
Variable costs: vary entirely based on production/sales volume.

Mixed costs: vary somewhat with production volume but also have a fixed cost component.



Categorical Data

Categorical data presents an analysis of different groups. Its purpose is to facilitate comparison across categories.

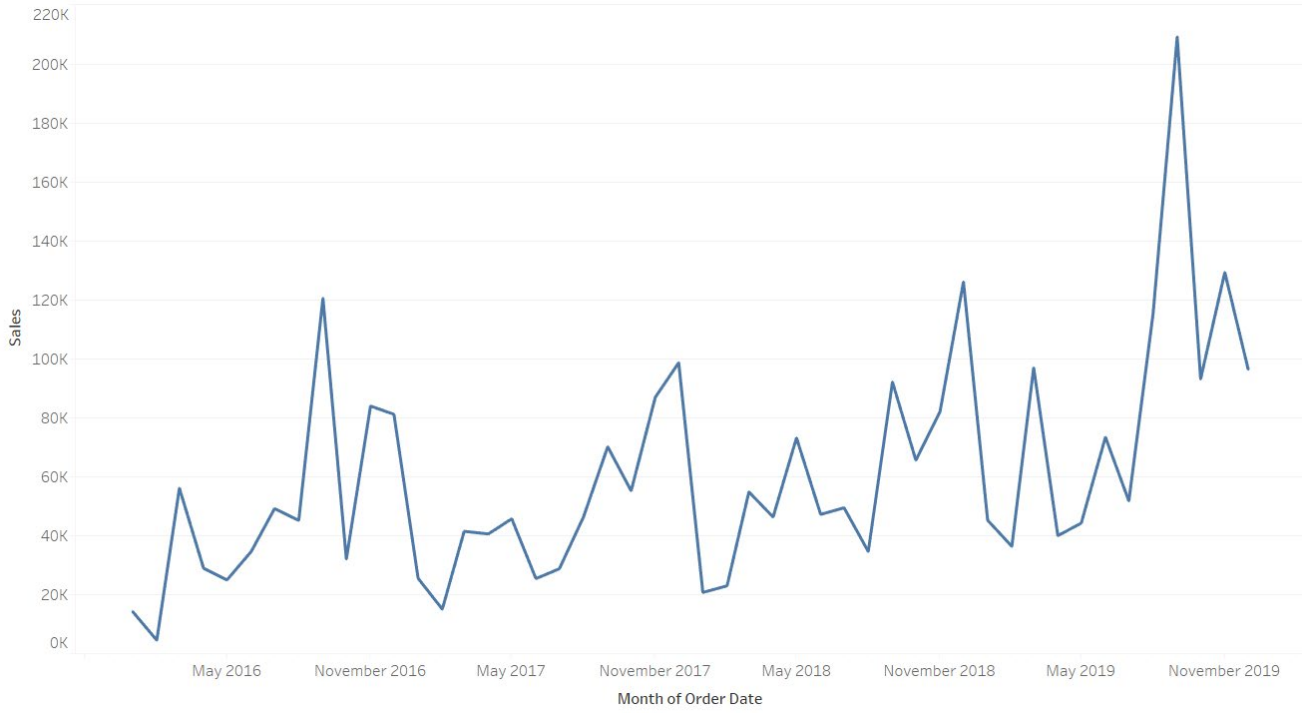


Bar chart

Color
Blue
Purple
Green
Red
Purple
Purple
Blue
Purple
Blue
Blue
Purple
Red
Black
Blue
...

