FOUNDATIONS OF

DATA LITERACY

BUILD CONFIDENCE AND MAKE **SMART**, **DATA-DRIVEN DECISIONS**

MAVEN® ANALYTICS

*Copyright Maven Analytics, LLC

COURSE OUTLINE

1	**Why Data Literacy?**	*Learn why data literacy skills are in high demand, and how they can be used to make smarter, more impactful decisions*
2	**Data Literacy 101**	*Learn the pillars of data literacy, address common misconceptions, and review the elements of a successful data ecosystem*
3	**Interpreting & Managing Data**	*Learn the basics of data profiling, practice identifying common quality issues, and review various data sources, storage modes and workflows*
4	**Analyzing Data**	*Learn proven frameworks for analytical thinking, and explore the differences between descriptive, diagnostic, predictive, and prescriptive analytics*
5	**Communicating with Data**	*Learn data visualization and storytelling best practices, review essential visuals and chart types, and discuss common visualization mistakes*
6	**Data Literacy + AI**	*Review common use cases for leveraging AI tools like ChatGPT, Gemini and Copilot, and learn prompt engineering tips along with pitfalls to avoid*
7	**Key Takeaways & Next Steps**	*Recap the most important takeaways from the course and suggest practical next steps in your data literacy journey*

*Copyright Maven Analytics, LLC

MEET YOUR INSTRUCTOR

CHRIS DUTTON

Chris is a best-selling instructor and Founder of Maven Analytics, with 15+ years specializing in data literacy & business intelligence.

At Maven Analytics, our mission is to empower everyday people with life-changing data skills. Named one of the top EdTech companies by USA Today, we've helped more than 1,000,000 learners around the world launch or accelerate their careers.

Where top companies train their teams

KPMG · Quest Diagnostics · SIEMENS · Pfizer · UKG · CATERPILLAR · crocs · Coca-Cola · CVSHealth

MAVEN ANALYTICS

SETTING EXPECTATIONS

This is a **high-level introduction** to the world of data literacy

- *This is a non-technical, tool-agnostic course, so we won't cover advanced or specialized topics*

Our goal is to provide a **foundation and roadmap** to help you build the confidence to make smart, data-driven decisions

- *We'll share proven frameworks and best practices for data management, analysis, storytelling and more*

This course is **designed for everyone**, not just data professionals

- *Data literacy can help you accelerate your career and make a bigger impact, no matter your role*

*Copyright Maven Analytics, LLC

WHY DATA LITERACY?

*Copyright Maven Analytics, LLC

DATA SKILLS ARE IN DEMAND

82%

of decision-makers **expect data literacy from employees in every department***

11%

of employees are **fully confident in their data literacy skills***

* https://www.tableau.com/sites/default/files/2022-03/Forrester_Building_Data_Literacy_Tableau_Mar2022.pdf
* https://www.qlik.com/us/-/media/files/resource-library/global-us/direct/datasheets/ds-data-literacy-the-upskilling-evolution-en.pdf

*Copyright Maven Analytics, LLC

DATA SKILLS ARE IN DEMAND

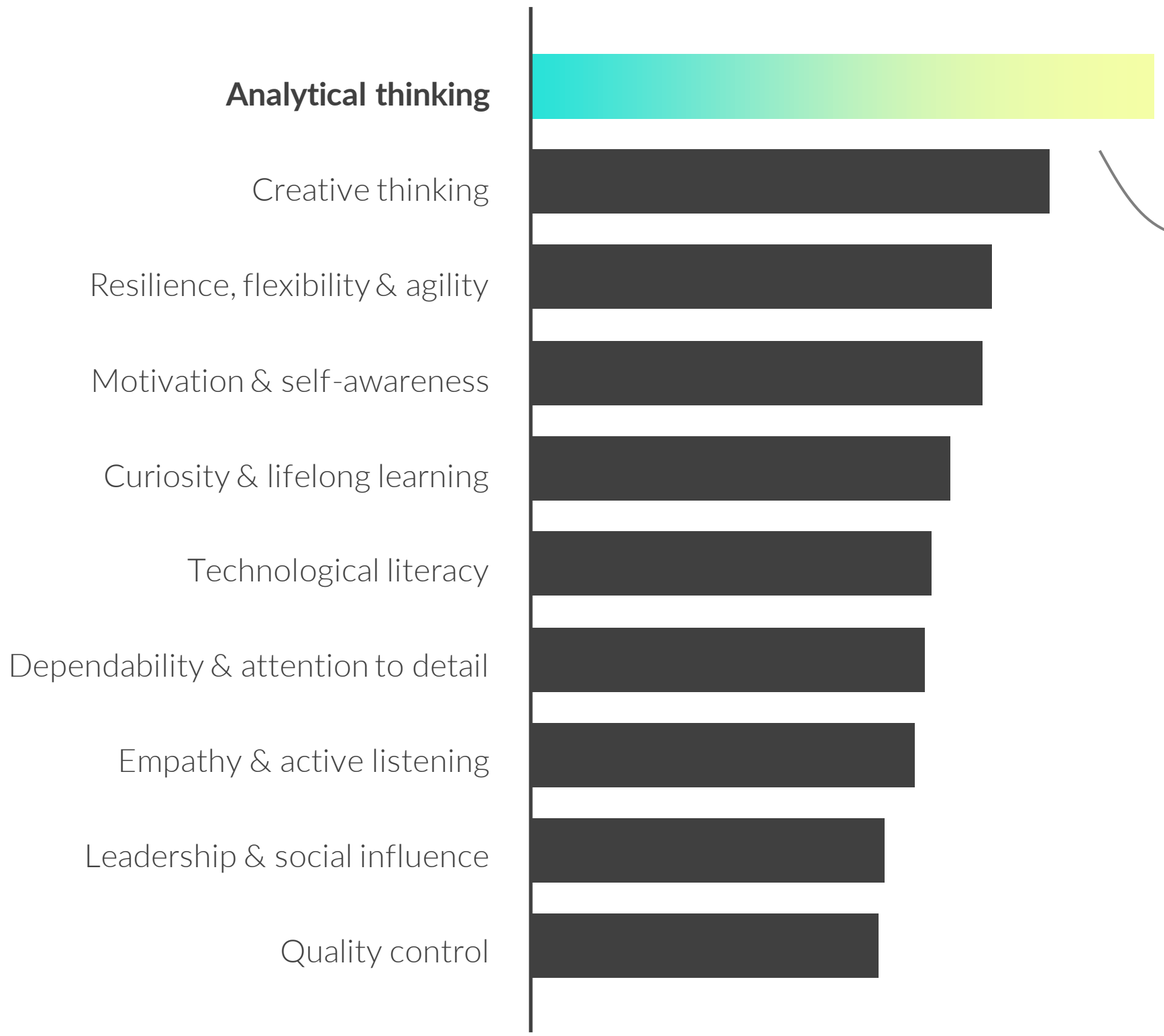
> **85% of executives** believe that being data literate will be **as vital in the future as the ability to use a computer is today***

In a Forrester study, "**basic data skills**" **was the most important factor for employee success** across all departments, but less than half of companies reported that they offer data skills training

* https://www.qlik.com/us/-/media/files/resource-library/global-us/direct/datasheets/ds-data-literacy-the-upskilling-evolution-en.pdf
* https://www.tableau.com/sites/default/files/2022-03/Forrester_Building_Data_Literacy_Tableau_Mar2022.pdf

*Copyright Maven Analytics, LLC

DATA SKILLS ARE IN DEMAND



According to the World Economic Forum's 2023 Future of Jobs report, **analytical thinking** is the **#1 core skill** required by companies today

Analytical thinking is also the **highest priority for skills training**, and is expected to account for ~10% of upskilling initiatives in 2023-2027, on average*

* https://www3.weforum.org/docs/WEF_Future_of_Jobs_2023.pdf

*Copyright Maven Analytics, LLC

THE WORLD RUNS ON DATA

$100B

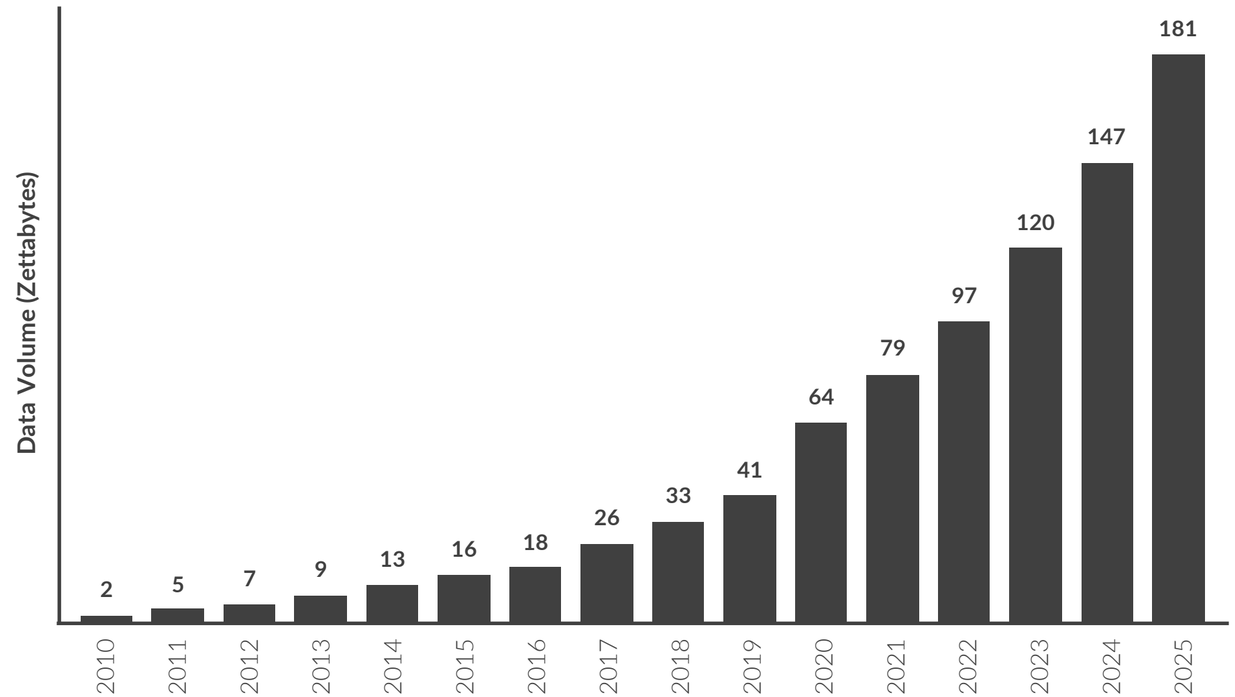
Size of the global **Data & Analytics** market in 2022*

90%

Of the world's data was produced in the **last 2 years***

2.5

Quintillion bytes of data produced **every day***



We live in a world that runs on data; every time you send an email, request an Uber, browse a website, type a Google search or watch a Netflix show, **you are generating data**

** https://www.globaldata.com/store/report/data-and-analytics-technology-market-analysis*
** https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read*
** https://www.statista.com/statistics/871513/worldwide-data-created/*

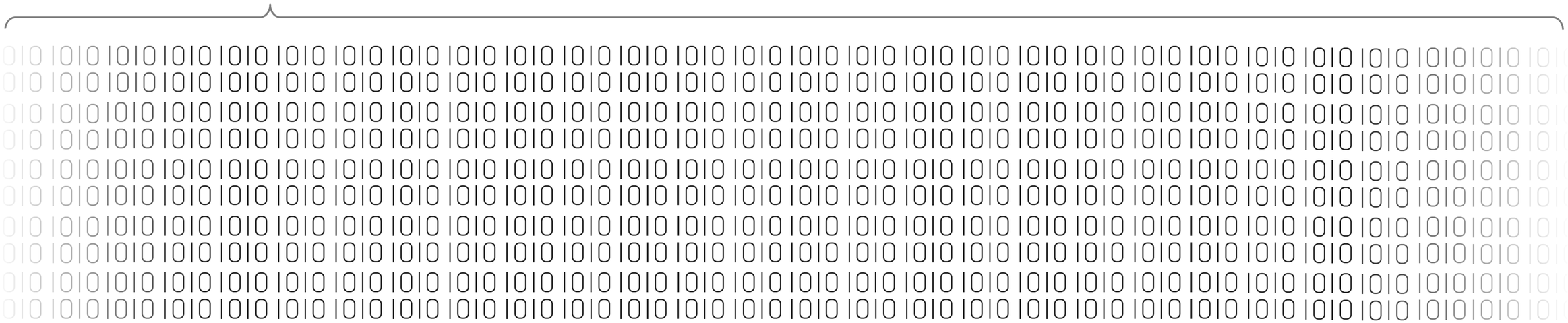
**Copyright Maven Analytics, LLC*

THE WORLD RUNS ON DATA

"Bigger" data isn't necessarily better; what matters most is that it can be used to derive **meaningful insights** and inform **data-driven decisions**

120 ZB

Global data creation in 2023 is estimated to be **~120 zettabytes**



* https://www.statista.com/statistics/871513/worldwide-data-created/

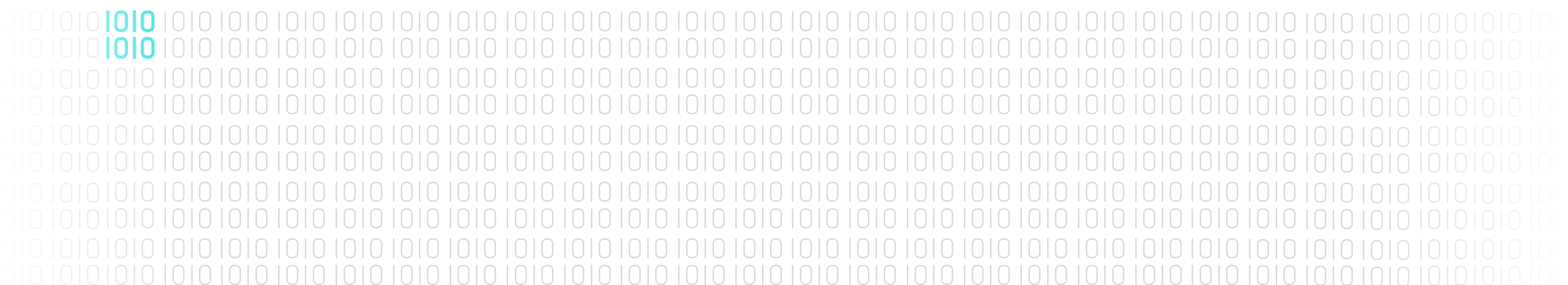
*Copyright Maven Analytics, LLC

THE WORLD RUNS ON DATA

"Bigger" data isn't necessarily better; what matters most is that it can be used to derive **meaningful insights** and inform **data-driven decisions**

<1%

But only a tiny fraction of the data in existence is actually **used for analysis**



* https://www.statista.com/statistics/871513/worldwide-data-created/

*Copyright Maven Analytics, LLC

DATA-DRIVEN DECISION MAKING

Data is one of the most valuable assets a business can own, but data literacy is what **brings it to life**

- Data literacy is about translating raw data into meaningful information, and using that information to inform **data-driven decisions**

The goal of any analysis is to **inspire action**. Charts and graphs may look pretty, but they are worthless until they inform a decision

*Copyright Maven Analytics, LLC

DATA-DRIVEN DECISION MAKING

VALUE

DATA

In its raw form, **data often has very little meaning or value**. Like oil, it must be mined, refined, and repackaged before it becomes useful for analysis

INFO

Once data is processed, organized and stored, it can be analyzed to produce **useful information** which adds context and clarity

INSIGHT

Data-driven insights are meaningful findings which can help inform or influence business decisions and recommendations

ACTION

Data is meaningless until it **inspires stakeholders to act**; no matter how insightful the analysis, it only adds value if it impacts real-world decisions

*Copyright Maven Analytics, LLC

DATA-DRIVEN DECISION MAKING

VALUE

DATA

*"There were **173** transactions in January"*

INFO

*"There were 173 transactions in January, **up 75% over December**. **Fitness equipment** and **athletic apparel** saw the largest month-over-month gains"*

INSIGHT

*"Every January we see an uptick in sales and revenue, **driven primarily by new customers looking to prioritize health and fitness in the new year**"*

ACTION

*"Based on this data, **we recommend increasing advertising budgets and testing campaigns to promote top-selling health and fitness products in January**"*

*Copyright Maven Analytics, LLC

DATA-DRIVEN DECISION MAKING

The **world's most successful companies leverage data** for everything from product design to pricing strategy, risk mitigation, financial modeling, process optimization and more

- Data is how **Netflix** and **Spotify** know which movies and songs to recommend, how **Amazon** knows what products you're likely to buy, how **sports teams** evaluate talent, how **banks** detect fraud, how **Starbucks** manages a global supply chain, and how **Uber** connects drivers with passengers in real-time

*Copyright Maven Analytics, LLC

DATA SKILLS ARE FOR EVERYONE

Data literacy is no longer a specialized skill, but a **core competency** that companies expect from employees across virtually all roles and departments

- **Everyone works with data to some degree**: marketers use it to optimize media mix, sales reps use it to target leads, HR departments use it to predict churn, product teams use it to create better user experiences, etc.

You can benefit from data literacy skills if you:

- ✓ Have ever worked with or encountered data as part of your job
- ✓ Want to become more fluent with data or use it more effectively
- ✓ Read or interpret data outputs (charts, tables, etc.) produced by others
- ✓ Are uncomfortable managing data or using tools like Excel
- ✓ Manage a team that works with data
- ✓ Want to build a powerful, highly versatile skill set

Data literacy skills are critical for both those who **conduct** analyses and those who **consume** them

*Copyright Maven Analytics, LLC

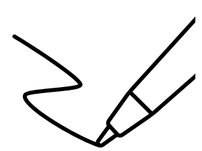
DATA SKILLS ARE FOR EVERYONE

Data literacy is no longer a specialized skill, but a **core competency** that companies expect from employees across virtually all roles and departments

- **Everyone works with data to some degree**: marketers use it to optimize media mix, sales reps use it to target leads, HR departments use it to predict churn, product teams use it to create better user experiences, etc.

What it takes to become data literate*:

BE **CURIOUS**



BE **CREATIVE**

THINK **CRITICALLY**

** Based on a framework originally introduced by Jordan Morrow*

**Copyright Maven Analytics, LLC*

DATA SKILLS ARE FOR EVERYONE

- ✓ A **teacher** uses data to track test scores, identify knowledge gaps, and revise lesson plans to help students better comprehend the material

- ✓ A **sales manager** uses data to monitor team performance, and creates pacing charts to keep reps motivated and on track to hit targets

- ✓ An **HR rep** uses data to identify unusually high employee churn, and analyzes exit surveys to identify work/life balance as a root cause

- ✓ A **marketer** for an online retailer uses data to understand which products customers bundle together and identify new cross-sell opportunities

*Copyright Maven Analytics, LLC

KEY BENEFITS OF DATA LITERACY

- Make **smarter**, **more impactful decisions**
- Work more efficiently and **supercharge your productivity**
- Make an impact and help **drive meaningful business outcomes**
- Accelerate your career and **earn a higher salary**
- Build a **future-proof**, **highly versatile skill set**

*Copyright Maven Analytics, LLC

DATA LITERACY 101

*Copyright Maven Analytics, LLC

DATA LITERACY DEFINITION

DATA LITERACY

The ability to **interpret**, **manage**, **analyze** and **communicate** with data

INTERPRET

Comprehend and make sense of data presented to you in the form of tables, charts, or metrics

MANAGE

Work with raw data to store, clean, transform and prepare it for exploration and analysis

ANALYZE

Use analytical tools and techniques to identify patterns and insights in the data

COMMUNICATE

Use visualization and storytelling to create clear, data-driven narratives to help influence decisions

COMMON MISCONCEPTIONS

Data literacy is about:

- Helping everyday people make smarter, data-driven decisions
- Democratizing access to data and learning resources
- Empowering all employees with the skills they need to thrive

Data literacy is NOT about:

- Forcing everyone to learn complex math or programming languages
- Replacing dedicated analytics and data science teams
- Transforming everyone into a data scientist or database engineer

Artificial intelligence tools like ChatGPT, Microsoft Copilot and Google Gemini are proving to be valuable tools for data-related tasks, **but human intuition and foundational data literacy skills are still critical**

*Copyright Maven Analytics, LLC

LEVELS OF DATA LITERACY

CONVERSATIONAL

You understand common data concepts and can interpret basic analytical outputs and ask smart questions, but don't have the ability to manipulate data or perform a technical analysis yourself

DATA INTERPRETERS

Non-analysts who are consuming and potentially making decisions based on analytical outputs

LITERATE

You can interpret, manage and analyze data for relatively simple descriptive or diagnostic analyses, but have limited experience conducting advanced analyses or presenting to stakeholders

CITIZEN ANALYSTS

People who aren't necessarily analytics specialists, but leverage data in their role

FLUENT

You have the skills to think strategically, design and conduct complex analyses, and clearly communicate insights and implications to both technical and non-technical audiences

DATA SPECIALISTS

Seasoned, specialized data professionals including BI Analysts, data scientists or data engineers

Data literacy is **not about making everyone fluent**; it's about empowering employees at all levels with the data skills they need to succeed in *their* specific role (conversational skills may be enough!)