Time Series Data

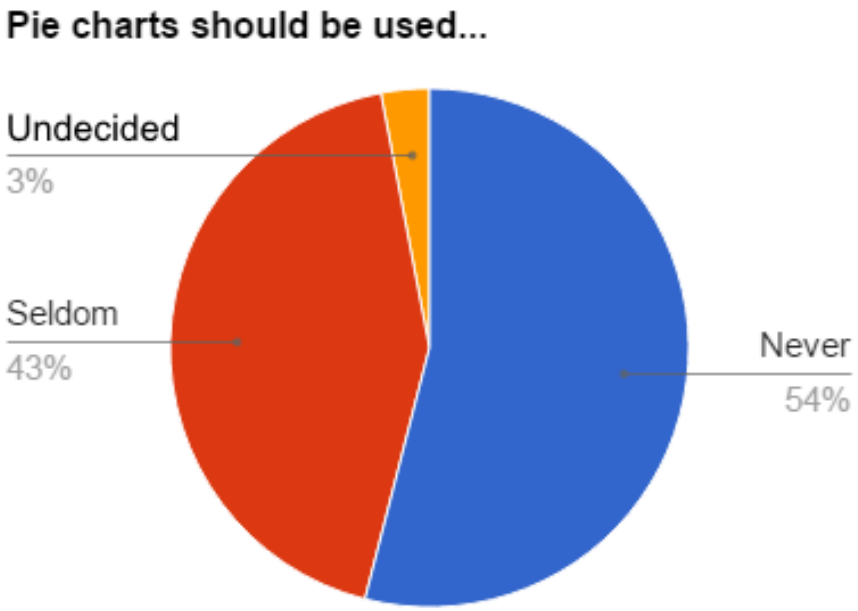
Data is considered a '**time series**' if it captures differences observations within a unit over time. Its purpose is to assess trends.



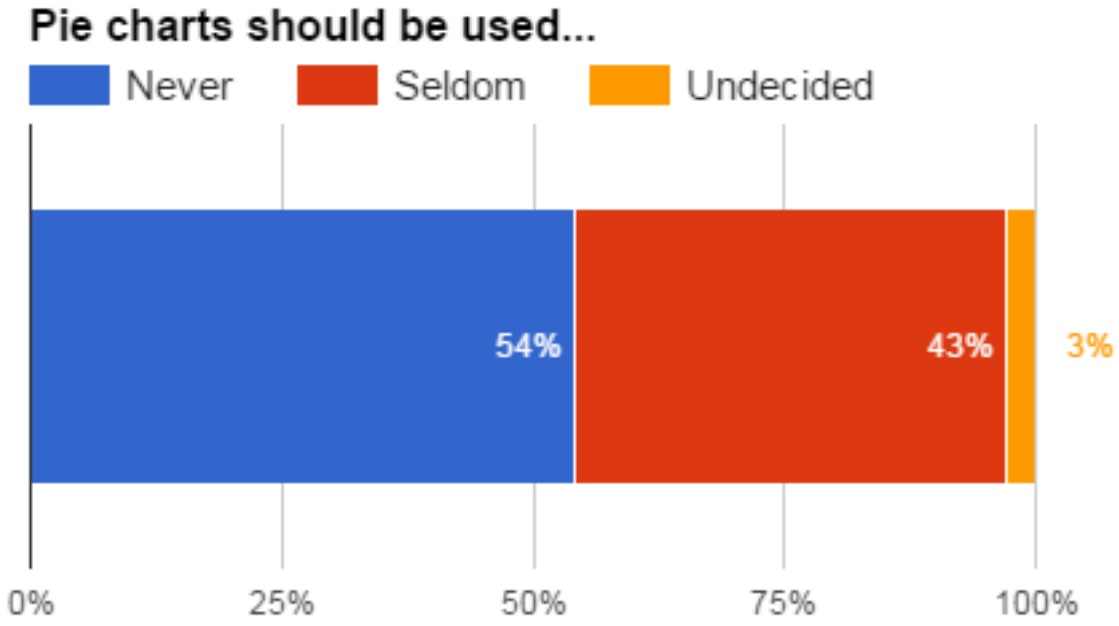
Line Charts

Proportional Data

Proportional data tracks response variables that express percentages or fractions. Its purpose is to evaluate parts to a whole.



Pie Charts



100% Stacked Bar Charts

3 Ingredients to Highly Effective Visuals

- Appropriate type of visual is used
- Visual elements are simplified
- Takeaways are emphasized

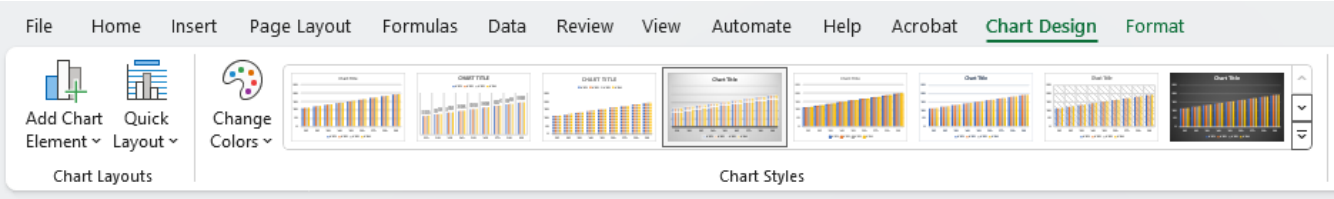
Simplification:

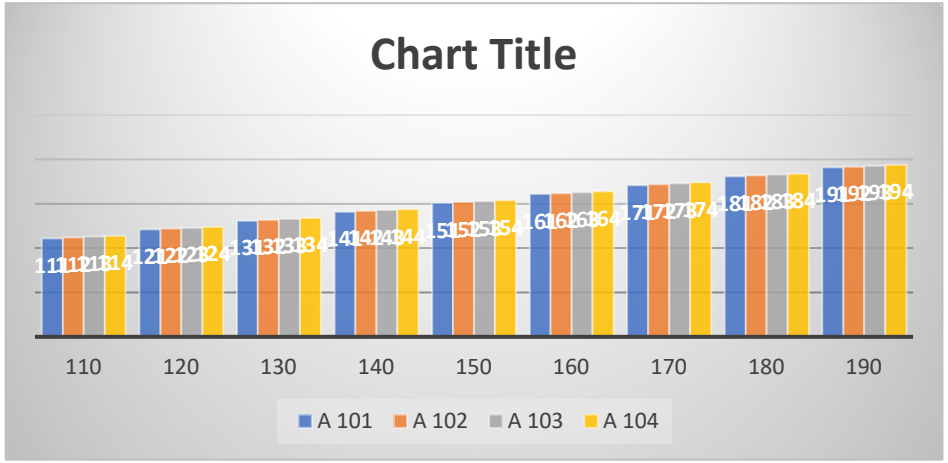
Making something easier to understand.

Data-to-Ink Ratio

$$\frac{Data}{Ink} = \frac{Information}{Chart\ Elements}$$

Strive to maximize the data-to-ink ratio. Convey the information/takeaway using only necessary chart elements. Less is more.





Emphasis:

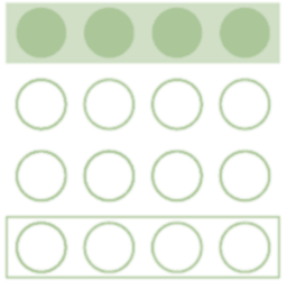
the art of **making important things stand-out**
and unimportant things disappear

Gestalt Principles (how do we see groups)

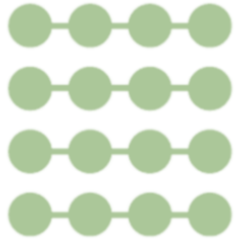
Proximity: when objects are close together.

Similarity: when objects exhibit similar attributes

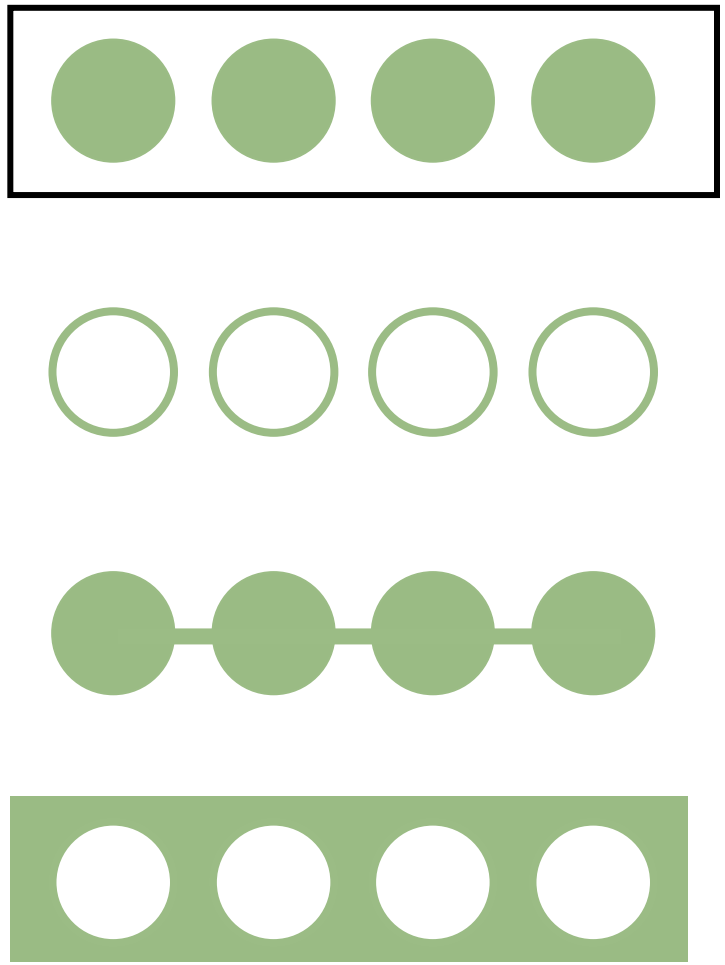
Enclosure: objects within a common boundary