** Loosely based on Gartner's 5 levels of data literacy: https://www.gartner.com/en/documents/3602517*

**Copyright Maven Analytics, LLC*

LEVELS OF DATA LITERACY

	INTERPRET	MANAGE	ANALYZE	COMMUNICATE
CONVERSATIONAL	Read and understand basic tables, charts and metrics	Access data in spreadsheets and perform basic actions like manipulating rows, columns and cells	Apply basic filtering and sorting techniques to explore tabular data	Participate in data conversations, share relevant facts and figures, and ask thoughtful questions
LITERATE				
FLUENT				

*Copyright Maven Analytics, LLC

LEVELS OF DATA LITERACY

	INTERPRET	MANAGE	ANALYZE	COMMUNICATE
CONVERSATIONAL	Read and understand basic tables, charts and metrics	Access data in spreadsheets and perform basic actions like manipulating rows, columns and cells	Apply basic filtering and sorting techniques to explore tabular data	Participate in data conversations, share relevant facts and figures, and ask thoughtful questions
LITERATE	Understand and draw conclusions from data presented in basic visuals and reports	Interact with multiple data sources, identify and correct common quality issues, and structure tables for analysis	Use tools like formulas, charts and pivot tables to identify patterns and trends	Use data and basic visuals to share findings, support decisions or communicate information and ideas
FLUENT				

*Copyright Maven Analytics, LLC

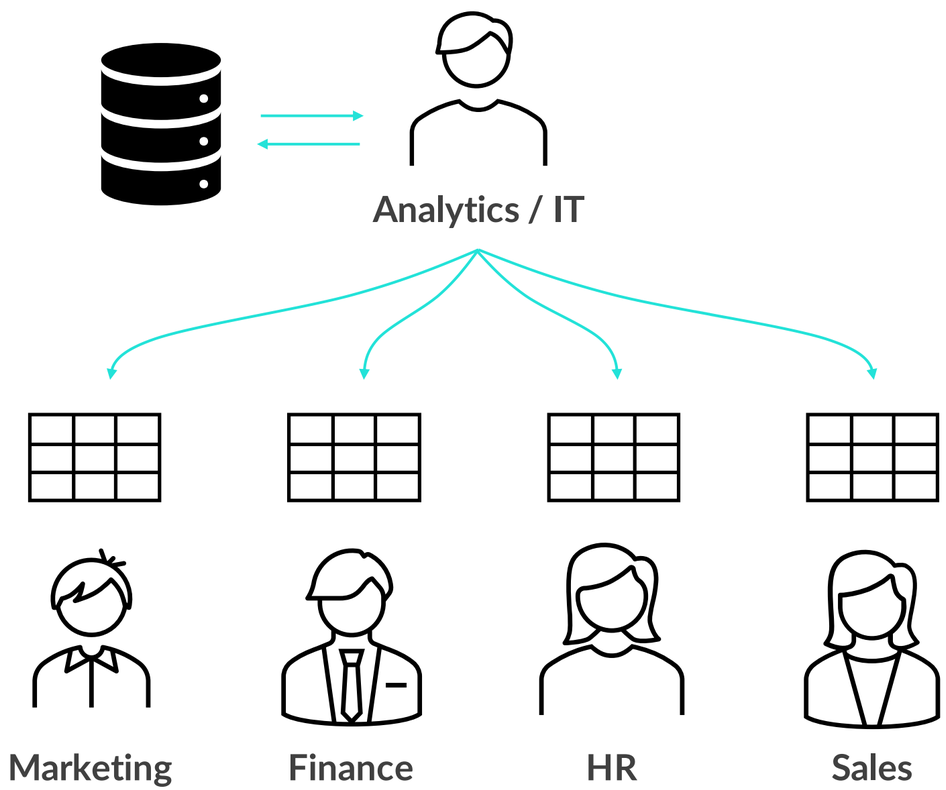
LEVELS OF DATA LITERACY

	INTERPRET	MANAGE	ANALYZE	COMMUNICATE
CONVERSATIONAL	Read and understand basic tables, charts and metrics	Access data in spreadsheets and perform basic actions like manipulating rows, columns and cells	Apply basic filtering and sorting techniques to explore tabular data	Participate in data conversations, share relevant facts and figures, and ask thoughtful questions
LITERATE	Understand and draw conclusions from data presented in basic visuals and reports	Interact with multiple data sources, identify and correct common quality issues, and structure tables for analysis	Use tools like formulas, charts and pivot tables to identify patterns and trends	Use data and basic visuals to share findings, support decisions or communicate information and ideas
FLUENT	Interpret data presented in virtually any format, including dashboards, complex charts or models	Design ETL workflows, build relational models, and engineer features for data analysis or machine learning	Leverage a wide range of tools, techniques, languages and models to analyze large or complex datasets	Use data visualization and storytelling techniques to communicate insights and inspire stakeholders to act

*Copyright Maven Analytics, LLC

DATA DEMOCRATIZATION

Data democratization describes the process of empowering employees throughout an organization with the access and resources they need to make data-driven decisions



In many organizations, analytics or IT teams act as "gatekeepers" who have direct access to data sources and provide data to other teams by request.

While this helps with security and governance, it limits access and **discourages data-driven decision making throughout the organization**

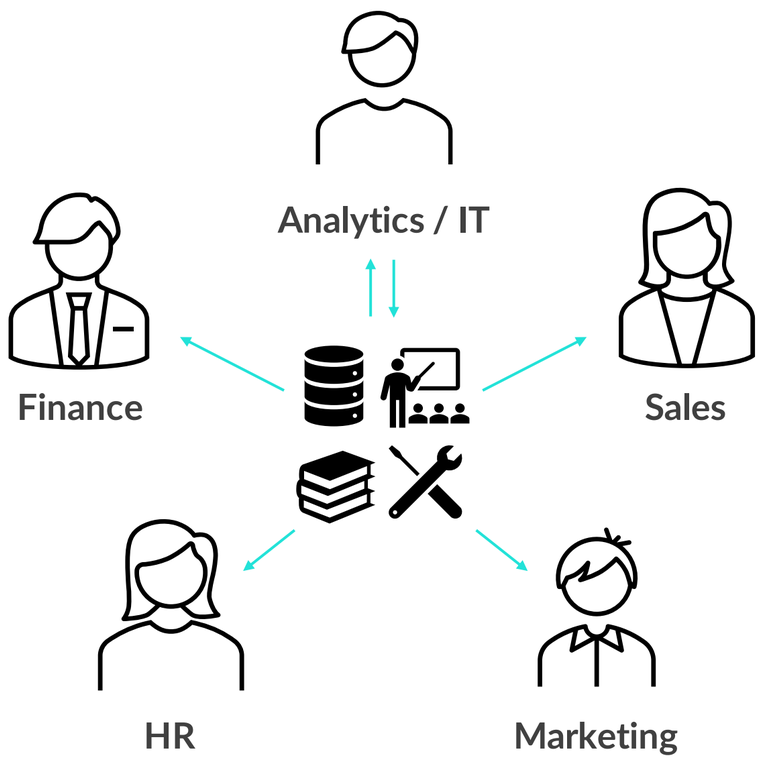
Common pain points:

- *I don't have access to the data I need*
- *I don't have the skills to make use of the data*
- *I don't have the right tools for the job*
- *I can't get the help I need when I need it*

*Copyright Maven Analytics, LLC

DATA DEMOCRATIZATION

Data democratization describes the process of empowering employees throughout an organization with the access and resources they need to make data-driven decisions



When you democratize access to data and provide data literacy skills training, you **empower employees to build confidence, work smarter, and make better decisions**.

Data democratization **isn't just about data access**, it's an ongoing process and a cultural shift, to enable people with the **skills**, **tools** and **resources** they need to utilize data responsibly and effectively

*Copyright Maven Analytics, LLC

DATA ECOSYSTEM

DATA **STRATEGY**

Overarching vision for how data will be used for the business

Data strategy is like the **blueprints for a building**, which ensure that everyone working on the project is following the same plan

DATA **ARCHITECTURE**

Technology used to collect, manage, process and connect data

Data architecture is like the **foundation and plumbing**, which creates a solid structure and allows data to flow from place to place

DATA **GOVERNANCE**

Rules and frameworks to make sure data is secure, accurate, and usable

Governance is like the **building codes** that ensure workers have the proper permits and follow the same universal quality standards

*Copyright Maven Analytics, LLC

KEY TAKEAWAYS

- Data literacy is about helping people make **smart, data-driven decisions**
- Data literacy is **NOT about turning everyone into data scientists**
- There are **multiple levels of data literacy**, from conversational to fluent
- **Data democratization** is a critical step to building a data literate workforce
- A strong data ecosystem requires **strategy**, **architecture** and **governance**

*Copyright Maven Analytics, LLC

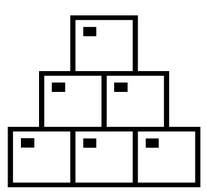
INTERPRETING & MANAGING DATA

*Copyright Maven Analytics, LLC

INTERPRETING & MANAGING DATA

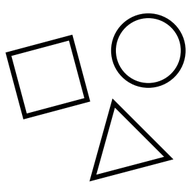
One of the most foundational data literacy skills is the ability to **interpret & manage** data

- **Interpreting** data means making sense of information presented as tables or charts, and knowing how to apply data profiling and QA techniques to understand the characteristics of a dataset
- **Managing** data means understanding how to access and work with different formats and storage modes, and how to extract, transform, load and prepare data for analysis
- As the **volume**, **variety** and **velocity** of data increases, these skills will be more important than ever



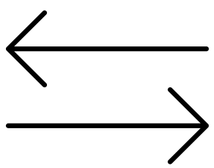
VOLUME

90% of the world's data was produced in the last 2 years



VARIETY

Data is captured in **more formats** than ever *(text, image, video, etc.)*



VELOCITY

Production and **processing speed** has never been faster

*Copyright Maven Analytics, LLC

INTERPRETING TABULAR DATA

EXAMPLE | COFFEE SHOP SALES

transaction_id	date	time	quantity	unit_price	product_id	product_name
1	1/1/24	7:06:11	2	$3.00	3	Gourmet brewed coffee
2	1/1/24	7:08:56	2	$3.10	9	Brewed Chai tea
3	1/1/24	7:20:24	1	$2.00	1	Drip coffee
4	1/1/24	7:22:41	2	$3.10	9	Brewed Chai tea
5	1/1/24	7:22:41	1	$3.00	12	Scone
6	1/1/24	7:25:49	1	$2.00	1	Drip coffee
7	1/1/24	7:33:34	2	$2.00	3	Gourmet brewed coffee
8	1/1/24	7:39:13	1	$4.25	5	Barista Espresso
9	1/1/24	7:43:05	1	$2.55	9	Brewed Chai tea
10	1/1/24	7:44:35	2	$3.50	3	Gourmet brewed coffee
11	1/1/24	7:45:51	1	$3.00	8	Brewed Black tea
12	1/1/24	7:48:19	1	$3.10	9	Brewed Chai tea
13	1/1/24	7:52:36	2	$3.00	5	Barista Espresso
14	1/1/24	7:59:58	2	$3.00	7	Brewed Green tea
15	1/1/24	7:59:58	1	$3.75	12	Scone
16	1/1/24	8:00:18	1	$2.50	6	Brewed herbal tea
17	1/1/24	8:17:27	2	$3.50	3	Gourmet brewed coffee
18	1/1/24	8:24:26	2	$2.55	9	Brewed Chai tea
19	1/1/24	8:24:26	1	$3.25	11	Biscotti
20	1/1/24	8:29:38	1	$2.55	9	Brewed Chai tea
21	1/1/24	8:31:23	1	$3.75	5	Barista Espresso
22	1/1/24	8:33:08	1	$3.00	6	Brewed herbal tea
23	1/1/24	8:33:08	1	$3.50	11	Biscotti
24	1/1/24	8:35:03	2	$3.00	6	Brewed herbal tea
25	1/1/24	8:35:03	1	$3.75	12	Pastry

This data is in a **tabular format**, which is a set of related data points stored in a table

*Copyright Maven Analytics, LLC

INTERPRETING TABULAR DATA

EXAMPLE | COFFEE SHOP SALES

transaction_id	date	time	quantity	unit_price	product_id	product_name
1	1/1/24	7:06:11	2	$3.00	3	Gourmet brewed coffee
2	1/1/24	7:08:56	2	$3.10	9	Brewed Chai tea
3	1/1/24	7:20:24	1	$2.00	1	Drip coffee
4	1/1/24	7:22:41	2	$3.10	9	Brewed Chai tea
5	1/1/24	7:22:41	1	$3.00	12	Scone
6	1/1/24	7:25:49	1	$2.00	1	Drip coffee
7	1/1/24	7:33:34	2	$2.00	3	Gourmet brewed coffee
8	1/1/24	7:39:13	1	$4.25	5	Barista Espresso
9	1/1/24	7:43:05	1	$2.55	9	Brewed Chai tea
10	1/1/24	7:44:35	2	$3.50	3	Gourmet brewed coffee
11	1/1/24	7:45:51	1	$3.00	8	Brewed Black tea
12	1/1/24	7:48:19	1	$3.10	9	Brewed Chai tea
13	1/1/24	7:52:36	2	$3.00	5	Barista Espresso
14	1/1/24	7:59:58	2	$3.00	7	Brewed Green tea
15	1/1/24	7:59:58	1	$3.75	12	Scone
16	1/1/24	8:00:18	1	$2.50	6	Brewed herbal tea
17	1/1/24	8:17:27	2	$3.50	3	Gourmet brewed coffee
18	1/1/24	8:24:26	2	$2.55	9	Brewed Chai tea
19	1/1/24	8:24:26	1	$3.25	11	Biscotti
20	1/1/24	8:29:38	1	$2.55	9	Brewed Chai tea
21	1/1/24	8:31:23	1	$3.75	5	Barista Espresso
22	1/1/24	8:33:08	1	$3.00	6	Brewed herbal tea
23	1/1/24	8:33:08	1	$3.50	11	Biscotti
24	1/1/24	8:35:03	2	$3.00	6	Brewed herbal tea
25	1/1/24	8:35:03	1	$3.75	13	Pastry

This data is in a **tabular format**, which is a set of related data points stored in a table

The **header row** gives us information about the data our table contains (in this case coffee shop transactions)

*Copyright Maven Analytics, LLC

INTERPRETING TABULAR DATA

EXAMPLE | COFFEE SHOP SALES

transaction_id	date	time	quantity	unit_price	product_id	product_name
1	1/1/24	7:06:11	2	$3.00	3	Gourmet brewed coffee
2	1/1/24	7:08:56	2	$3.10	9	Brewed Chai tea
3	1/1/24	7:20:24	1	$2.00	1	Drip coffee
4	1/1/24	7:22:41	2	$3.10	9	Brewed Chai tea
5	1/1/24	7:22:41	1	$3.00	12	Scone
6	1/1/24	7:25:49	1	$2.00	1	Drip coffee
7	1/1/24	7:33:34	2	$2.00	3	Gourmet brewed coffee
8	1/1/24	7:39:13	1	$4.25	5	Barista Espresso
9	1/1/24	7:43:05	1	$2.55	9	Brewed Chai tea
10	1/1/24	7:44:35	2	$3.50	3	Gourmet brewed coffee
11	1/1/24	7:45:51	1	$3.00	8	Brewed Black tea
12	1/1/24	7:48:19	1	$3.10	9	Brewed Chai tea
13	1/1/24	7:52:36	2	$3.00	5	Barista Espresso
14	1/1/24	7:59:58	2	$3.00	7	Brewed Green tea
15	1/1/24	7:59:58	1	$3.75	12	Scone
16	1/1/24	8:00:18	1	$2.50	6	Brewed herbal tea
17	1/1/24	8:17:27	2	$3.50	3	Gourmet brewed coffee
18	1/1/24	8:24:26	2	$2.55	9	Brewed Chai tea
19	1/1/24	8:24:26	1	$3.25	11	Biscotti
20	1/1/24	8:29:38	1	$2.55	9	Brewed Chai tea
21	1/1/24	8:31:23	1	$3.75	5	Barista Espresso
22	1/1/24	8:33:08	1	$3.00	6	Brewed herbal tea
23	1/1/24	8:33:08	1	$3.50	11	Biscotti
24	1/1/24	8:35:03	2	$3.00	6	Brewed herbal tea
25	1/1/24	8:35:03	1	$3.75	13	Pastry

This data is in a **tabular format**, which is a set of related data points stored in a table

The **header row** gives us information about the data our table contains (in this case coffee shop transactions)

Tables consist of **rows** which represent individual records, **columns** (or **fields**) which contain information about each record, and **cells** containing individual data points

*Copyright Maven Analytics, LLC

INTERPRETING TABULAR DATA

EXAMPLE | COFFEE SHOP SALES

transaction_id	date	time	quantity	unit_price	product_id	product_name
1	1/1/24	7:06:11	2	$3.00	3	Gourmet brewed coffee
2	1/1/24	7:08:56	2	$3.10	9	Brewed Chai tea
3	1/1/24	7:20:24	1	$2.00	1	Drip coffee
4	1/1/24	7:22:41	2	$3.10	9	Brewed Chai tea
5	1/1/24	7:22:41	1	$3.00	12	Scone
6	1/1/24	7:25:49	1	$2.00	1	Drip coffee
7	1/1/24	7:33:34	2	$2.00	3	Gourmet brewed coffee
8	1/1/24	7:39:13	1	$4.25	5	Barista Espresso
9	1/1/24	7:43:05	1	$2.55	9	Brewed Chai tea
10	1/1/24	7:44:35	2	$3.50	3	Gourmet brewed coffee
11	1/1/24	7:45:51	1	$3.00	8	Brewed Black tea
12	1/1/24	7:48:19	1	$3.10	9	Brewed Chai tea
13	1/1/24	7:52:36	2	$3.00	5	Barista Espresso
14	1/1/24	7:59:58	2	$3.00	7	Brewed Green tea
15	1/1/24	7:59:58	1	$3.75	12	Scone
16	1/1/24	8:00:18	1	$2.50	6	Brewed herbal tea
17	1/1/24	8:17:27	2	$3.50	3	Gourmet brewed coffee
18	1/1/24	8:24:26	2	$2.55	9	Brewed Chai tea
19	1/1/24	8:24:26	1	$3.25	11	Biscotti
20	1/1/24	8:29:38	1	$2.55	9	Brewed Chai tea
21	1/1/24	8:31:23	1	$3.75	5	Barista Espresso
22	1/1/24	8:33:08	1	$3.00	6	Brewed herbal tea
23	1/1/24	8:33:08	1	$3.50	11	Biscotti
24	1/1/24	8:35:03	2	$3.00	6	Brewed herbal tea
25	1/1/24	8:35:03	1	$3.75	12	Pastry

This data is in a **tabular format**, which is a set of related data points stored in a table

The **header row** gives us information about the data our table contains (in this case coffee shop transactions)

Tables consist of **rows** which represent individual records, **columns** (or **fields**) which contain information about each record, and **cells** containing individual data points

Columns can contain **numerical values** (or **metrics**) which can be counted or aggregated, or **descriptive attributes** which can be used to filter or explore the data

*Copyright Maven Analytics, LLC

INTERPRETING TABULAR DATA

EXAMPLE | COFFEE SHOP SALES

transaction_id	date	time	quantity	unit_price	product_id	product_name
1	1/1/24	7:06:11	2	$3.00	3	Gourmet brewed coffee
2	1/1/24	7:08:56	2	$3.10	9	Brewed Chai tea
3	1/1/24	7:20:24	1	$2.00	1	Drip coffee
4	1/1/24	7:22:41	2	$3.10	9	Brewed Chai tea
5	1/1/24	7:22:41	1	$3.00	12	Scone
6	1/1/24	7:25:49	1	$2.00	1	Drip coffee
7	1/1/24	7:33:34	2	$2.00	3	Gourmet brewed coffee
8	1/1/24	7:39:13	1	$4.25	5	Barista Espresso
9	1/1/24	7:43:05	1	$2.55	9	Brewed Chai tea
10	1/1/24	7:44:35	2	$3.50	3	Gourmet brewed coffee
11	1/1/24	7:45:51	1	$3.00	8	Brewed Black tea
12	1/1/24	7:48:19	1	$3.10	9	Brewed Chai tea
13	1/1/24	7:52:36	2	$3.00	5	Barista Espresso
14	1/1/24	7:59:58	2	$3.00	7	Brewed Green tea
15	1/1/24	7:59:58	1	$3.75	12	Scone
16	1/1/24	8:00:18	1	$2.50	6	Brewed herbal tea
17	1/1/24	8:17:27	2	$3.50	3	Gourmet brewed coffee
18	1/1/24	8:24:26	2	$2.55	9	Brewed Chai tea
19	1/1/24	8:24:26	1	$3.25	11	Biscotti
20	1/1/24	8:29:38	1	$2.55	9	Brewed Chai tea
21	1/1/24	8:31:23	1	$3.75	5	Barista Espresso
22	1/1/24	8:33:08	1	$3.00	6	Brewed herbal tea
23	1/1/24	8:33:08	1	$3.50	11	Biscotti
24	1/1/24	8:35:03	2	$3.00	6	Brewed herbal tea
25	1/1/24	8:35:03	1	$3.75	12	Pastry

This data is in a **tabular format**, which is a set of related data points stored in a table

The **header row** gives us information about the data our table contains (in this case coffee shop transactions)

Tables consist of **rows** which represent individual records, **columns** (or **fields**) which contain information about each record, and **cells** containing individual data points