Connection: objects appear attached



Gestalt Principles (how do we see groups)



Gestalt Principles (how do we see groups)

Symmetry: objects are symmetrical.

Continuity: objects are aligned

Closure: our preference to perceive a meaningful whole.

Gestalt Principles (how do we see groups)

Figure/Ground: the tendency to organize objects relative to a background.

Pre-Attentive Attributes

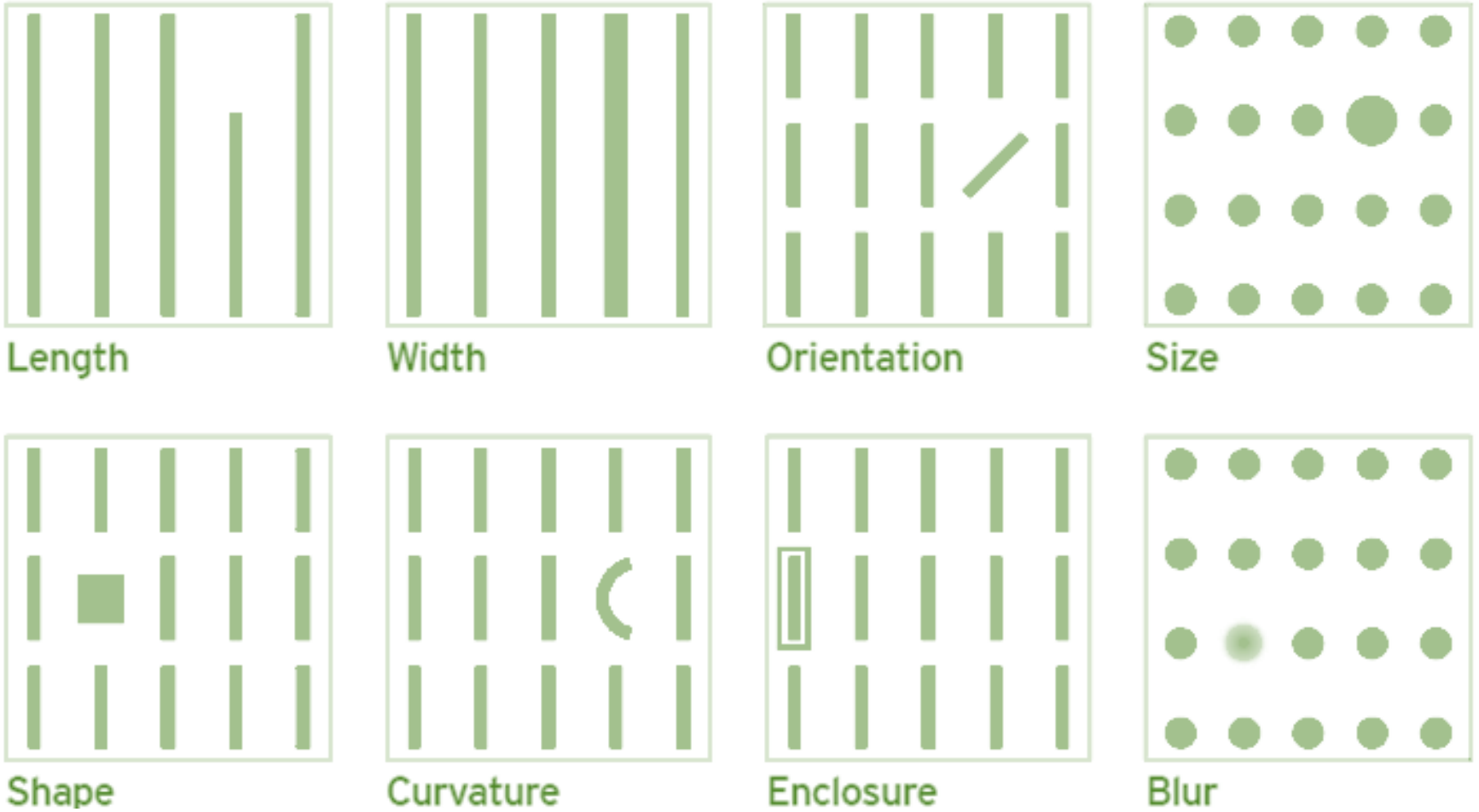
Pre-attentive processing refers to how our brains engage in subconscious perception based on a variety of characteristics, occurring at near instantaneous speed.

Two categories of these pre-attentive attributes:

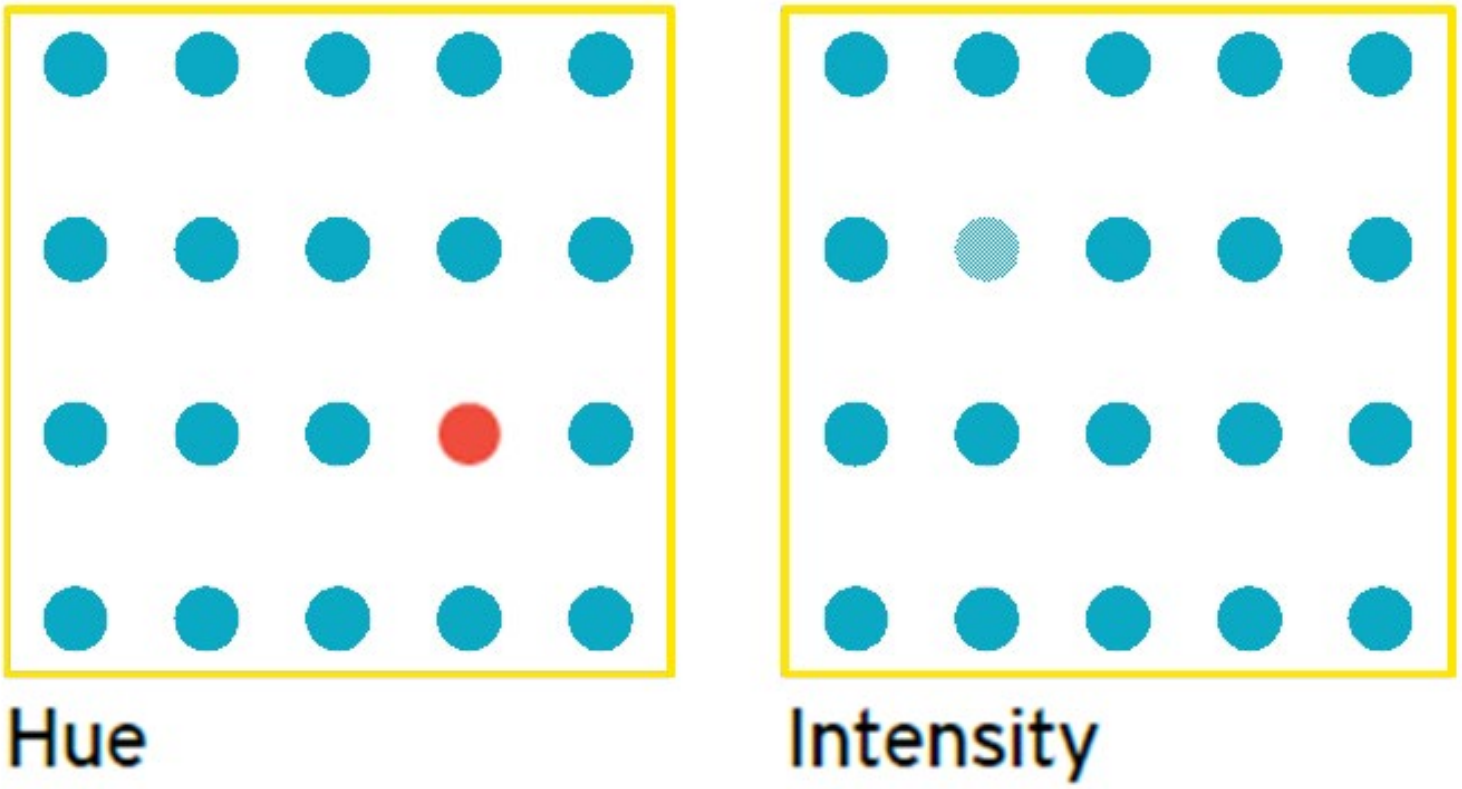
- Form
- Color

Form

The shape and dimensions of how objects are represented



Emphasis can be achieved through differences in **hue** (shades) and **intensity** (saturation)