Columns can contain **numerical values** (or **metrics**) which can be counted or aggregated, or **descriptive attributes** which can be used to filter or explore the data

Some columns are known as **keys**, which can be used to create connections or relationships between tables

*Copyright Maven Analytics, LLC

INTERPRETING TABULAR DATA

EXAMPLE | COFFEE SHOP SALES

transaction_id	date	time	quantity	unit_price	product_id	product_name
1	1/1/24	7:06:11	2	$3.00	3	Gourmet brewed coffee
2	1/1/24	7:08:56	2	$3.10	9	Brewed Chai tea
3	1/1/24	7:20:24	1	$2.00	1	Drip coffee
4	1/1/24	7:22:41	2	$3.10	9	Brewed Chai tea
5	1/1/24	7:22:41	1	$3.00	12	Scone
6	1/1/24	7:25:49	1	$2.00	1	Drip coffee
7	1/1/24	7:33:34	2	$2.00	3	Gourmet brewed coffee
8	1/1/24	7:39:13	1	$4.25	5	Barista Espresso
9	1/1/24	7:43:05	1	$2.55	9	Brewed Chai tea
10	1/1/24	7:44:35	2	$3.50	3	Gourmet brewed coffee
11	1/1/24	7:45:51	1	$3.00	8	Brewed Black tea
12	1/1/24	7:48:19	1	$3.10	9	Brewed Chai tea
13	1/1/24	7:52:36	2	$3.00	5	Barista Espresso
14	1/1/24	7:59:58	2	$3.00	7	Brewed Green tea
15	1/1/24	7:59:58	1	$3.75	12	Scone
16	1/1/24	8:00:18	1	$2.50	6	Brewed herbal tea
17	1/1/24	8:17:27	2	$3.50	3	Gourmet brewed coffee
18	1/1/24	8:24:26	2	$2.55	9	Brewed Chai tea
19	1/1/24	8:24:26	1	$3.25	11	Biscotti
20	1/1/24	8:29:38	1	$2.55	9	Brewed Chai tea
21	1/1/24	8:31:23	1	$3.75	5	Barista Espresso
22	1/1/24	8:33:08	1	$3.00	6	Brewed herbal tea
23	1/1/24	8:33:08	1	$3.50	11	Biscotti
24	1/1/24	8:35:03	2	$3.00	6	Brewed herbal tea
25	1/1/24	8:35:03	1	$3.75	13	Pastry

product_id	product_name	product_category	SKU	unit_price	calories
1	Drip coffee	Coffee	18624	$2.00	30
2	Organic brewed coffee	Coffee	63065	$3.00	30
3	Gourmet brewed coffee	Coffee	28292	$3.50	35
4	Premium brewed coffee	Coffee	46008	$4.00	40
5	Barista Espresso	Coffee	67307	$4.25	10
6	Brewed herbal tea	Tea	88117	$3.00	80
7	Brewed Green tea	Tea	70382	$3.00	15
8	Brewed Black tea	Tea	82122	$3.00	10
9	Brewed Chai tea	Tea	68396	$3.10	50
10	Hot chocolate	Drinking Chocolate	67637	$2.50	125
11	Biscotti	Bakery	42183	$3.25	145
12	Scone	Bakery	95376	$3.75	75
13	Pastry	Bakery	77454	$3.75	150

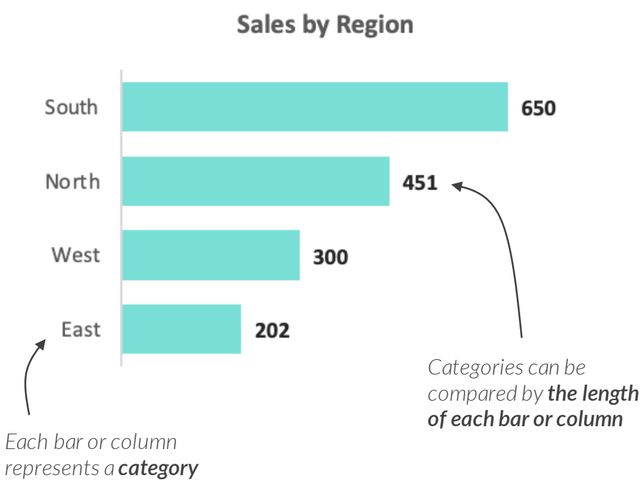
Some columns are known as **keys**, which can be used to create connections or relationships between tables

*Copyright Maven Analytics, LLC

INTERPRETING COMMON CHARTS

BAR/COLUMN CHART

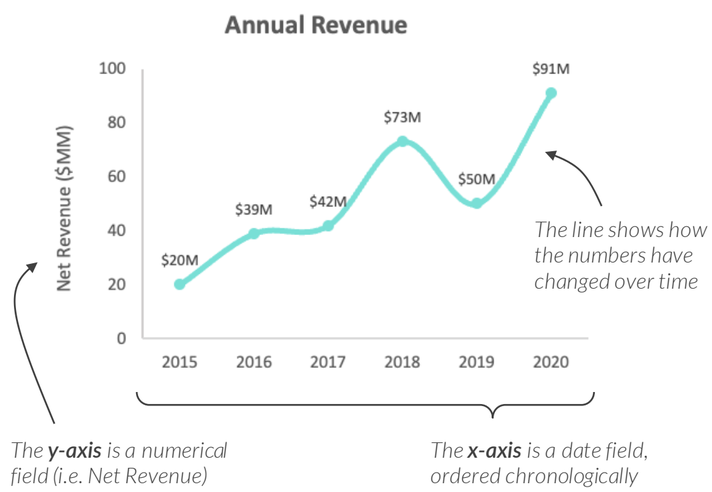
Used to show **comparisons between categories**



*In this example, the **South** region drove the most sales (650), and the **East** region drove the fewest*

LINE CHART

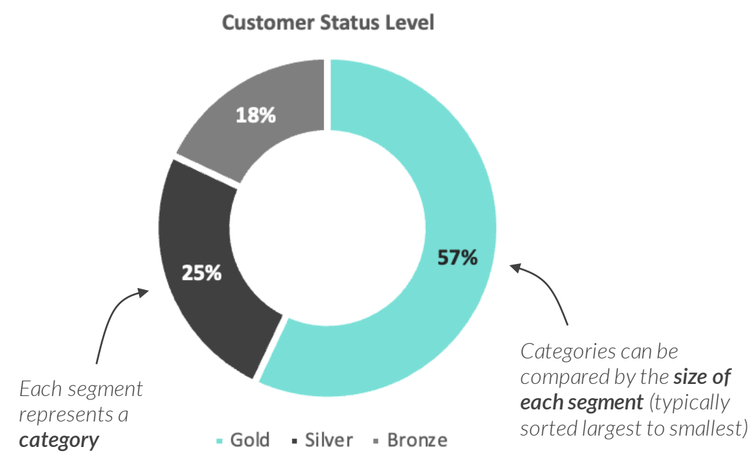
Used to show **trends over time**



*In this example, net revenue climbed to **$73M** in 2018, dipped in 2019, and peaked at **$91M** in 2020*

PIE/DONUT CHART

Used to show composition as **parts of a whole**



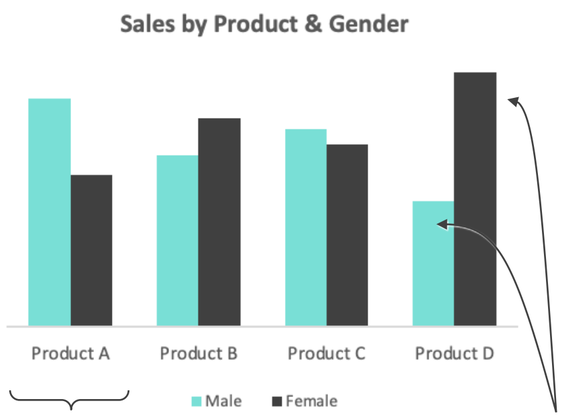
*In this example, **57%** of customers have **Gold** status, followed by **Silver** (25%) and **Bronze** (18%)*

*Copyright Maven Analytics, LLC

INTERPRETING COMMON CHARTS

CLUSTERED BAR/COLUMN

Used to show **composition across categories**



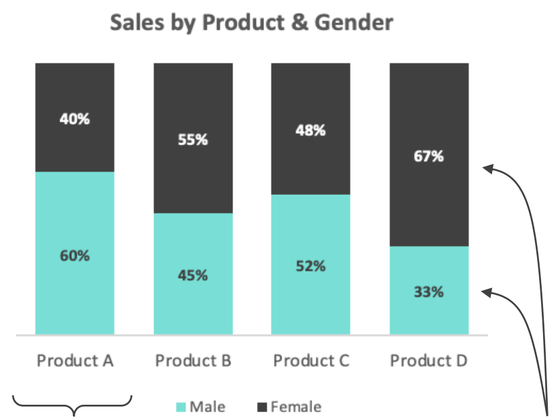
*Each cluster represents a **category**, and each individual bar or column represents a **subcategory***

*Subcategories can be compared by **the length of each bar or column***

*In this example, **products A & C** skew towards **male customers** while products B & D skew towards females*

STACKED BAR/COLUMN

Used to show **composition across categories**



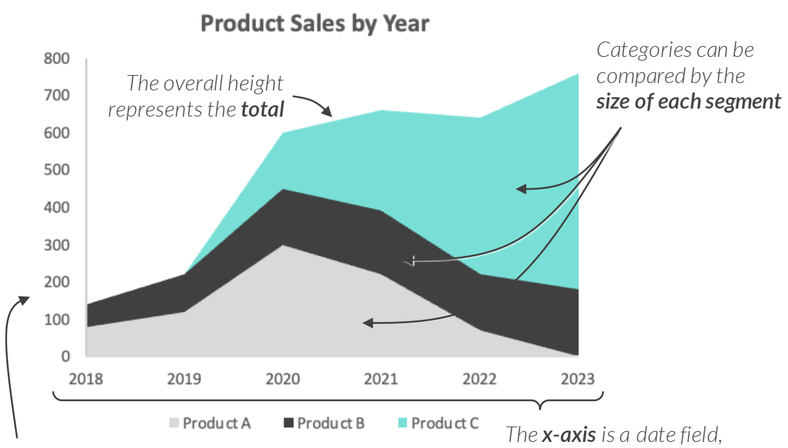
*Each bar or column represents a **category**, and each segment represents a **subcategory***

*Subcategories can be compared by the **size of each segment** (either volume or % of total)*

*In this example, male customers account for **60%** of sales for **product A** but only **33%** of sales for **product D***

AREA CHART

Used to show **changes in composition** over time



*The overall height represents the **total***

*Categories can be compared by the **size of each segment***

*The **y-axis** is a numerical field*

*The **x-axis** is a date field, ordered chronologically*

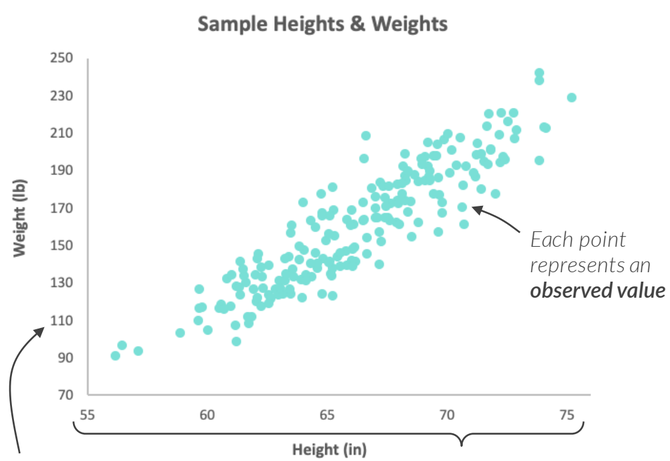
*In this example, **product C** launched in 2020, grew steadily, and accounted for the **majority of sales by 2023***

*Copyright Maven Analytics, LLC

INTERPRETING COMMON CHARTS

SCATTER PLOT

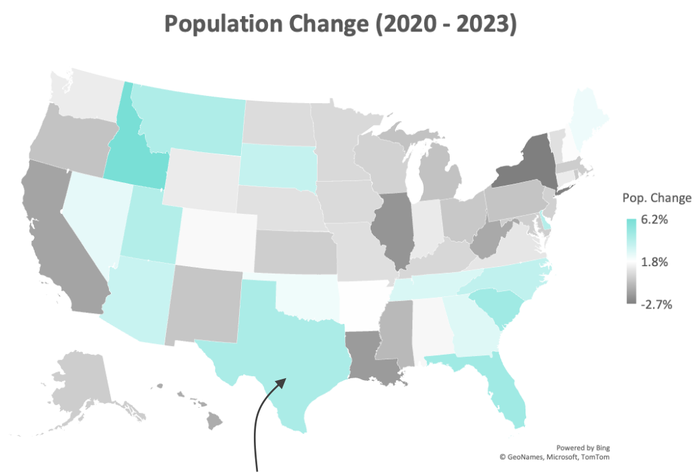
Used to show **relationships between variables**



*In this example, we see a **strong positive correlation** between height and weight (as height increases, weight tends to increase as well)*

FILLED MAP

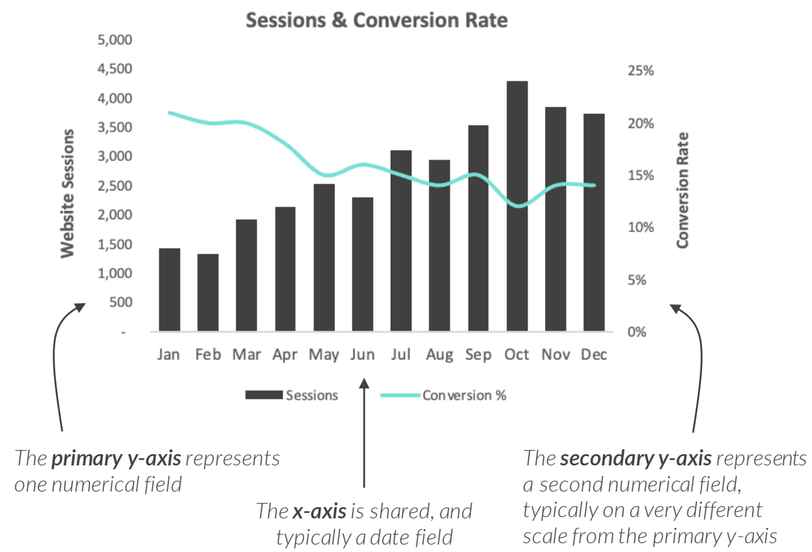
Used to visualize **geospatial data**



*In this example, **Idaho** and **Florida** saw the fastest population growth from 2020-2023, while **New York** and **Illinois** populations declined*

DUAL-AXIS CHART

Used to show **data on multiple scales**



*In this example, website sessions **trended up from Jan – Dec** and peaked in October while **conversion rate dropped under 15%***

*Copyright Maven Analytics, LLC

BASIC DATA PROFILING

Data profiling is about exploring and understanding the key characteristics of a dataset

- This can take many forms, but typically involves using a combination of descriptive statistics and visuals like histograms, bar charts, box plots or scatter plots to develop a "profile" of the data

Count = **20**
(# of transactions)

Distinct Count = **3**
(# of distinct product categories)

Conditional Count
(# records by category)

Coffee = **8**
Tea = **9**
Bakery = **3**

transaction_id	product_id	product_category	product_name	quantity	unit_price
1	3	Coffee	Gourmet brewed coffee	2	$3.00
2	9	Tea	Brewed Chai tea	2	$3.10
3	1	Coffee	Drip coffee	1	$2.00
4	9	Tea	Brewed Chai tea	2	$3.10
5	12	Bakery	Scone	1	$3.00
6	1	Coffee	Drip coffee	1	$2.00
7	3	Coffee	Gourmet brewed coffee	2	$2.00
8	5	Coffee	Barista Espresso	1	$4.25
9	9	Tea	Brewed Chai tea	1	$2.55
10	3	Coffee	Gourmet brewed coffee	2	$3.50
11	8	Tea	Brewed Black tea	1	$3.00
12	9	Tea	Brewed Chai tea	1	$3.10
13	5	Coffee	Barista Espresso	2	$3.00
14	7	Tea	Brewed Green tea	2	$3.00
15	12	Bakery	Scone	1	$3.75
16	6	Tea	Brewed herbal tea	1	$2.50
17	3	Coffee	Gourmet brewed coffee	2	$3.50
18	9	Tea	Brewed Chai tea	2	$2.55
19	11	Bakery	Biscotti	1	$3.25
20	9	Tea	Brewed Chai tea	1	$2.55

Range = **$2.00 - $4.25**
(max & min unit price)

Mean = **$2.94**
(average unit price)

Median = **$3.00**
("middle" unit price)

Mode = **$3.00**
(most frequently observed unit price)

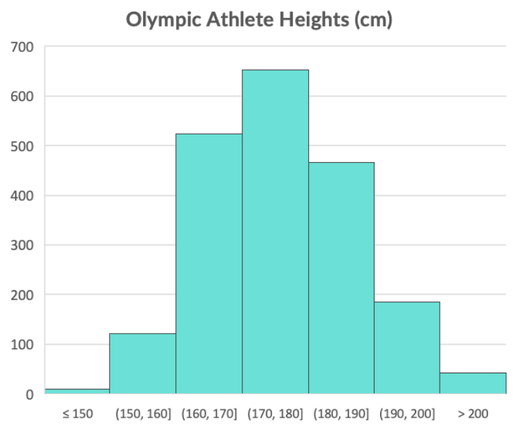
75th Percentile = **$3.14**
(75% of observations fall below this value)

*Copyright Maven Analytics, LLC

BASIC DATA PROFILING

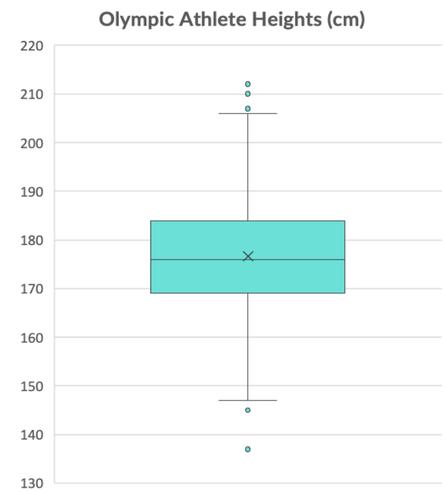
Data profiling is about exploring and understanding the key characteristics of a dataset

- This can take many forms, but typically involves using a combination of descriptive statistics and visuals like histograms, bar charts, box plots or scatter plots to develop a "profile" of the data



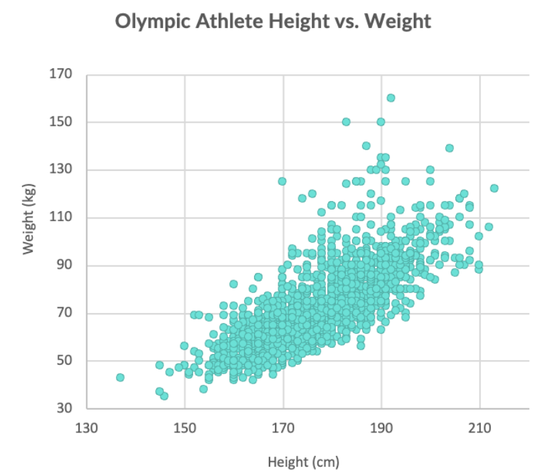
Histogram

Shows the frequency distribution, and can be used to identify outliers or compare metrics like variance or skew



Box Plot

Shows the min, max, median, mean, and middle 50% of values, and can be used to plot outliers or compare distributions across categories



Scatter Plot

Shows the relationship between two numerical variables, and can be used to identify correlation or impute missing values

*Copyright Maven Analytics, LLC

QUALITY ASSURANCE

Quality assurance (QA) is a critical step in every data analysis workflow, and involves identifying and correcting issues that could potentially skew your results or findings

*"Are there any **missing** or **empty** values in the data?"*

*"Are all fields and data types **encoded properly**?"*

*"Is there any **risk of bias** in the data collection process?"*

*"Do any text fields need to be **cleaned or standardized**?"*

*"Are there any **outliers** that might skew our results?"*

*Copyright Maven Analytics, LLC

COMMON DATA QUALITY ISSUES

Quality assurance (QA) is a critical step in every data analysis workflow, and involves identifying and correcting issues that could potentially skew your results or findings

First Name	Last Name	City	State	Postal Code	Telephone	HH Income	Birthdate
Jennifer	Mcgrath	Fieldon	IL	62031	6,183,763,064	$32,000	7/27/99
Susan	Rodriguez	Maplewood	NJ	7040	9,082,641,670	$45,000	
Wayne	Nielson	Dahlgren	VA	22448	5,406,447,658	$75,000	
John	Depaul	Baton Rouge	LA	70815	2,252,744,802	$60,000	8/27/67
Joseph	Martinez	Portland	OR	97205	5,034,022,075	$120,000	12/10/60
Joseph	Martinez	Portland	OR	97205	5,034,022,075	$120,000	12/10/60
Veronica	Comerford	Tuskegee	AL	36083	3,347,257,343	$50,000	
Beverly	Nixon	Long Island	NY	11101	3,476,260,700	$45,000,000	3/5/62
Ivan	Layton	Albany	New York	12207	5,184,317,602	$105,000	
Debra	Ponce	Dallas	TX	75247	2,544,883,212	$75,000	
Meg	Bryner	Providence	RI	2905	4,012,222,097	$90,000	
Sue	Gay	Cataract	WI	54620	6,082,728,021	$100,000	
Mark	Flores	Saginaw	MI	48607	9,896,694,705	$55,000	
Julie	Purington	Piscataway	NJ	8854	7,329,814,756	$25,000	3/11/93
Grace	Renninger	Chicago	IL	60661	7,732,504,788	$95,000	
James	Weiss	Houston	TX	77060	8,322,324,190	$135,000	
Addie	Stevenson	Jacksonville			9,042,482,500	$80,000	
Laura	Hailey	Ephraim	Utah	84627	4,352,831,109	$62,000	1/13/94
Daniel	Long						
Queen	Watson	Los Angeles	CA	90017	7,604,889,781	$110,000	5/19/68

ISSUE: Incorrect data types

SOLUTION:

- Format or re-encode fields to ensure accurate and consistent data types *(numeric, string, date, etc.)*

Pay extra attention to fields containing **dates**, **zip codes** or **phone numbers**, which are often encoded incorrectly

*Copyright Maven Analytics, LLC

COMMON DATA QUALITY ISSUES

Quality assurance (QA) is a critical step in every data analysis workflow, and involves identifying and correcting issues that could potentially skew your results or findings

First Name	Last Name	City	State	Postal Code	Telephone	HH Income	Birthdate
Jennifer	Mcgrath	Fieldon	IL	62031	6,183,763,064	$32,000	7/27/99
Susan	Rodriguez	Maplewood	NJ	7040	9,082,641,670	$45,000	
Wayne	Nielson	Dahlgren	VA	22448	5,406,447,658	$75,000	
John	Depaul	Baton Rouge	LA	70815	2,252,744,802	$60,000	8/27/67
Joseph	Martinez	Portland	OR	97205	5,034,022,075	$120,000	12/10/60
Joseph	Martinez	Portland	OR	97205	5,034,022,075	$120,000	12/10/60
Veronica	Comerford	Tuskegee	AL	36083	3,347,257,343	$50,000	
Beverly	Nixon	Long Island	NY	11101	3,476,260,700	$45,000,000	3/5/62
Ivan	Layton	Albany	New York	12207	5,184,317,602	$105,000	
Debra	Ponce	Dallas	TX	75247	2,544,883,212	$75,000	
Meg	Bryner	Providence	RI	2905	4,012,222,097	$90,000	
Sue	Gay	Cataract	WI	54620	6,082,728,021	$100,000	
Mark	Flores	Saginaw	MI	48607	9,896,694,705	$55,000	
Julie	Purington	Piscataway	NJ	8854	7,329,814,756	$25,000	3/11/93
Grace	Renninger	Chicago	IL	60661	7,732,504,788	$95,000	
James	Weiss	Houston	TX	77060	8,322,324,190	$135,000	
Addie	Stevenson	Jacksonville			9,042,482,500	$80,000	
Laura	Hailey	Ephraim	Utah	84627	4,352,831,109	$62,000	1/13/94
Daniel	Long						
Queen	Watson	Los Angeles	CA	90017	7,604,889,781	$110,000	5/19/68

⚠ **ISSUE:** Duplicate records

SOLUTION:

- Delete redundant records to avoid inflation or bias, unless duplicates serve a specific, valid purpose *(i.e. up-sampling)*

In most cases you'll only want to remove records where **all fields are duplicates** (not just some columns)

*Copyright Maven Analytics, LLC

COMMON DATA QUALITY ISSUES

Quality assurance (QA) is a critical step in every data analysis workflow, and involves identifying and correcting issues that could potentially skew your results or findings

First Name	Last Name	City	State	Postal Code	Telephone	HH Income	Birthdate
Jennifer	Mcgrath	Fieldon	IL	62031	6,183,763,064	$32,000	7/27/99
Susan	Rodriguez	Maplewood	NJ	7040	9,082,641,670	$45,000	
Wayne	Nielson	Dahlgren	VA	22448	5,406,447,658	$75,000	
John	Depaul	Baton Rouge	LA	70815	2,252,744,802	$60,000	8/27/67
Joseph	Martinez	Portland	OR	97205	5,034,022,075	$120,000	12/10/60
Joseph	Martinez	Portland	OR	97205	5,034,022,075	$120,000	12/10/60
Veronica	Comerford	Tuskegee	AL	36083	3,347,257,343	$50,000	
Beverly	Nixon	Long Island	NY	11101	3,476,260,700	$45,000,000	3/5/62
Ivan	Layton	Albany	New York	12207	5,184,317,602	$105,000	
Debra	Ponce	Dallas	TX	75247	2,544,883,212	$75,000	
Meg	Bryner	Providence	RI	2905	4,012,222,097	$90,000	
Sue	Gay	Cataract	WI	54620	6,082,728,021	$100,000	
Mark	Flores	Saginaw	MI	48607	9,896,694,705	$55,000	
Julie	Purington	Piscataway	NJ	8854	7,329,814,756	$25,000	3/11/93
Grace	Renninger	Chicago	IL	60661	7,732,504,788	$95,000	
James	Weiss	Houston	TX	77060	8,322,324,190	$135,000	
Addie	Stevenson	Jacksonville			9,042,482,500	$80,000	
Laura	Hailey	Ephraim	Utah	84627	4,352,831,109	$62,000	1/13/94
Daniel	Long						
Queen	Watson	Los Angeles	CA	90017	7,604,889,781	$110,000	5/19/68

ISSUE: Inconsistent categories

SOLUTION:

- Determine a standard and find & replace inconsistent values
- Preserve the original data and add a new, standardized version for analysis

Using **distinct counts** or **lists of unique values** is a great way to identify inconsistencies in categorical fields

*Copyright Maven Analytics, LLC

COMMON DATA QUALITY ISSUES

Quality assurance (QA) is a critical step in every data analysis workflow, and involves identifying and correcting issues that could potentially skew your results or findings

First Name	Last Name	City	State	Postal Code	Telephone	HH Income	Birthdate
Jennifer	Mcgrath	Fieldon	IL	62031	6,183,763,064	$32,000	7/27/99
Susan	Rodriguez	Maplewood	NJ	7040	9,082,641,670	$45,000	
Wayne	Nielson	Dahlgren	VA	22448	5,406,447,658	$75,000	
John	Depaul	Baton Rouge	LA	70815	2,252,744,802	$60,000	8/27/67
Joseph	Martinez	Portland	OR	97205	5,034,022,075	$120,000	12/10/60
Joseph	Martinez	Portland	OR	97205	5,034,022,075	$120,000	12/10/60
Veronica	Comerford	Tuskegee	AL	36083	3,347,257,343	$50,000	
Beverly	Nixon	Long Island	NY	11101	3,476,260,700	$45,000,000	3/5/62
Ivan	Layton	Albany	New York	12207	5,184,317,602	$105,000	
Debra	Ponce	Dallas	TX	75247	2,544,883,212	$75,000	
Meg	Bryner	Providence	RI	2905	4,012,222,097	$90,000	
Sue	Gay	Cataract	WI	54620	6,082,728,021	$100,000	
Mark	Flores	Saginaw	MI	48607	9,896,694,705	$55,000	
Julie	Purington	Piscataway	NJ	8854	7,329,814,756	$25,000	3/11/93
Grace	Renninger	Chicago	IL	60661	7,732,504,788	$95,000	
James	Weiss	Houston	TX	77060	8,322,324,190	$135,000	
Addie	Stevenson	Jacksonville			9,042,482,500	$80,000	
Laura	Hailey	Ephraim	Utah	84627	4,352,831,109	$62,000	1/13/94
Daniel	Long						
Queen	Watson	Los Angeles	CA	90017	7,604,889,781	$110,000	5/19/68

ISSUE: Empty values

SOLUTION:

- **KEEP** if the empty values are legitimate and won't skew your analysis
- **IMPUTE** *(substitute)* if you can accurately populate or estimate the data
- **REMOVE** if you have a large volume of data and can't impute the empty values

Never replace an empty or null value with a **zero** unless you know it's a valid replacement and won't skew your analysis

*Copyright Maven Analytics, LLC

COMMON DATA QUALITY ISSUES

Quality assurance (QA) is a critical step in every data analysis workflow, and involves identifying and correcting issues that could potentially skew your results or findings

First Name	Last Name	City	State	Postal Code	Telephone	HH Income	Birthdate
Jennifer	Mcgrath	Fieldon	IL	62031	6,183,763,064	$32,000	7/27/99
Susan	Rodriguez	Maplewood	NJ	7040	9,082,641,670	$45,000	
Wayne	Nielson	Dahlgren	VA	22448	5,406,447,658	$75,000	
John	Depaul	Baton Rouge	LA	70815	2,252,744,802	$60,000	8/27/67
Joseph	Martinez	Portland	OR	97205	5,034,022,075	$120,000	12/10/60
Joseph	Martinez	Portland	OR	97205	5,034,022,075	$120,000	12/10/60
Veronica	Comerford	Tuskegee	AL	36083	3,347,257,343	$50,000	
Beverly	Nixon	Long Island	NY	11101	3,476,260,700	$45,000,000	3/5/62
Ivan	Layton	Albany	New York	12207	5,184,317,602	$105,000	
Debra	Ponce	Dallas	TX	75247	2,544,883,212	$75,000	
Meg	Bryner	Providence	RI	2905	4,012,222,097	$90,000	
Sue	Gay	Cataract	WI	54620	6,082,728,021	$100,000	
Mark	Flores	Saginaw	MI	48607	9,896,694,705	$55,000	
Julie	Purington	Piscataway	NJ	8854	7,329,814,756	$25,000	3/11/93
Grace	Renninger	Chicago	IL	60661	7,732,504,788	$95,000	
James	Weiss	Houston	TX	77060	8,322,324,190	$135,000	
Addie	Stevenson	Jacksonville			9,042,482,500	$80,000	
Laura	Hailey	Ephraim	Utah	84627	4,352,831,109	$62,000	1/13/94
Daniel	Long						
Queen	Watson	Los Angeles	CA	90017	7,604,889,781	$110,000	5/19/68

⚠ **ISSUE:** Statistical outliers

SOLUTION:

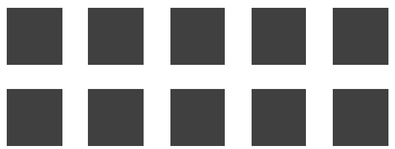
- **KEEP** the outlier if it's valid and you want to capture its impact on the data
- **IMPUTE** *(substitute)* outliers with caps or replacement values
- **REMOVE** the outlier if it's an impossible value or significantly skews the analysis
- **TRANSFORM** values to reduce the effect of the outlier

In this case we see an average household income of ***$2,400,000*** *when we include the outlier. Does that feel accurate to you?*

*Copyright Maven Analytics, LLC

STRUCTURED VS. UNSTRUCTURED DATA

STRUCTURED



- ✓ Typically organized in a **tabular format** (CSV, Excel, XML) and stored in spreadsheets, relational databases or data warehouses
- ✓ Contains **quantitative** data (numbers and values) which can easily be searched or analyzed using tools like Excel, SQL or business intelligence platforms

UNSTRUCTURED

- ✓ Can take a **huge variety of formats** (PDFs, images, audio, videos, etc.), typically stored as native files or documents in data lakes
- ✓ Contains **qualitative** data (multimedia files, emails, PDFs, etc.) which contain valuable information but can be very difficult to analyze using common tools or methods

It's estimated that **only ~20% of enterprise data is structured**, so there's a massive opportunity to extract value and insight from the massive amount of unstructured data in existence

** https://www.sinequa.com/resources/blog/structured-vs-unstructured-data-the-ultimate-guide*

**Copyright Maven Analytics, LLC*

ACCESSING DATA

OPEN-SOURCE

Public data sources typically managed by government organizations or non-profits

EXAMPLES: Data.gov, U.S. Census Bureau, World Health Organization, World Bank, etc.

PUBLIC API

Live connections that allow you to extract or access data from public internet sources

EXAMPLES: Google Maps, Wikipedia, Open Weather Map, Yahoo Finance, etc.

INTERNAL

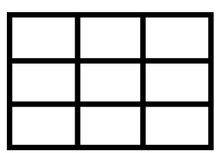
Private data containing business-specific details like customer information or transactional records

EXAMPLES: Customer database (CRM), transactional records, web analytics platform, etc.

If you are looking for free downloadable datasets to practice or showcase your skills, check out Maven Analytics' **Data Playground** and **Guided Project** library, or browse other resources like Kaggle or data.world

*Copyright Maven Analytics, LLC

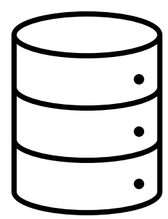
STORING DATA



FLAT FILE

Static data extracts saved to a local machine or cloud drive, typically used for one-off projects or quick, ad hoc analysis

EXAMPLES: CSV, Excel



DATABASE

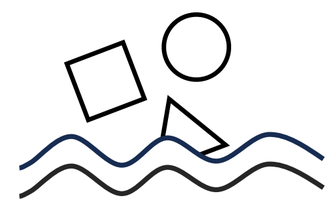
Collection of related tables stored in a database management system (DBMS), typically used for Online Transactional Processing (OLTP)

EXAMPLES: MySQL, Amazon RDS, Microsoft Access

DATA WAREHOUSE

A database or collection of data sourced from multiple databases, structured to support analytics needs and optimized for Online Analytical Processing (OLAP)

EXAMPLES: RedShift, Google BigQuery, Snowflake



DATA LAKE

A repository of structured and unstructured data typically stored in a raw, unprocessed state

EXAMPLES: AWS S3, Azure Data Lake, Google Cloud

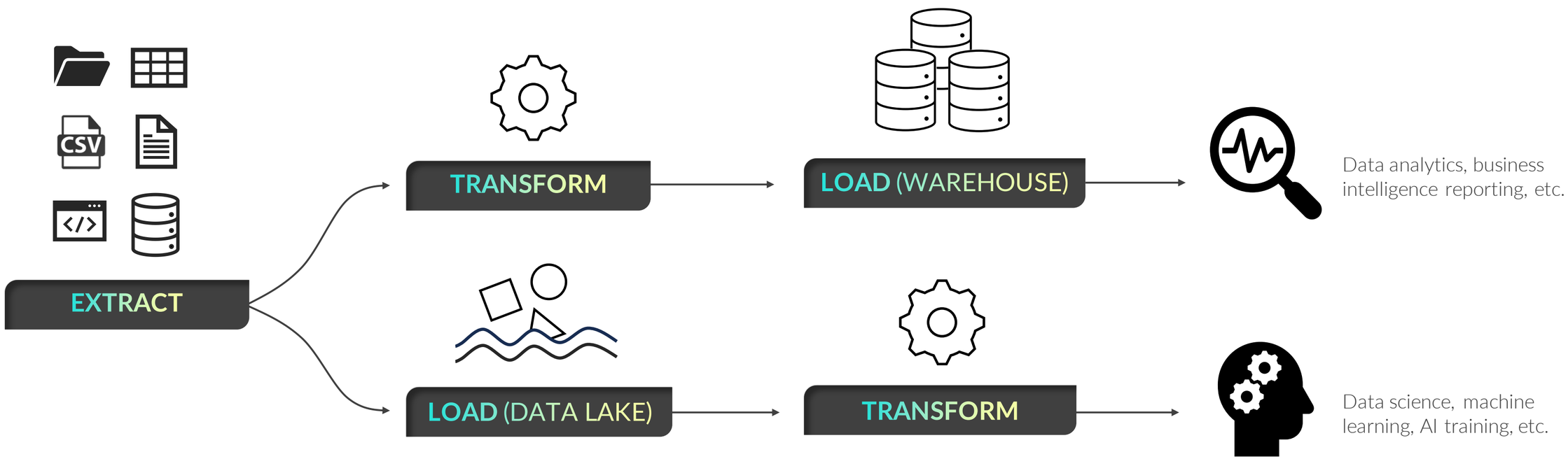
In most cases, **data engineers** are the ones responsible for setting up these storage systems, as well as the pipelines and automations that help move data from one place to another

*Copyright Maven Analytics, LLC

ETL & ELT

Raw data typically needs to be **extracted**, **transformed** and **loaded** for analysis

- Data can be transformed before loading (**ETL**) or loaded directly into storage (**ELT**), depending on the use case; ETL is common for data analytics and BI, and ELT is usually used for machine learning or AI



*Copyright Maven Analytics, LLC

KEY TAKEAWAYS

- The ability to **interpret** and **manage** data is more important than ever
- **Profiling** and **QA** help you understand and describe the qualities of a dataset
- Data can be **structured** or **unstructured**, accessed via internal or public sources
- Data is typically stored in **flat files**, **databases**, **warehouses** or **data lakes**
- Raw data is **extracted**, **transformed** and **loaded** for analysis

*Copyright Maven Analytics, LLC

ANALYZING DATA

*Copyright Maven Analytics, LLC

DATA ANALYSIS

Data analysis is about identifying insights that help inspire action or inform data-driven decisions, and can generally be categorized as either **exploratory** or **explanatory**

EXPLORATORY (EDA)

- Goal is to **explore and profile** the data to see what insights emerge
- Unbiased, open-minded analysis *(no specific objective or assumptions)*
- Helps you understand the data and identify interesting patterns and trends

EXPLANATORY

- Goal is to **tell a specific story** or explain what happened and why
- Purpose-driven, root-cause analysis *(answers a specific question)*
- Identifies key business drivers and delivers insights & recommendations

*Copyright Maven Analytics, LLC

EXPLORATORY VS. EXPLANATORY

EXPLORATORY (EDA)

"How does performance compare across different segments or categories?"

"Are there any interesting correlations or relationships between variables?"

"How is the data distributed, and are there any outliers or missing values?"

"How are key metrics changing over time? Are there any predictable seasonal trends?"

EXPLANATORY

"Why did our website conversion rate drop by 25% last month?"

"Which products have seen the largest increase in gross revenue this year?"

"Which user segments deliver the highest customer lifetime value (LTV)?"

"Based on geographical trends, where should we open our next store location?"

*Copyright Maven Analytics, LLC

DATA ANALYSIS FRAMEWORK

To transform data into insight and action, **you need to think like an analyst**

- Analytical thinking means identifying the problem, setting clear expectations, collecting and analyzing the information you need, and leveraging insights and findings to help influence decisions

KEY QUESTIONS

*What **specific problem** am I trying to solve?*

*Who are the **key stakeholders**, and how will this help them?*

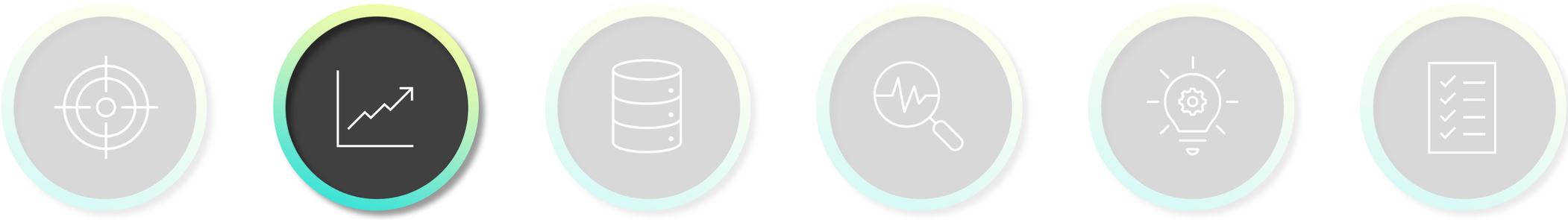
*Does this align with broader **business priorities and strategy**?*

*Copyright Maven Analytics, LLC

DATA ANALYSIS FRAMEWORK

To transform data into insight and action, **you need to think like an analyst**

- Analytical thinking means identifying the problem, setting clear expectations, collecting and analyzing the information you need, and leveraging insights and findings to help influence decisions



KEY QUESTIONS

*What exactly does **a successful outcome** look like?*

*Which **specific metrics** will help me quantify success?*

*What data will I need to **capture and track key metrics**?*

*Copyright Maven Analytics, LLC

DATA ANALYSIS FRAMEWORK

To transform data into insight and action, **you need to think like an analyst**

- Analytical thinking means identifying the problem, setting clear expectations, collecting and analyzing the information you need, and leveraging insights and findings to help influence decisions

KEY QUESTIONS

*Where is the data stored, and **how can I access it**?*

*Are there **data quality issues** that might skew my analysis?*

*Will I need to **transform the data** to support this analysis?*

*Copyright Maven Analytics, LLC

DATA ANALYSIS FRAMEWORK

To transform data into insight and action, **you need to think like an analyst**

- Analytical thinking means identifying the problem, setting clear expectations, collecting and analyzing the information you need, and leveraging insights and findings to help influence decisions