Color

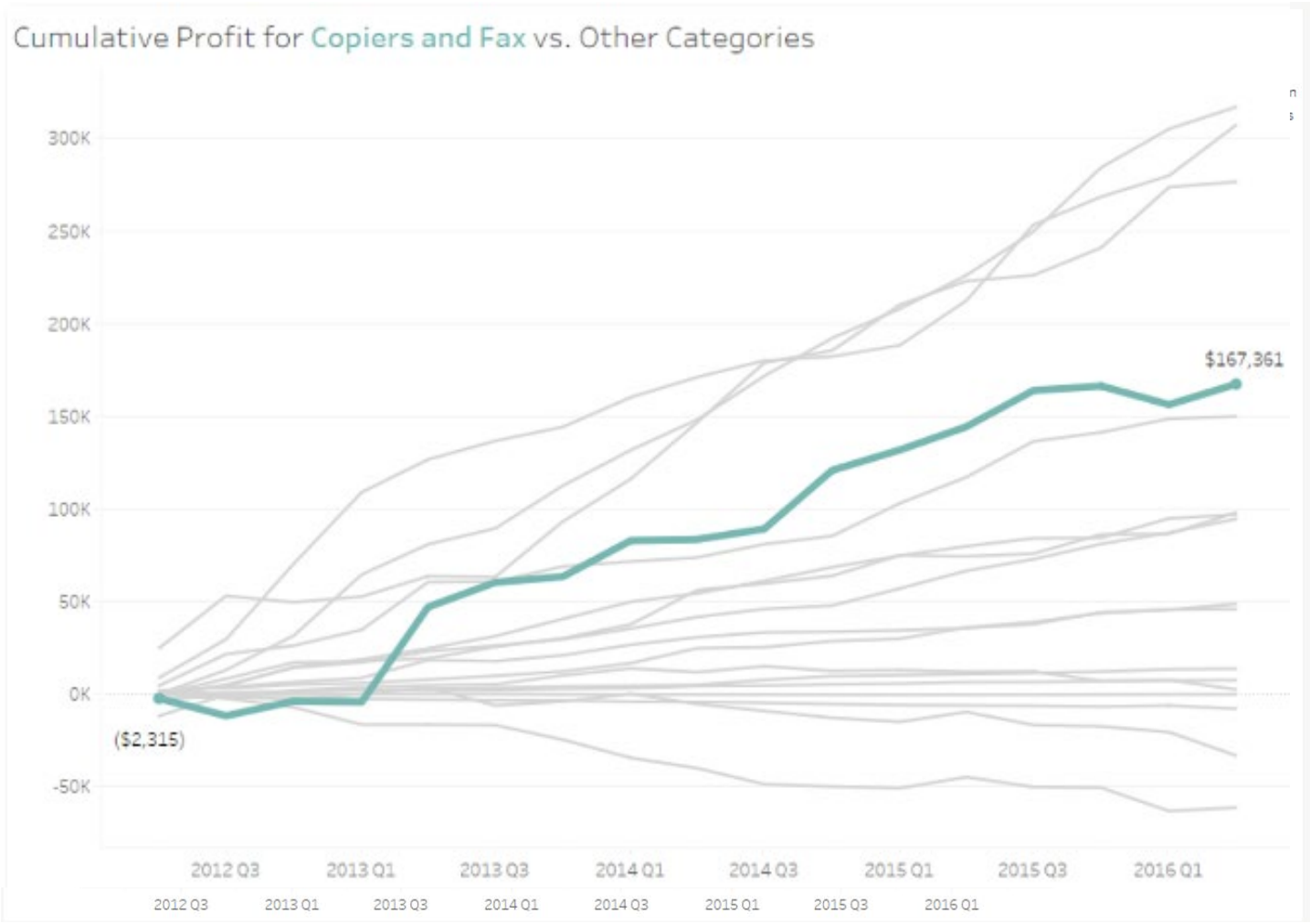
Line Charts

Avoid multiple y-axes

Limit the number of lines

Data labels can be helpful but used sparingly

Avoid awkward legends