KEY QUESTIONS

*Which **views of the data** can help support my analysis?*

*What types of **patterns and trends** are emerging?*

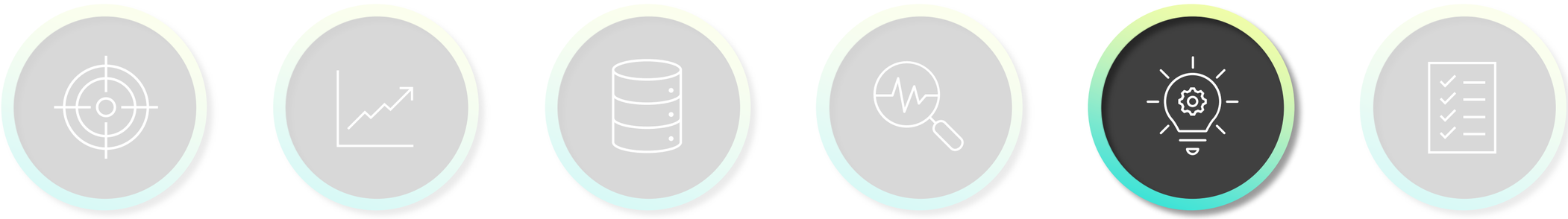
*Are there **actionable insights** that could help drive success outcomes?*

*Copyright Maven Analytics, LLC

DATA ANALYSIS FRAMEWORK

To transform data into insight and action, **you need to think like an analyst**

- Analytical thinking means identifying the problem, setting clear expectations, collecting and analyzing the information you need, and leveraging insights and findings to help influence decisions



KEY QUESTIONS

*How can I **craft a narrative** to clearly summarize my analysis?*

*What's the most effective way to **visually communicate** my findings?*

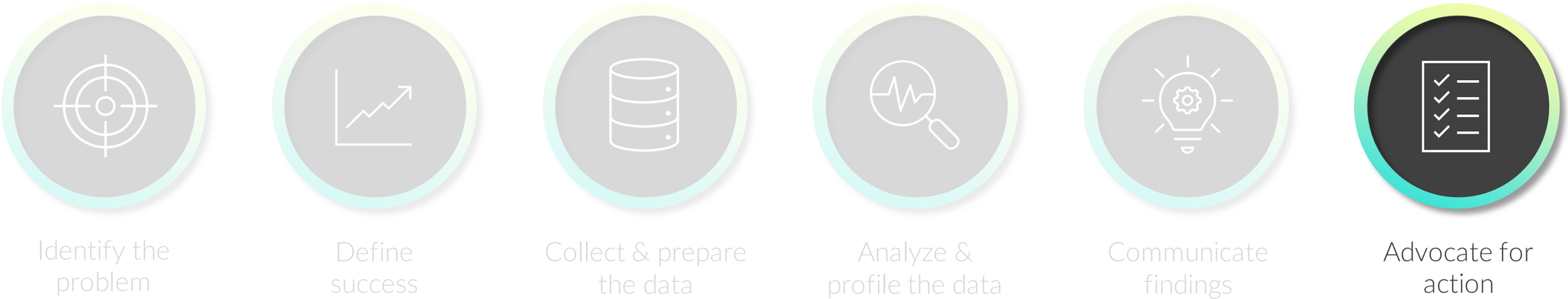
*Are my insights **compelling, and supported by the data**?*

*Copyright Maven Analytics, LLC

DATA ANALYSIS FRAMEWORK

To transform data into insight and action, **you need to think like an analyst**

- Analytical thinking means identifying the problem, setting clear expectations, collecting and analyzing the information you need, and leveraging insights and findings to help influence decisions



KEY QUESTIONS

*What **data-driven recommendations** am I proposing?*

*Do they **directly impact the outcomes** I care most about?*

*How could we potentially **quantify the business impact**?*

*Copyright Maven Analytics, LLC

FOUR CATEGORIES OF ANALYTICS

Data analytics techniques are commonly described as part of four distinct categories: **descriptive**, **diagnostic**, **predictive** and **prescriptive**

DESCRIPTIVE
What happened?

DIAGNOSTIC
Why did it happen?

PREDICTIVE
What will happen next?

PRESCRIPTIVE
What's our course of action?

These categories are often plotted on a linear "value" scale, which implies that the goal is to evolve from one category to the next. The truth is that **each type of analytics plays an important role and can add value when applied in the right context**.

*Copyright Maven Analytics, LLC

FOUR CATEGORIES OF ANALYTICS

EXAMPLE | VISITING THE DOCTOR

DESCRIPTIVE

"Your cholesterol level is 215"

DIAGNOSTIC

"Your cholesterol level is 215. This is on the high end and likely due to lack of exercise and too much saturated fat in your diet."

PREDICTIVE

"If you maintain your current diet and lifestyle, your cholesterol level will continue to rise and increase your risk of cardiovascular disease."

PRESCRIPTIVE

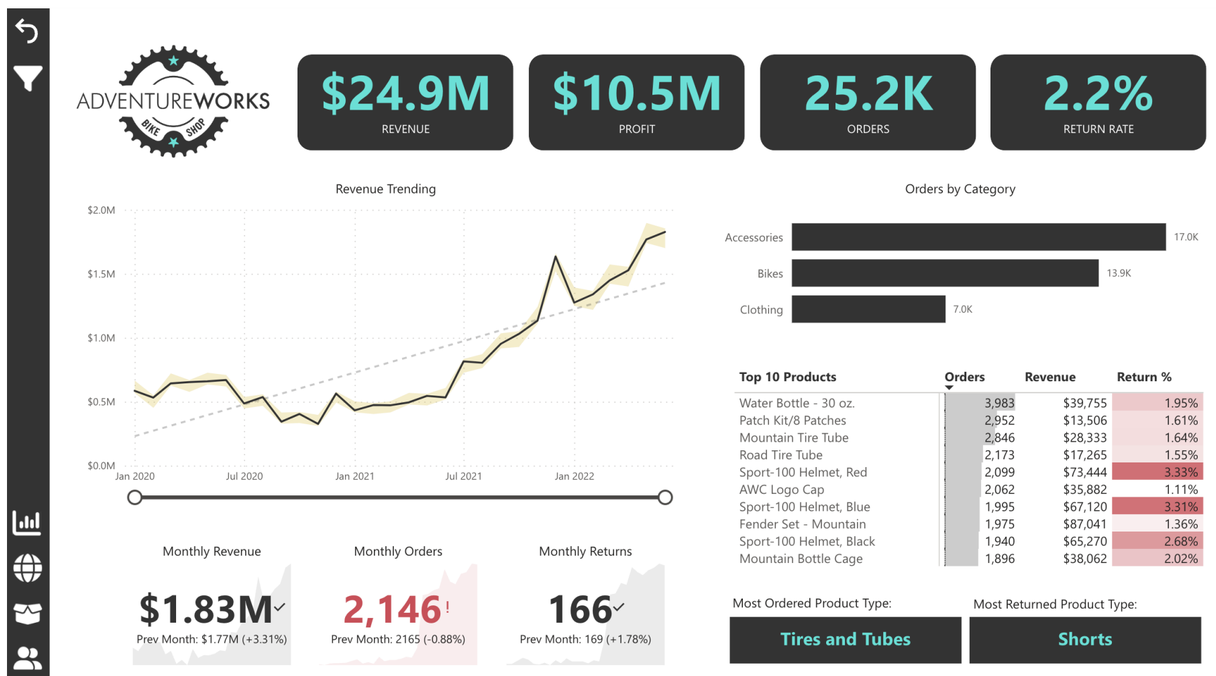
"Based on your test results, I've prescribed statins and recommended a new diet to lower your cholesterol and reduce the risk of heart disease."

*Copyright Maven Analytics, LLC

DESCRIPTIVE ANALYTICS

DESCRIPTIVE

Understanding what happened or is happening



Example business performance dashboard (Adventure Works)

EXAMPLES:

- KPI reporting
- Data visualization
- Exploratory analysis
- Data profiling

PEOPLE & ROLES:

- Business analysts
- Data visualization specialists
- Citizen analysts (non-specialists)

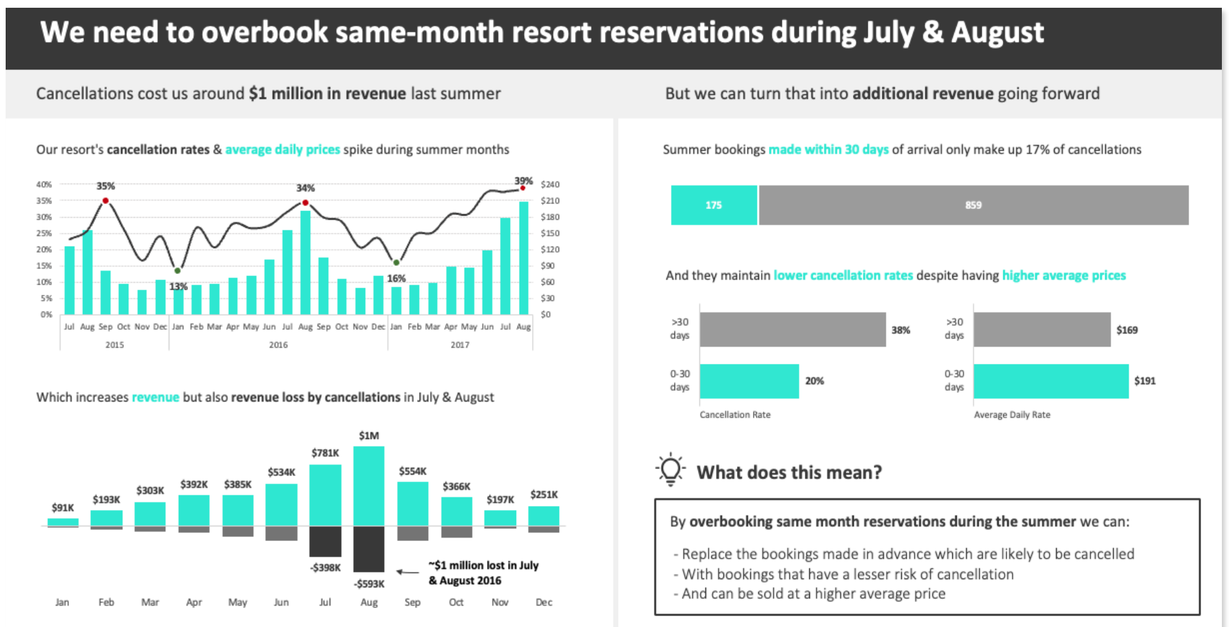
TOOLS:

*Copyright Maven Analytics, LLC

DIAGNOSTIC ANALYTICS

DIAGNOSTIC

Understanding why something happened



Example explanatory report (hotel booking analysis)

EXAMPLES:

- Explanatory analysis
- Regression modeling
- Root cause analysis
- Hypothesis testing

PEOPLE & ROLES:

- Business Intelligence Analysts
- Data Analysts
- Data Scientists

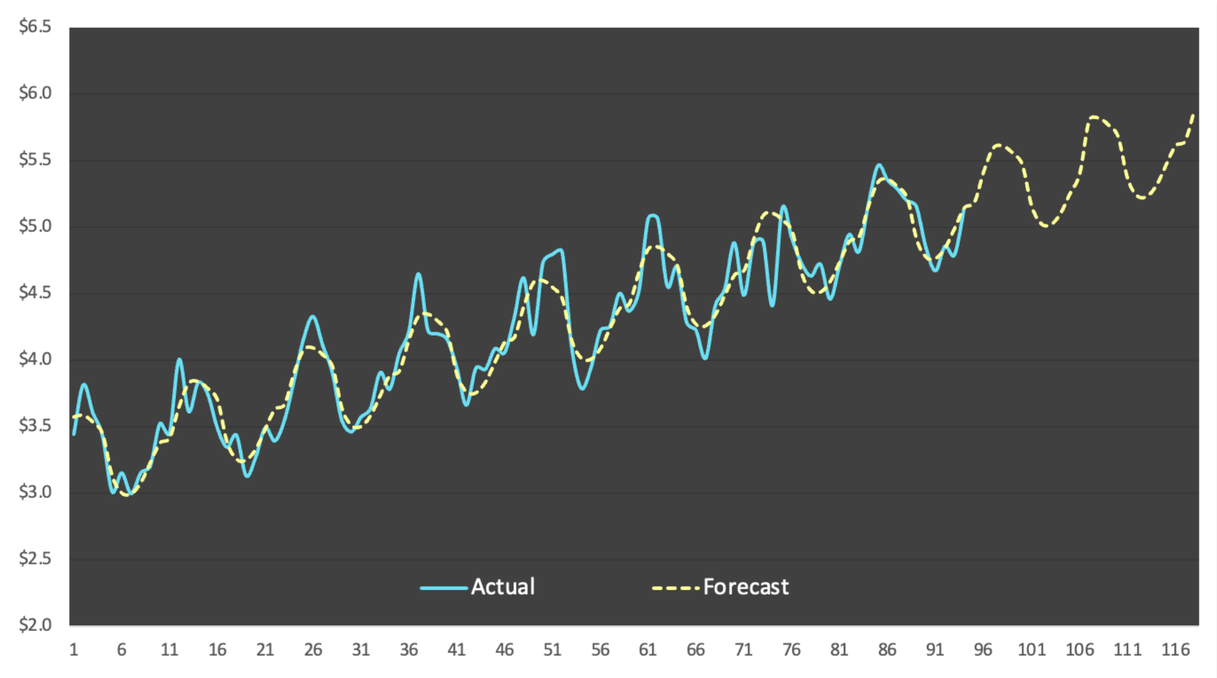
TOOLS:

*Copyright Maven Analytics, LLC

PREDICTIVE ANALYTICS

PREDICTIVE

Understanding what will likely happen next



Example forecast model with seasonality and trending

EXAMPLES:

- Time-series forecasting
- Machine learning
- Regression modeling
- Classification modeling

PEOPLE & ROLES:

- Data Scientists
- Machine Learning Engineers
- Business Intelligence Analysts

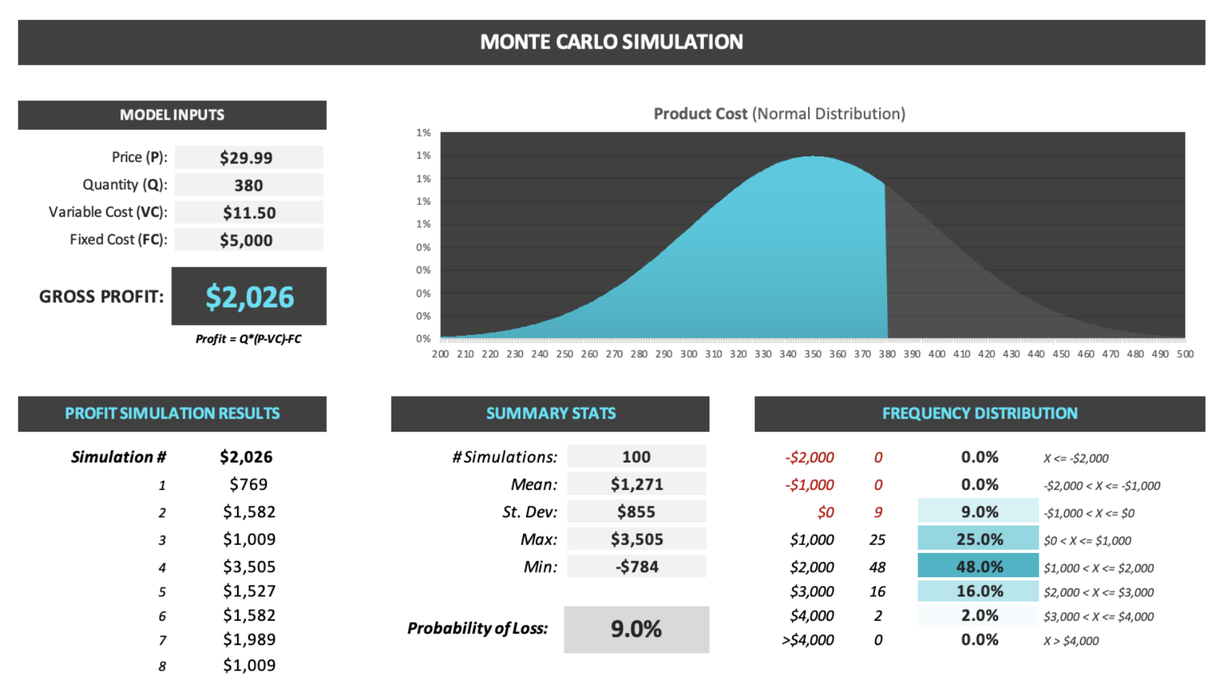
TOOLS:

*Copyright Maven Analytics, LLC

PRESCRIPTIVE ANALYTICS

PRESCRIPTIVE

Understanding the optimal course of action



MONTE CARLO SIMULATION

MODEL INPUTS

Price (P):	$29.99
Quantity (Q):	380
Variable Cost (VC):	$11.50
Fixed Cost (FC):	$5,000

GROSS PROFIT: $2,026

Profit = Q(P-VC)-FC*

PROFIT SIMULATION RESULTS

Simulation #	$2,026
1	$769
2	$1,582
3	$1,009
4	$3,505
5	$1,527
6	$1,582
7	$1,989
8	$1,009

SUMMARY STATS

# Simulations:	100
Mean:	$1,271
St. Dev:	$855
Max:	$3,505
Min:	-$784
Probability of Loss:	9.0%

FREQUENCY DISTRIBUTION

-$2,000	0	0.0%	X <= -$2,000
-$1,000	0	0.0%	-$2,000 < X <= -$1,000
$0	9	9.0%	-$1,000 < X <= $0
$1,000	25	25.0%	$0 < X <= $1,000
$2,000	48	48.0%	$1,000 < X <= $2,000
$3,000	16	16.0%	$2,000 < X <= $3,000
$4,000	2	2.0%	$3,000 < X <= $4,000
>$4,000	0	0.0%	X > $4,000

Example simulation model for scenario testing and profit optimization

EXAMPLES:

- Scenario testing
- Generative AI
- Reinforcement learning
- Recommendation engines

PEOPLE & ROLES:

- Data Scientists
- Machine Learning Engineers
- AI Engineers

TOOLS:

*Copyright Maven Analytics, LLC

KEY TAKEAWAYS

- Data analysis is generally considered either **exploratory** or **explanatory**
- Every analysis starts by **identifying the problem** and **defining success**
- Remember that the goal of any analysis is to **impact real-world outcomes**
- Analytics can be described as **descriptive**, **diagnostic**, **predictive** or **prescriptive**
- Every form of analytics can add value **when applied in the right context**

*Copyright Maven Analytics, LLC

COMMUNICATING WITH DATA

*Copyright Maven Analytics, LLC

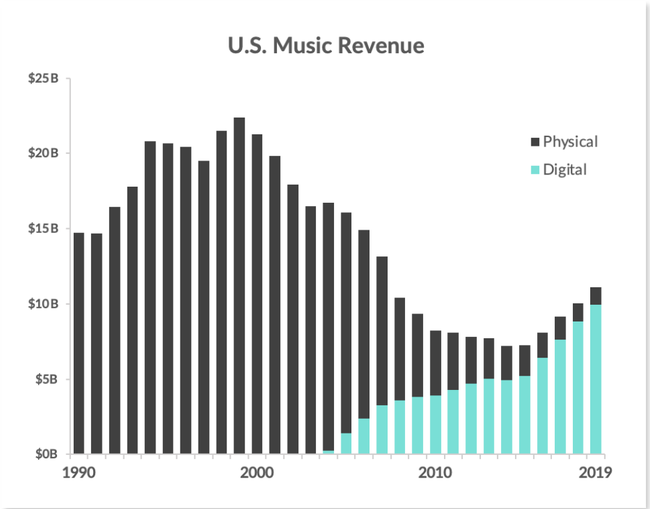
DATA VISUALIZATION & STORYTELLING

Data visualization is the graphical representation of data and information

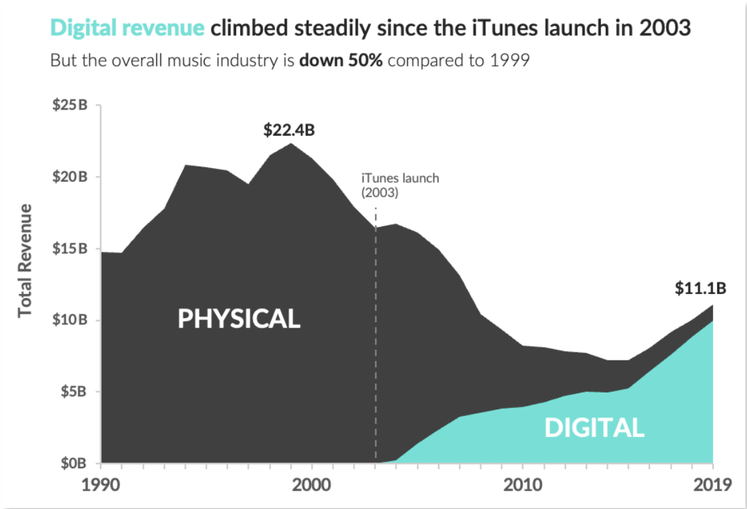
- The human brain isn't optimized to interpret numbers and tables; we need **clear patterns** and **visual cues** to help us quickly make sense of complex information
- **Data storytelling** is an extension of visualization, focused on creating narratives designed to communicate insights, influence decisions, or support specific data-driven recommendations

	REVENUE ($MM)		
Year	Physical	Digital	Total
1990	$14,751	$0	$14,751
1991	$14,705	$0	$14,705
1992	$16,444	$0	$16,444
1993	$17,775	$0	$17,775
1994	$20,818	$0	$20,818
1995	$20,668	$0	$20,668
1996	$20,423	$0	$20,423
1997	$19,492	$0	$19,492
1998	$21,505	$0	$21,505
1999	$22,381	$0	$22,381
2000	$21,266	$0	$21,266
2001	$19,836	$0	$19,836
2002	$17,926	$0	$17,926
2003	$16,471	$0	$16,471
2004	$16,450	$258	$16,708
2005	$14,655	$1,433	$16,088
2006	$12,515	$2,398	$14,913
2007	$9,849	[illegible]	[illegible]

Raw data



Data Visualization



Data Storytelling

*Copyright Maven Analytics, LLC

THE 10-SECOND RULE

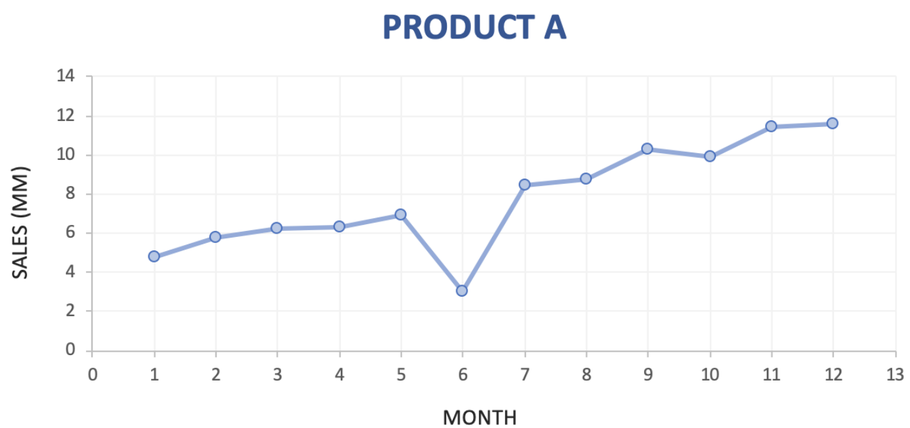
In **10 seconds**, what can you learn from the data below?

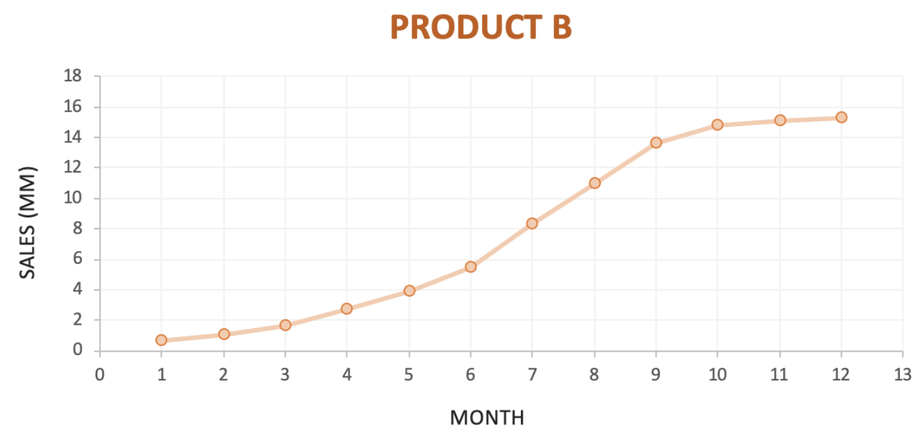
Product A		Product B		Product C		Product D	
Month	Sales (MM)	Month	Sales (MM)	Month	Sales (MM)	Month	Sales (MM)
1	4.80	1	0.67	1	4.53	1	8.35
2	5.78	2	1.05	2	4.61	2	7.72
3	6.24	3	1.62	3	4.74	3	12.05
4	6.34	4	2.67	4	5.10	4	7.70
5	6.95	5	3.91	5	5.32	5	7.05
6	3.02	6	5.49	6	5.70	6	11.05
7	8.45	7	8.36	7	5.77	7	6.95
8	8.79	8	10.99	8	6.32	8	6.39
9	10.30	9	13.58	9	6.56	9	9.50
10	9.93	10	14.81	10	6.64	10	4.83
11	11.40	11	15.13	11	18.50	11	4.03
12	11.56	12	15.26	12	19.80	12	8.03

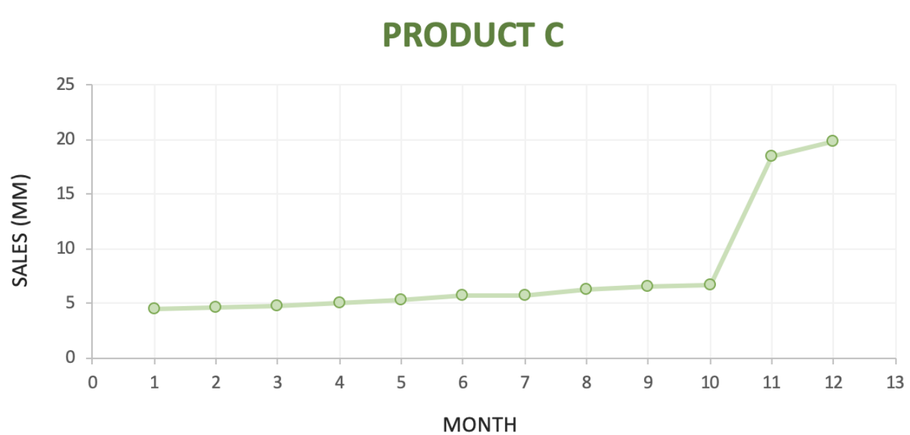
*Copyright Maven Analytics, LLC

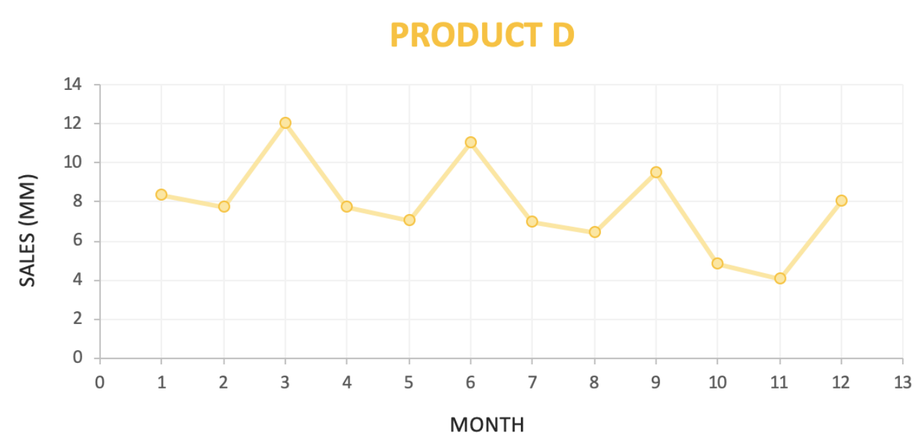
THE 10-SECOND RULE

How about now?









*Copyright Maven Analytics, LLC

3 KEY QUESTIONS

1

What **type of data** are you working with?

- Geospatial? Time-series? Hierarchical? Financial?

2

What do you want to **communicate**?

- Comparison? Composition? Relationship? Distribution?

3

Who is the **audience** and what do they need?

- Analyst? Manager? Executive? General public?

*Copyright Maven Analytics, LLC

3 KEY QUESTIONS

1 What **type of data** are you working with?

- Time-series
- Financial
- Geospatial
- Textual
- Categorical
- Funnel
- Hierarchical
- Survey

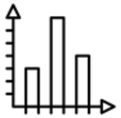
The type of data you're working with often determines **which type of visual will best represent it**; for example, using maps to represent geospatial data, line charts for time-series data, or bar/column charts for categorical data

*Copyright Maven Analytics, LLC

3 KEY QUESTIONS

2 What do you want to **communicate**?

COMPARISON



Used to **compare values over time or across categories**

Common visuals:

- Bar Chart
- Column Chart
- Clustered Column/Bar
- Line/Area Chart *(time series)*

COMPOSITION

Used to **break down the component parts of a whole**

Common visuals:

- Stacked Bar/Column Chart
- Pie/Donut Chart
- Stacked Area *(time series)*
- Tree Map/sunburst *(hierarchies)*

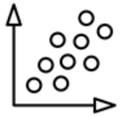
DISTRIBUTION

Used to **show the frequency of values within a series**

Common visuals:

- Histogram
- Box & Whisker
- Density Plot
- Heat Map

RELATIONSHIP



Used to **show correlation between multiple variables**

Common visuals:

- Scatter Plot
- Bubble Chart
- Heat Map
- Correlation Matrix

Keep it simple! Basic variations of **bar**, **column** & **line charts** are often the simplest – and clearest – way to visualize data

*Copyright Maven Analytics, LLC

3 KEY QUESTIONS

3 Who is the **audience** and what do they need?

THE **ANALYST**

Likes to see details and understand exactly what's happening at a granular level

- Tables or combo charts
- Granular detail to support root-cause analysis

THE **MANAGER**

Wants summarized data with clear, actionable insights to help operate the business

- Common charts & graphs
- Some detail, but only when it supports a specific insight

THE **EXECUTIVE**

Needs high-level, crystal clear KPIs to track business health and topline performance

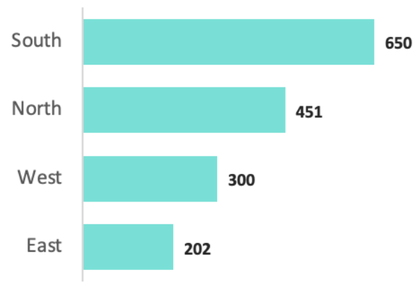
- KPI cards or simple charts
- Minimal detail, unless it adds critical context to KPIs

How you visualize and present your data is a function of **who will be consuming it**; an analyst or data specialist may want to see granular details, while managers and executives often prefer topline metrics and clear, data-driven insights

*Copyright Maven Analytics, LLC

ESSENTIAL VISUALS

BAR/COLUMN CHART

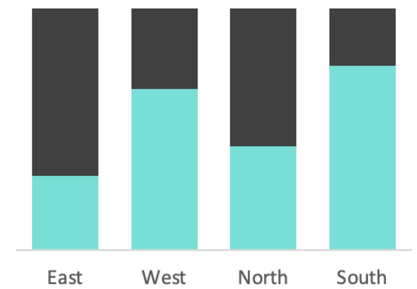


Used to show **comparisons between categories**

LINE CHART

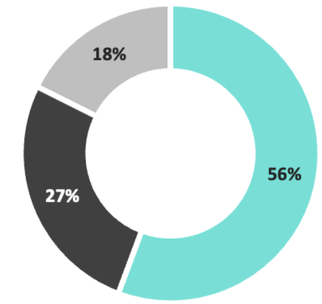
Used to show **trends over time**

STACKED BAR/COLUMN



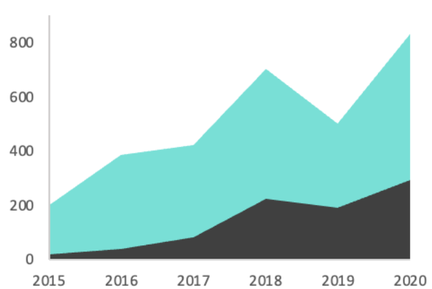
Used to show **composition across categories**

PIE/DONUT CHART



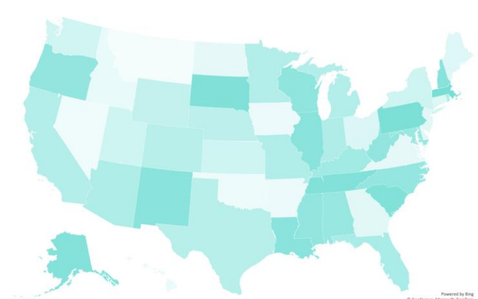
Used to visualize composition as **parts of a whole**

AREA CHART



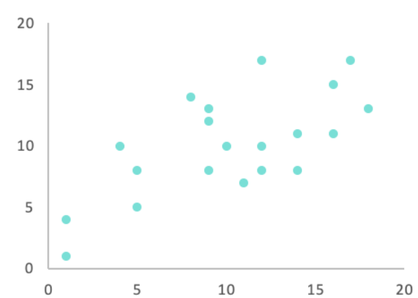
Used to show **changes in composition** over time

FILLED MAP



Used to visualize **geospatial data**

SCATTER PLOT



Used to show **relationships between variables**

TABLE/MATRIX

	A	B	C
East	96	77	13
West	6	39	39
North	25	75	9
South	26	7	90

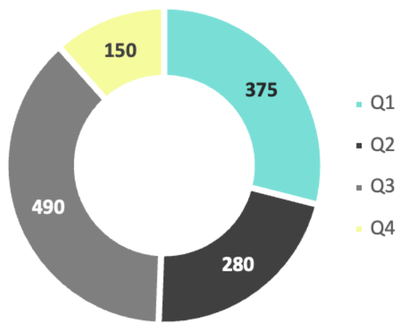
Used to show **numerical values** or highlight patterns

*Copyright Maven Analytics, LLC

COMMON VISUALIZATION MISTAKES

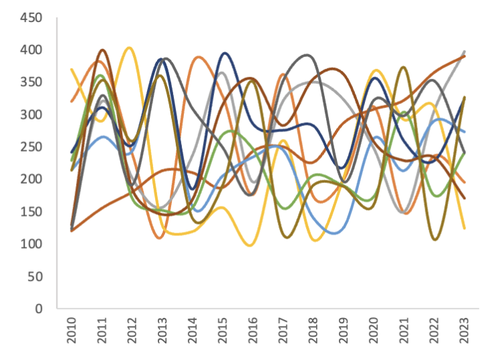
WRONG CHART TYPE

Using pie charts for time series data, line charts for categorical comparisons, etc.



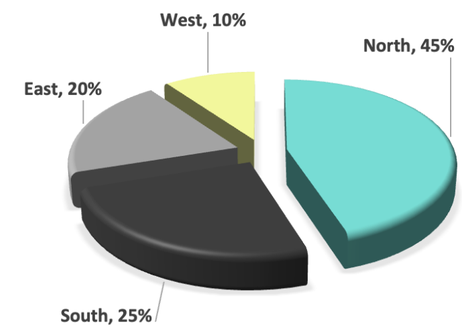
LACK OF FOCUS

Showing too many series or values, with no clear pattern or purpose



NOISE & DISTRACTIONS

Using unnecessary effects like 3D formats or background images

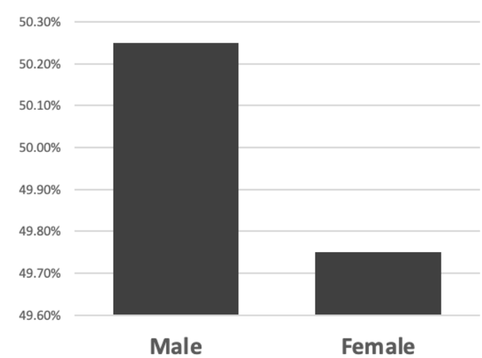


FORM OVER FUNCTION

Prioritizing design and visual aesthetics over clarity

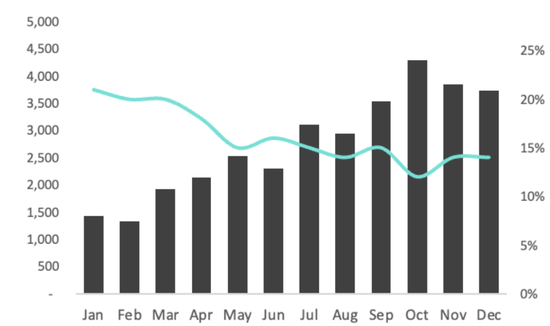
MISLEADING NARRATIVE

Exaggerating trends to make them appear more significant



MISSING CONTEXT

Failing to include series names, data labels, axis titles, etc.

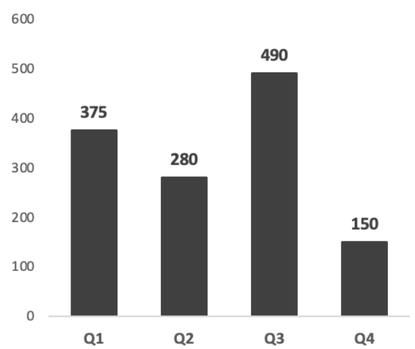


*Copyright Maven Analytics, LLC

COMMON VISUALIZATION MISTAKES

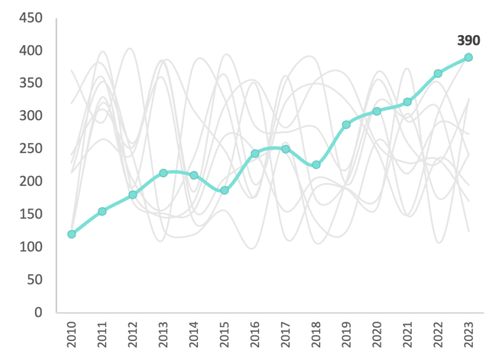
WRONG CHART TYPE

Using pie charts for time series data, line charts for categorical comparisons, etc.



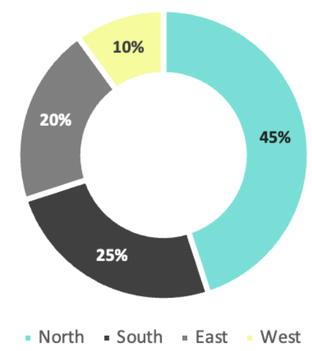
LACK OF FOCUS

Showing too many series or values, with no clear pattern or purpose



NOISE & DISTRACTIONS

Using unnecessary effects like 3D formats or background images



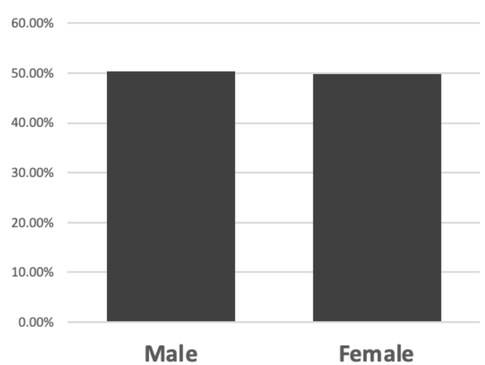
FORM OVER FUNCTION

Prioritizing design and visual aesthetics over clarity

Office	Team	Employee Name
Boston	Sales	Lisa T.
Boston	Sales	Phil B.
Boston	Sales	Marc G.
Boston	IT	Andrea T.
Boston	IT	Lauren W.
Boston	Product	Bill B.
New York	IT	Wendy L.
New York	IT	Gina C.
New York	Product	John R.
New York	Product	Darren S.
Miami	Sales	Vicky A.
Miami	IT	Dennis M.

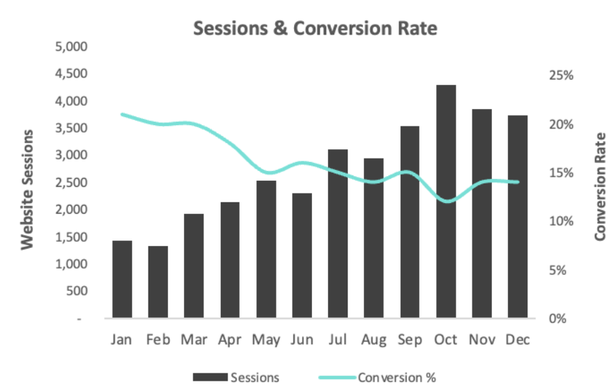
MISLEADING NARRATIVE

Exaggerating trends to make them appear more significant



MISSING CONTEXT

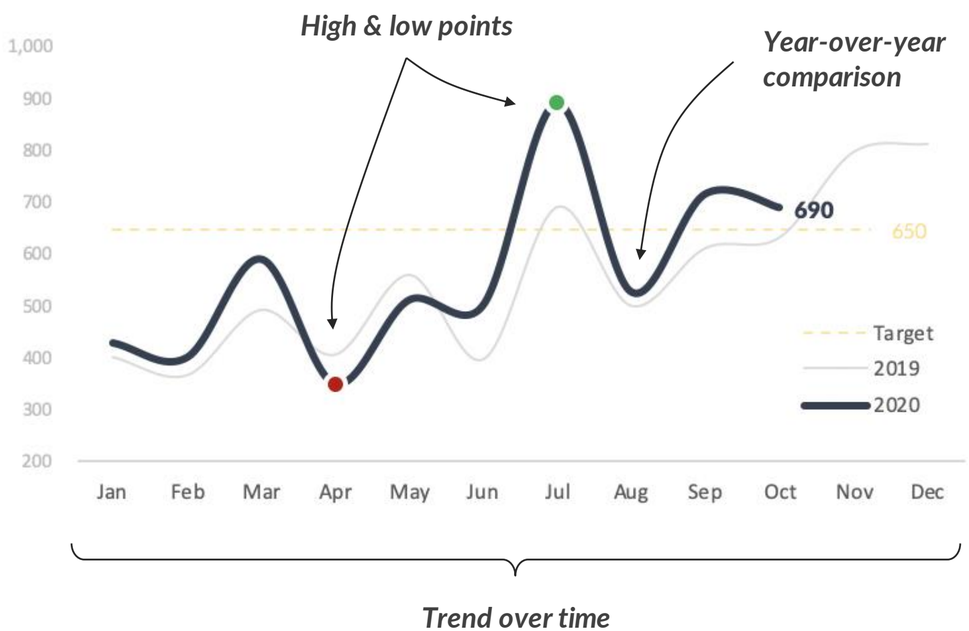
Failing to include series names, data labels, axis titles, etc.