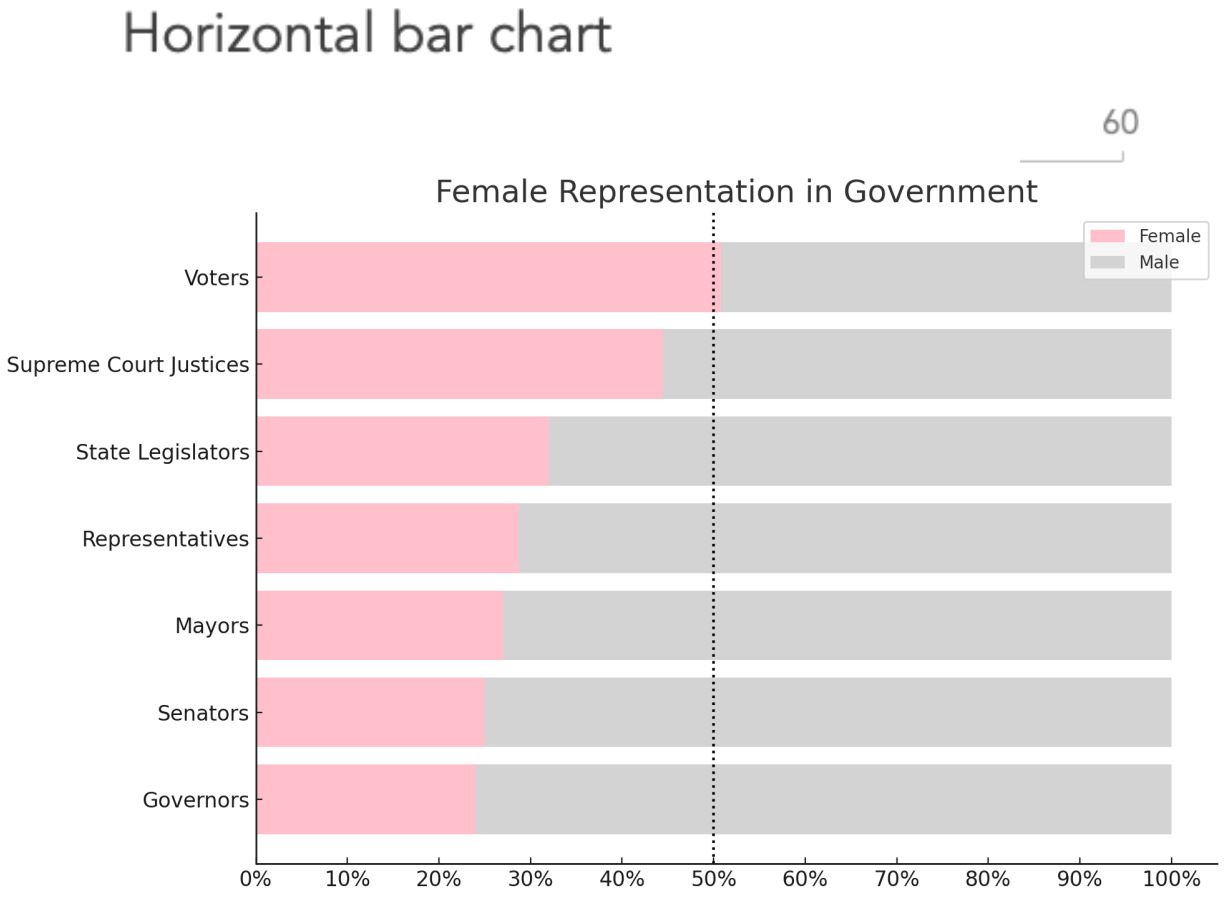
Bar Charts

Vertical: few categories, including limited time series (e.g., before and after).
Horizontal: use for longer category names or many categories.