*Copyright Maven Analytics, LLC

CONTEXT IS KEY

When you are presenting or visualizing data, remember that **context is key**; context is what gives numbers meaning, and helps users interpret them accurately



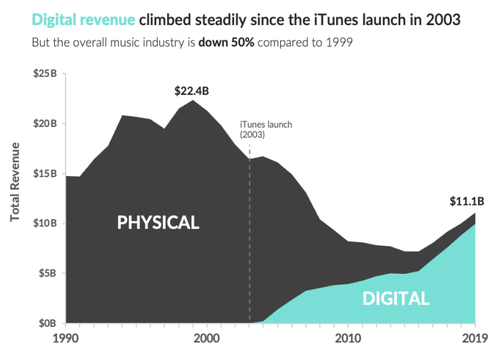
- ✓ There were **690 transactions** in Oct
- ✓ Sales are **down 4%** vs. last month, but **up 9%** year-over-year
- ✓ We **exceeded our target** in Sep/Oct after falling short in August
- ✓ We see an **upward trend** in 2020, with a significant **July peak**
- ✓ Based on seasonality, we can expect a **strong Nov/Dec**

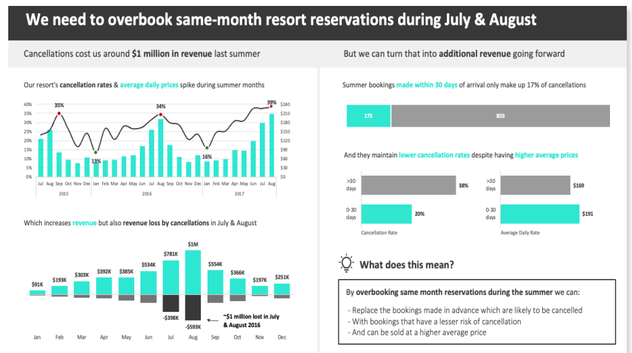
*Copyright Maven Analytics, LLC

STORYTELLING & DASHBOARD DESIGN

Data stories are commonly told through ad hoc reports or explanatory dashboards

- **Dashboards** are tools designed to consolidate data, track key metrics, and facilitate data-driven decisions





1. Define the purpose
2. Choose the right metrics
3. Present the data effectively
4. Eliminate clutter & noise
5. Use layout to focus attention
6. Tell a clear story

*Copyright Maven Analytics, LLC

CASE STUDY: DATA STORYTELLING

THE SITUATION

You've just been hired as a Business Intelligence Analyst for **Maven Toys**, a nationwide chain of toy stores.

THE ASSIGNMENT

Your assignment is to **design an explanatory dashboard for Regional Sales Managers**, to help them understand how their regions are performing.

They review the dashboard monthly, and need information on revenue trends, product-level performance, and lost revenue due to inventory shortages.

THE OBJECTIVES

1. Define the purpose
2. Choose the right metrics
3. Present the data effectively
4. Eliminate clutter & noise
5. Use layout to focus attention
6. Tell a clear story

MAVEN TOYS

*Copyright Maven Analytics, LLC

STEP 1: DEFINE THE PURPOSE

1. Who is your audience?
   - ***Regional sales managers***

2. What are their business goals and objectives?
   - ***Increase revenue, minimize revenue loss due to lack of inventory***

3. What specific questions do they need answers to?
   - ***How is revenue trending? Which products are performing best/worst? Are products out of stock?***

4. How often will the dashboard be reviewed?
   - ***Once a month***

*Copyright Maven Analytics, LLC

STEP 2: CHOOSE THE RIGHT METRICS

Business Goal:	Metric:	Level of Detail:
Increase revenue	• **Total Revenue**	• Month, Region, Category, Store & Product
	• **MoM Revenue Change**	• Store & Product
	• **MoM % Revenue Change**	• Region, Category
	• **YoY % Revenue Change**	• Region
Minimize loss from lack of inventory	• **Stock on Hand**	• Store & Product
	• **Estimated Monthly Revenue Loss**	• Region

*Copyright Maven Analytics, LLC

STEP 3: PRESENT THE DATA EFFECTIVELY

Metric & Level of Detail:		Visual Type:
• Total Revenue by Month	→	**Line Chart**
• Total Revenue by Region	→	**KPI Card**
• Total Revenue by Category	→	**Bar Chart**
• Total Revenue by Product	→	**Table**
• MoM Revenue Change by Product	→	**Table**
• MoM % Revenue Change by Region	→	**KPI Card**
• MoM % Revenue Change by Category	→	**Bar Chart**
• YoY % Revenue Change by Region	→	**KPI Card**
• Stock on Hand by Store-Product	→	**Table**
• Estimated Monthly Revenue Loss by Region	→	**KPI Card**

*Copyright Maven Analytics, LLC

STEP 3: PRESENT THE DATA EFFECTIVELY

MAVEN TOYS

REGIONAL SALES DASHBOARD

September 2021

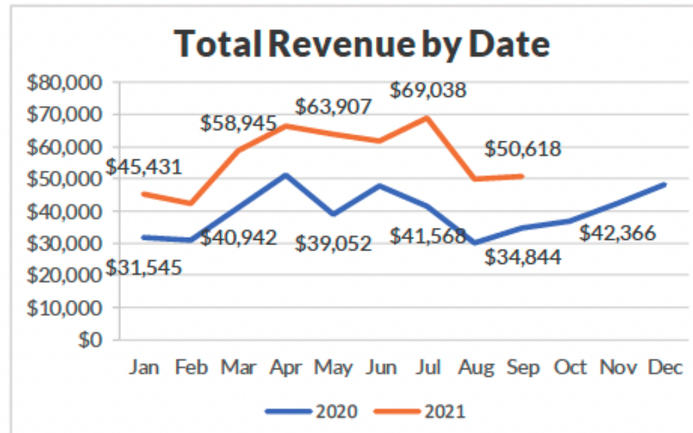
Region: **New York**

$50,618
Total Monthly Revenue

1.6%
M-o-M Revenue % Change

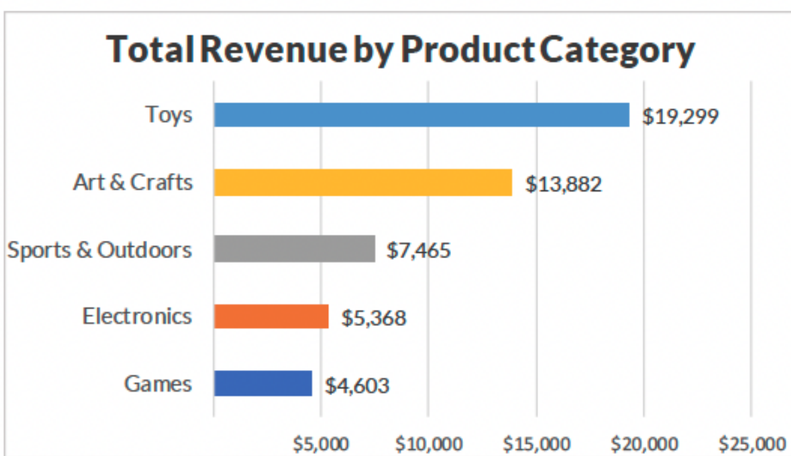
$1,640
Potential Monthly Revenue Loss

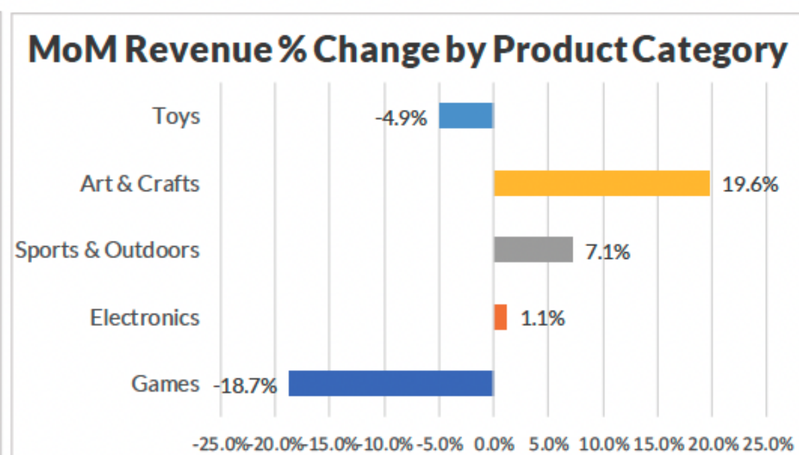
45.3%
Y-o-Y Revenue % Change



Products with 0 Stock by Store

Store Name	Product Name	Stock
JFK Airport	Gamer Headphones	0
JFK Airport	Hot Wheels 5-Pack	0
Times Square	Dino Egg	0
Times Square	Playfoam	0





Top 5 Products by MoM Revenue Change

Category	Product	Revenue	Δ Revenue
Art & Crafts	Playfoam	$3,352	$2,011
Toys	Dinosaur Figures	$2,893	$989
Games	Monopoly	$900	$740
Art & Crafts	Magic Sand	$4,589	$688
Art & Crafts	Barrel O' Slime	$1,357	$551

Bottom 5 Products by MoM Revenue Change

Category	Product	Revenue	Δ Revenue
Games	Rubik's Cube	$640	-$1,359
Toys	Lego Bricks	$9,318	-$800
Toys	Mr. Potatohead	$460	-$719
Art & Crafts	Etch A Sketch	$882	-$525
Games	Glass Marbles	$989	-$517

*Copyright Maven Analytics, LLC

STEP 4: ELIMINATE CLUTTER & NOISE

MAVEN TOYS

REGIONAL SALES DASHBOARD

September 2021

Region: **New York**

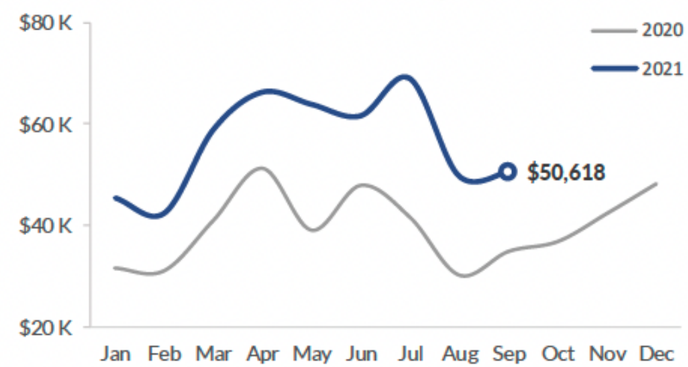
$50,618
Total Monthly Revenue

1.6%
M-o-M Revenue % Change

$1,640
Potential Monthly Revenue Loss

45.3%
Y-o-Y Revenue % Change

Total Revenue by Date



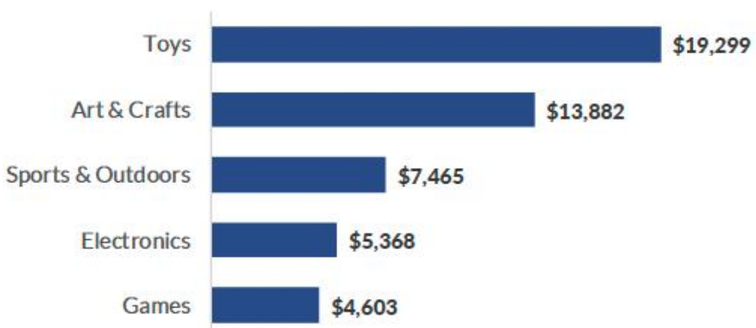
Products with 0 Stock by Store

Store Name	Product Name	Stock
JFK Airport	Gamer Headphones	0
JFK Airport	Hot Wheels 5-Pack	0
Times Square	Dino Egg	0
Times Square	Playfoam	0

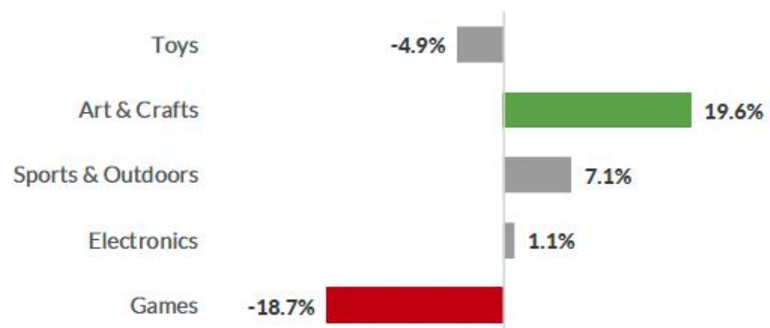
Top 5 Products by MoM Revenue Change

Category	Product	Revenue	Δ Revenue
Art & Crafts	Playfoam	$3,352	$2,011
Toys	Dinosaur Figures	$2,893	$989
Games	Monopoly	$900	$740
Art & Crafts	Magic Sand	$4,589	$688
Art & Crafts	Barrel O' Slime	$1,357	$551

Total Revenue by Product Category



MoM Revenue % Change by Product Category



Bottom 5 Products by MoM Revenue Change

Category	Product	Revenue	Δ Revenue
Games	Rubik's Cube	$640	-$1,359
Toys	Lego Bricks	$9,318	-$800
Toys	Mr. Potatohead	$460	-$719
Art & Crafts	Etch A Sketch	$882	-$525
Games	Glass Marbles	$989	-$517

*Copyright Maven Analytics, LLC

STEP 5: USE LAYOUT TO FOCUS ATTENTION

REGIONAL SALES DASHBOARD

September 2021

Region: **New York**

MAVEN TOYS

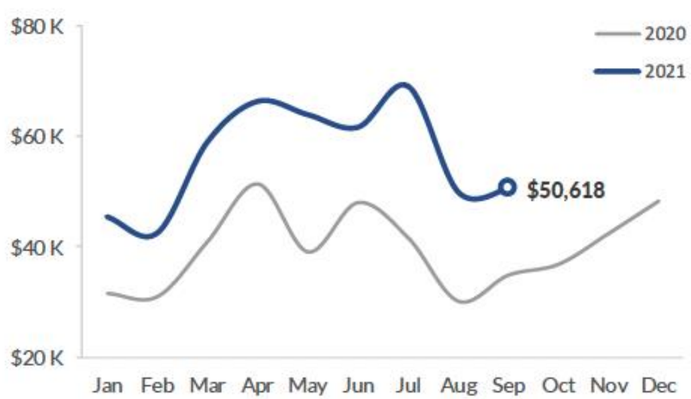
Total Revenue & Revenue % Changes

$50,618
Total Monthly Revenue

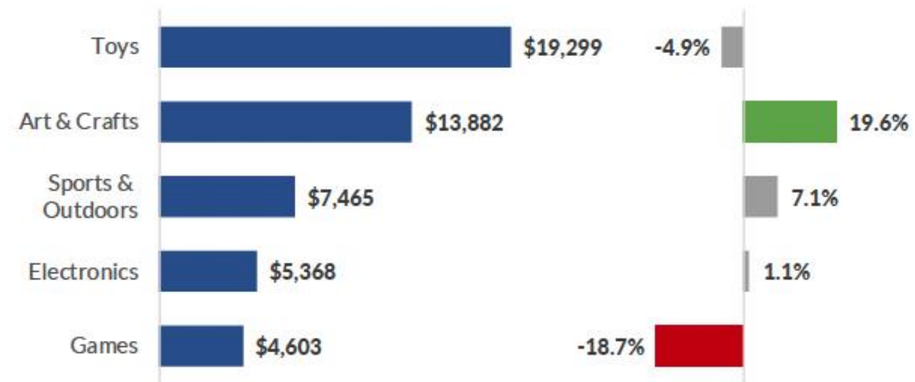
1.6%
M-o-M

45.3%
Y-o-Y

Total Revenue by Date



Revenue & MoM Revenue % Change by Product Category



Products with 0 Stock

Store Name	Product Name	Stock
JFK Airport	Gamer Headphones	0
JFK Airport	Hot Wheels 5-Pack	0
Times Square	Dino Egg	0
Times Square	Playfoam	0

$1,640
Potential Monthly Revenue **Loss**

Top & Bottom 5 Products by MoM Revenue Change

Category	Product	Revenue	Δ Revenue
Art & Crafts	Playfoam	$3,352	$2,011
Toys	Dinosaur Figures	$2,893	$989
Games	Monopoly	$900	$740
Art & Crafts	Magic Sand	$4,589	$688
Art & Crafts	Barrel O' Slime	$1,357	$551

Category	Product	Revenue	Δ Revenue
Games	Rubik's Cube	$640	-$1,359
Toys	Lego Bricks	$9,318	-$800
Toys	Mr. Potatohead	$460	-$719
Art & Crafts	Etch A Sketch	$882	-$525
Games	Glass Marbles	$989	-$517

*Copyright Maven Analytics, LLC

STEP 6: TELL A STORY

REGIONAL SALES DASHBOARD

How did the region perform in **September 2021**?

Region: **New York**

MAVEN TOYS

This is how much **revenue** we generated...

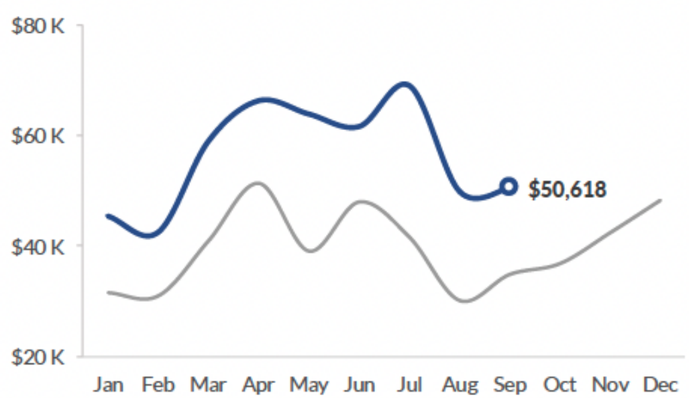
$50,618

Total Monthly Revenue

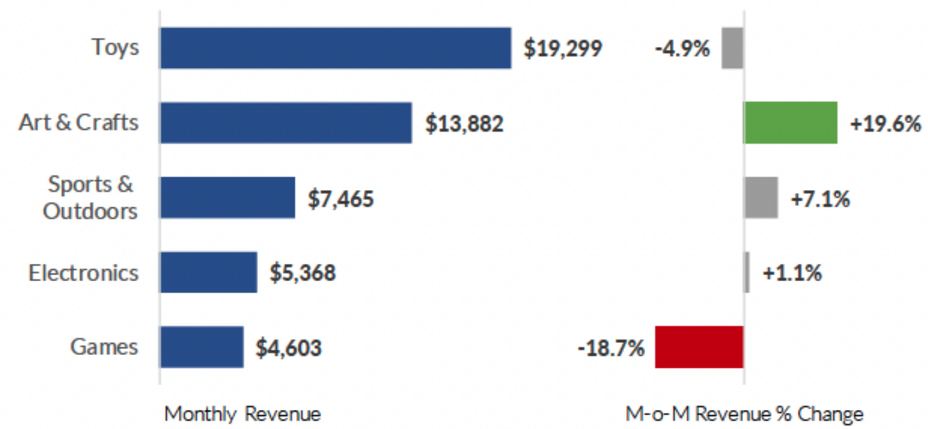
+1.6% M-o-M

+45.3% Y-o-Y

... and the trend in **2021** vs **2020**



This is the revenue split by **product category** this month...



We are **losing sales** here...