Bar/space width: not too thick or thin.

If stacked, limit the number of groups.

Sort the data on at least one category if appropriate.

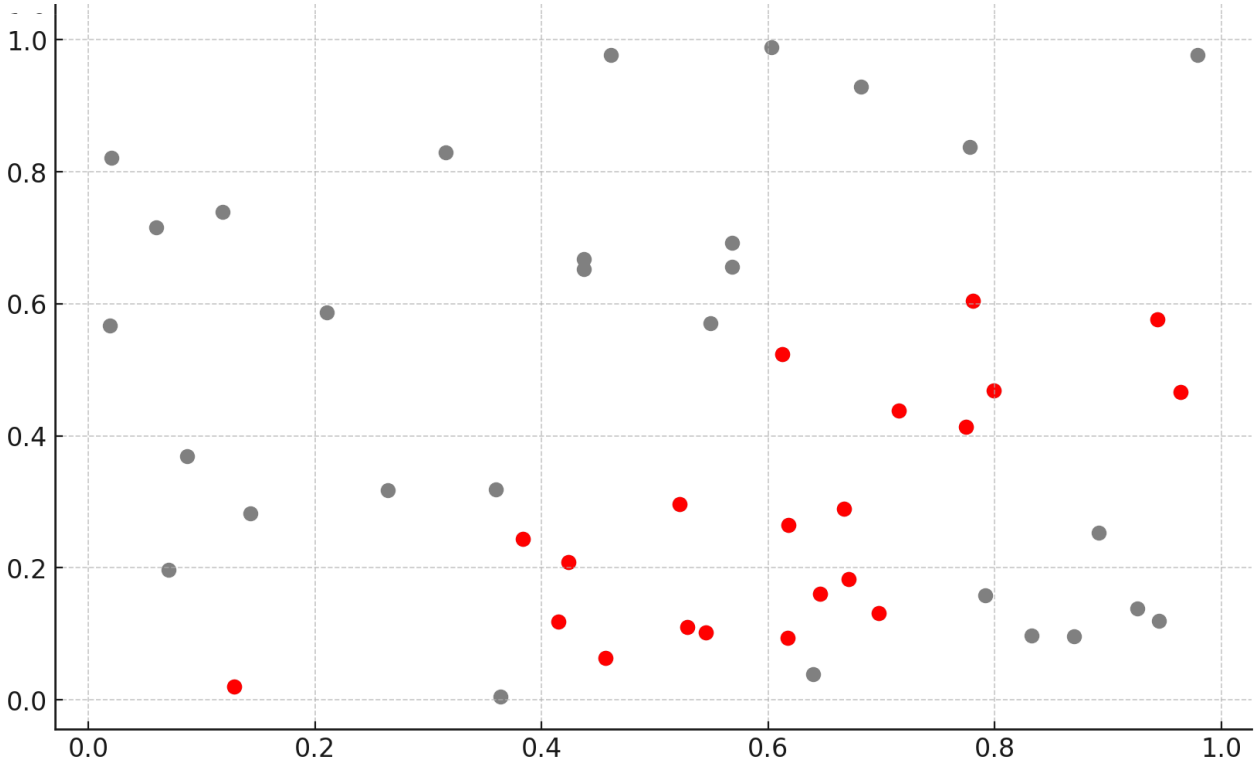


Color

Use color sparingly!

Understand contexts

Be consistent with color use across visuals.



More on Emphasis and Simplicity.

Legends can unnecessarily increase processing time.

Reduce redundancy. Axes titles are not always necessary.

Data labels, if not over-used, can directly provide quantitative info.

Use text to tell the story.

Gridlines should not dominate.

No 3D effects (or other “fancy” chart styles)