Store Name	Product Name	Stock
JFK Airport	Gamer Headphones	0
JFK Airport	Hot Wheels 5-Pack	0
Times Square	Dino Egg	0
Times Square	Playfoam	0

$1,640

Potential Monthly Revenue **Loss**

... where these **5 products** drove the **increases** & **decreases** in revenue

Category	Product	Revenue	Δ Revenue
Art & Crafts	Playfoam	$3,352	+$2,011
Toys	Dinosaur Figures	$2,893	+$989
Games	Monopoly	$900	+$740
Art & Crafts	Magic Sand	$4,589	+$688
Art & Crafts	Barrel O' Slime	$1,357	+$551

Category	Product	Revenue	Δ Revenue
Games	Rubik's Cube	$640	-$1,359
Toys	Lego Bricks	$9,318	-$800
Toys	Mr. Potatohead	$460	-$719
Art & Crafts	Etch A Sketch	$882	-$525
Games	Glass Marbles	$989	-$517

*Copyright Maven Analytics, LLC

KEY TAKEAWAYS

- **Data visualization** is the graphical representation of data and information
- **Data storytelling** is about creating narratives to communicate insights, influence decisions, or support specific data-driven recommendations
- Always consider the **three key questions**: What data are you working with? What do you want to communicate? Who is your audience and what do they need?
- **Common mistakes** include choosing the wrong chart, failing to focus, adding clutter and noise, and prioritizing form over function
- **Context** gives numbers meaning, and helps users interpret them accurately

*Copyright Maven Analytics, LLC

DATA LITERACY & AI

*Copyright Maven Analytics, LLC

WELCOME TO THE AGE OF AI

Artificial intelligence is quickly becoming mainstream technology, and tools like **ChatGPT**, **Google Gemini**, **Microsoft Copilot** and others are revolutionizing the way people work

Why you should embrace AI:

- ✓ Improve your performance & efficiency
- ✓ Automate routine, low-value tasks
- ✓ Spend more time on strategic work
- ✓ Learn new topics or generate new ideas
- ✓ Stay on the leading edge

Generative AI tools are more intuitive and approachable than ever; **you don't need technical skills or complex tools to harness the power of artificial intelligence**

*Copyright Maven Analytics, LLC

COMMON DATA USE CASES

THIS IS WHERE AI SHINES

While there are many ways that AI can be used to support data literacy and analytics, these are a few of the **most common and powerful use cases** we've explored:

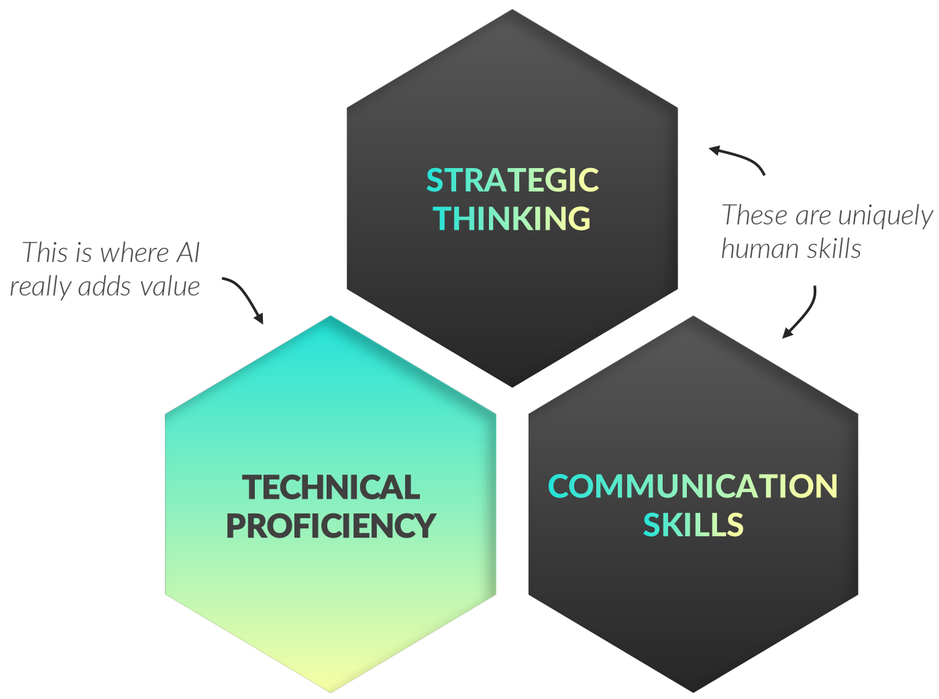
- Data profiling & quality assurance
- Basic exploratory analysis
- Data-driven insights & recommendations
- Step-by-step guides & tutorials
- Data analysis & visualization tips
- Conversational learning
- Manual task automation
- Technical research & discovery
- Formula & code generation
- Troubleshooting & debugging

*Copyright Maven Analytics, LLC

AI + HUMAN SKILLS

With the rise of AI, will data literacy skills become obsolete?

- **Data literacy is more important than ever**. Human expertise is critical for providing accurate data, writing effective prompts, checking and interpreting results, and providing intuition, context and creativity.



We call this the "**Analytics Trifecta**"

Since AI can fill many technical gaps, people with exceptional strategic thinking and communication skills will become more valuable than ever to organizations

AI cannot replace foundational skills; the real magic happens when you combine your human strengths with a technical foundation enhanced by AI

*Copyright Maven Analytics, LLC

AI + HUMAN SKILLS

AI STRENGTHS

- Automation
- Speed & Efficiency
- Technical Support
- Integration

+

HUMAN STRENGTHS

- Creativity
- Strategic Thinking
- Context & Intuition
- Trust & Empathy

*Copyright Maven Analytics, LLC

WARNING: COMMON PITFALLS

AI tools are known to **"hallucinate" facts** with total confidence

- *Remember that YOU are ultimately responsible for verifying the accuracy of model outputs*

Solutions provided **may be suboptimal, biased or entirely incorrect**

- *AI tools don't guarantee accuracy or efficiency, so be cautious integrating solutions into your work*

These tools are broad and often **lack specific domain knowledge**

- *They may not understand specific business context or the "why" behind the responses they produce*

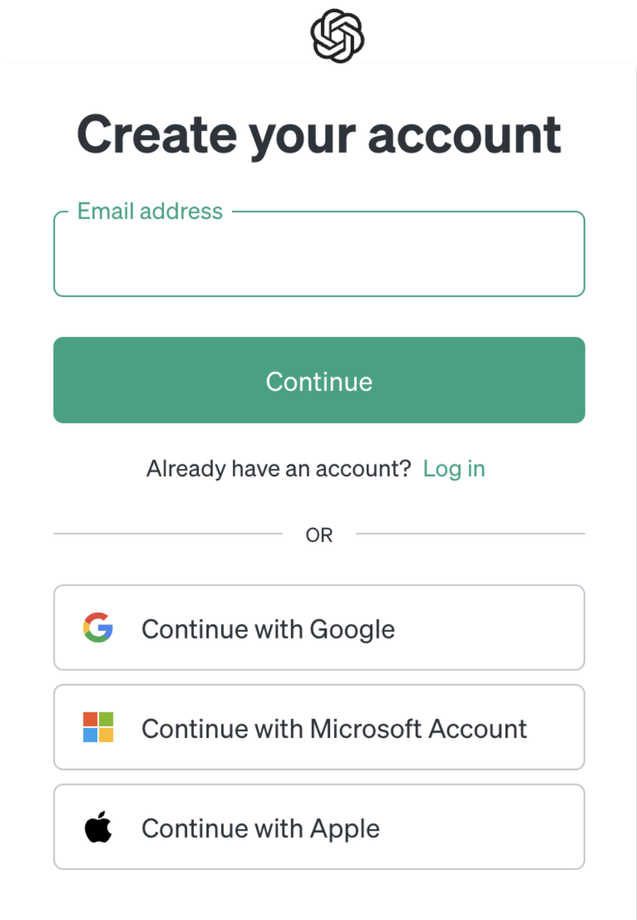
AI is NOT capable of **common sense** or **human judgement**

- *Models need specific, objective inputs, and may miss critical context that may seem obvious to humans*

*Copyright Maven Analytics, LLC

ACCESSING CHATGPT

You can sign up and access ChatGPT for free at **chat.openai.com**



- OpenAI currently allows anyone to sign up and use ChatGPT with a 100% free account
- **NOTE**: You can upgrade to **ChatGPT Plus** for $20/month, which includes access to faster, more powerful models and plug-ins like DALL-E (for image generation) and Data Analysis

Learn more about OpenAI, ChatGPT and the latest artificial intelligence news and developments on the OpenAI blog: **https://openai.com/blog**

*Copyright Maven Analytics, LLC

PROMPT ENGINEERING

Prompts are questions or instructions that you type to interact with models like ChatGPT

- **Prompt engineering** is the practice of constructing prompts to generate effective and accurate responses

This is a ***prompt***

Can you please explain how the OFFSET function works in Excel?

This is a ***response***

Certainly! The OFFSET function in Excel is a powerful tool that allows you to reference a range of cells based on a starting point and specified dimensions. It is commonly used in formulas to dynamically retrieve data or create dynamic ranges.

The syntax for the OFFSET function is as follows:

arduino Copy code

```
OFFSET(reference, rows, cols, [height], [width])
```

- `reference`: This is the starting point or cell from which you want to offset. It can be a cell reference or a named range.
- `rows`: This is the number of rows you want to move (up or down) from the reference cell. Positive values indicate moving down, while negative values indicate moving up.
- `cols`: This is the number of columns you want to move (left or right) from the reference cell. Positive values indicate moving to the right, while negative values indicate moving to the left.
- `[height]` (optional): This parameter specifies the height or number of rows you want the resulting range to have. If omitted, the function returns a reference to a single cell.
- `[width]` (optional): This parameter specifies the width or number of columns you want the resulting range to have. If omitted, the function returns a reference to a single cell.

Tips & Best Practices:

- ✓ Be as clear and specific as possible
- ✓ Provide context and examples
- ✓ Establish roles for specific use cases
- ✓ Set the tone or level of complexity
- ✓ Understand model limitations

You likely won't nail the "perfect" prompt right away; make sure to **refine and iterate** based on the quality of the responses

*Copyright Maven Analytics, LLC

TIP: BE CLEAR & SPECIFIC

ORIGINAL PROMPT

How do you calculate year over year growth?

Year over year (YoY) growth is calculated by comparing the performance of a certain metric or variable for two consecutive years. It measures the percentage change in that metric from one year to the next. To calculate YoY growth, follow these steps:

1. Determine the initial value: Identify the starting point for the metric you want to measure. This could be the revenue, sales, profit, or any other relevant figure for a particular year. Let's call this value "V1."
2. Determine the final value: Identify the value of the same metric for the subsequent year. Let's call this value "V2."
3. Calculate the difference: Subtract the initial value (V1) from the final value (V2). This gives you the difference or change in the metric from the first year to the second year. Difference = V2 - V1
4. Calculate the YoY growth rate: Divide the difference calculated in step 3 by the initial value (V1). Then, multiply the result by 100 to express it as a percentage. YoY growth rate = (Difference / V1) * 100

The resulting YoY growth rate represents the percentage change in the metric from the first year to the second year. A positive growth rate indicates an increase, while a negative growth rate indicates a decrease.

For example, let's say a company had revenues of $1 million in 2021 (V1) and $1.2 million in 2022 (V2). The YoY growth rate would be calculated as follows:

VS

BETTER PROMPT

Can you give me an Excel formula to calculate year over year growth?

Data for this year lives in cells C2:C100, and data for last year lives in cells B2:100.

Certainly! You can use the following Excel formula to calculate the year-over-year growth percentage:

excel Copy code

```
=((C2-B2)/B2)*100
```

Assuming you want to calculate the growth percentage for the first row (row 2) and drag the formula down to calculate for the rest of the rows, you can place the formula in cell D2 and then drag it down from cell D2 to D100 to apply the formula to the entire range.

This formula subtracts the value from the previous year (B2) from the value of the current year (C2), divides the result by the value from the previous year (B2), and then multiplies it by 100 to express the growth as a percentage.

When generating code, include details like **variable names**, **data types**, **cell ranges**, etc.

*Copyright Maven Analytics, LLC

TIP: PROVIDE CONTEXT

ORIGINAL PROMPT

What are common supply chain KPIs?

Key Performance Indicators (KPIs) in the supply chain are metrics used to assess the efficiency, effectiveness, and overall performance of the supply chain processes. While there can be variations depending on the specific industry and organization, here are some common supply chain KPIs:

1. On-time Delivery: Measures the percentage of orders delivered to customers within the promised or agreed-upon timeframe. It indicates how well the supply chain meets customer expectations and ensures timely fulfillment.
2. Order Cycle Time: Measures the time taken from when an order is placed until it is delivered to the customer. It assesses the speed and efficiency of order processing, including order entry, picking, packing, and shipping.
3. Perfect Order Rate: Measures the percentage of orders that are fulfilled without errors, including accuracy in product selection, packaging, documentation, and delivery. It reflects the ability to meet customer requirements accurately and completely.
4. Inventory Turnover: Calculates the number of times inventory is sold or consumed within a specific period. It indicates the efficiency of inventory management and how quickly goods are moving through the supply chain.
5. Fill Rate: Measures the percentage of customer orders that are filled completely from available stock without backorders or substitutions. It reflects the ability to meet customer demand without stockouts or delays.
6. Order Accuracy: Measures the percentage of orders that are error-free in terms of

VS

BETTER PROMPT

I'm a Data Analyst working on supply chain analytics for a multinational distribution company. What KPIs should I be analyzing? For each metric, please include a formula showing how it is calculated.

For example, we would calculate conversion rate as follows: Conversion rate = website orders / website sessions

As a Data Analyst working on supply chain analytics for a multinational distribution company, there are several key performance indicators (KPIs) that you should consider analyzing. Here are some relevant metrics along with their formulas:

1. On-time Delivery Rate:
   Formula: (Number of orders delivered on time / Total number of orders) x 100
2. Order Cycle Time:
   Formula: (Order delivery date - Order placement date) / Number of orders
3. Perfect Order Rate:
   Formula: (Number of perfect orders / Total number of orders) x 100
4. Inventory Turnover:
   Formula: Cost of Goods Sold (COGS) / Average Inventory Value
   COGS can be obtained from financial statements, and average inventory value can be calculated as (Beginning Inventory + Ending Inventory) / 2.
5. Fill Rate:
   Formula: (Number of complete orders / Total number of orders) x 100

*Copyright Maven Analytics, LLC

TIP: ESTABLISH ROLES

ORIGINAL PROMPT

What's the best way to clean data?

The best way to clean data depends on the specific characteristics of the dataset and the goals of your analysis. However, here are some general best practices for data cleaning:

1. Understand your data: Before cleaning the data, it's essential to thoroughly understand its structure, variables, and any potential issues or anomalies. This will help you make informed decisions during the cleaning process.
2. Define a data cleaning plan: Create a clear plan outlining the steps you'll take to clean the data. This plan should include identifying missing values, handling outliers, addressing inconsistencies, and resolving any other known issues.
3. Handle missing values: Decide on the most appropriate method for dealing with missing values. This can include removing rows or columns with a high proportion of missing values or imputing missing values by filling them in with reasonable estimates (e.g., using mean, median, or machine learning algorithms).
4. Address outliers: Identify outliers, which are data points significantly different from others in the dataset. Determine whether these outliers are valid data or errors. Depending on the situation, you can either correct the outliers, remove them if they are errors, or analyze them separately if they represent important information.
5. Check for duplicates: Identify and remove any duplicate records in your dataset. Duplicates can skew analysis results and lead to incorrect conclusions.
6. Standardize data formats: Ensure consistency by standardizing the formats of variables, such as dates, addresses, or units of measurement. This helps avoid confusion and facilitates accurate analysis.

VS

BETTER PROMPT

I'm a new employee for a retail business, and I've been tasked with cleaning and preparing some data for analysis.

Please play the role of a Data QA Engineer and provide a step-by-step plan for me to follow.

As a Data QA Engineer, I can guide you through the process of cleaning and preparing a raw dataset for analysis. Here's a step-by-step plan you can follow:

Step 1: Understand the Data

- Review the dataset documentation or data dictionary if available to understand the meaning of each variable.
- Identify the goals of the analysis and the specific questions you need to answer with the data.
- Determine the data types of each variable (numeric, categorical, datetime, etc.).

Step 2: Assess Data Quality

- Check for missing values: Identify variables with missing data and decide how to handle them. You can either impute missing values, remove rows with missing values, or create a separate category for missing data.
- Identify duplicate records: Look for duplicate rows in the dataset and decide whether to keep or remove them based on your analysis goals.
- Validate data range and outliers: Examine the minimum and maximum values of numeric variables to ensure they fall within logical ranges. Identify any potential outliers that may require further investigation.

*Copyright Maven Analytics, LLC

TIP: SET THE TONE

ORIGINAL PROMPT

Explain regression analysis

Regression analysis is a statistical method used to examine the relationship between a dependent variable and one or more independent variables. It aims to model and predict the value of the dependent variable based on the values of the independent variables. Regression analysis is widely used in various fields, including economics, social sciences, finance, and healthcare, to understand the factors that influence an outcome or to make predictions.

The dependent variable, also known as the response variable or outcome variable, is the variable of interest that we want to predict or explain. The independent variables, also called predictor variables or explanatory variables, are the variables that we believe have an impact on the dependent variable. The regression analysis determines the nature and strength of the relationship between the dependent variable and the independent variables.

The most common form of regression analysis is linear regression, where the relationship between the variables is assumed to be linear. In linear regression, the relationship is represented by a straight line equation of the form:

Y = b0 + b1*X1* + *b2*X2 + ... + bn*Xn

In this equation, Y is the dependent variable, b0 is the intercept (the value of Y when all the independent variables are zero), b1, b2, ..., bn are the coefficients representing the impact of each independent variable (X1, X2, ..., Xn) on the dependent variable. The coefficients indicate the change in the dependent variable for a one-unit change in the corresponding

VS

BETTER PROMPT

Explain regression analysis in two sentences to a C-Level executive at a marketing agency

Regression analysis is a statistical technique that helps identify the relationship between a dependent variable (such as sales or customer satisfaction) and independent variables (such as advertising spend or website traffic), allowing us to understand how changes in one variable affect the other. By analyzing this relationship, we can make data-driven decisions and optimize marketing strategies to maximize business outcomes.

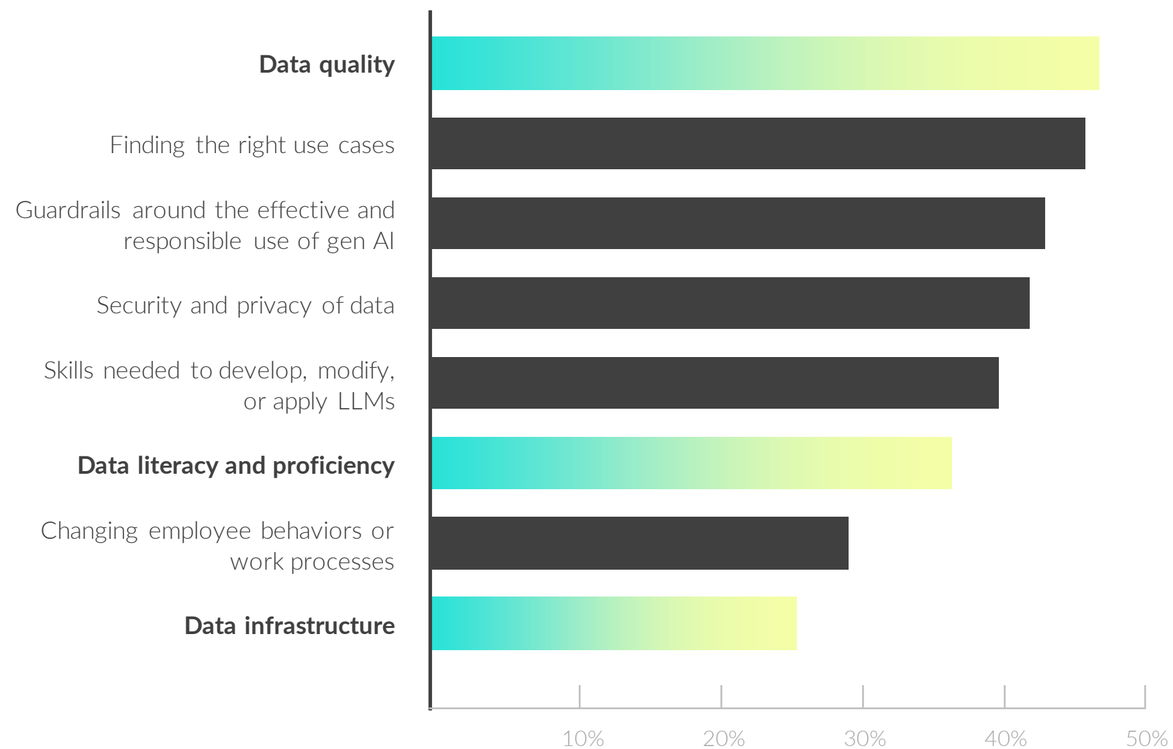
Use prompts like "**explain it to me like you are speaking to a child**" to get simplified descriptions of complex or highly technical topics

*Copyright Maven Analytics, LLC

CHALLENGES INTEGRATING GENERATIVE AI

In 2024 AWS surveyed **300+ data & AI leaders** to understand how their roles and priorities have evolved, and how they are adapting to new technologies like generative AI

What's the **biggest challenge** for your organization in realizing the potential of generative AI?*



Data quality, **data literacy & proficiency** and **data infrastructure** all emerged as some of the biggest challenges for realizing AI's potential

Companies will struggle to derive value from AI without a strong data foundation which includes strategy, architecture, governance and literacy

* https://aws.amazon.com/data/cdo-report/

*Copyright Maven Analytics, LLC

KEY TAKEAWAYS

- AI is revolutionizing the way we work, but **cannot replace foundational skills**
- Human strengths like **strategic thinking** and **creativity** are as important as ever
- **Prompt engineering** is the key to generating relevant, accurate responses
- **AI isn't perfect**, and requires human expertise and intuition to ensure accuracy
- Companies will struggle to derive value from AI without a **strong data ecosystem**

*Copyright Maven Analytics, LLC

KEY TAKEAWAYS & NEXT STEPS

*Copyright Maven Analytics, LLC

KEY TAKEAWAYS

- Data literacy is about helping everyday people build the skills and the confidence they need to make **smart, data-driven decisions**
- A strong data ecosystem consists of **strategy**, **architecture** and **governance**, with **democratized access** to data, tools and learning resources
- **Data analysis** is about identifying actionable insights that can help inform decisions and impact real-world outcomes
- **Data visualization** and **storytelling** help us quickly make sense of complex information using stories, patterns, and visual cues
- **AI is revolutionizing how we work**, but cannot replace foundational skills

*Copyright Maven Analytics, LLC

NEXT STEPS

Congratulations, you completed the course!

- We hope that you've enjoyed the journey, and that we were able to provide a solid foundation and a clear roadmap to help you interpret, manage, analyze and communicate with data
- As you continue along your path, take time to assess your strengths, embrace lifelong learning, seek opportunities to apply your skills, and build a reputation as an advocate for data literacy

Assess yourself to identify skills gaps

Embrace a **lifelong learning** mentality

Seek opportunities to use data in your role

Become an advocate for data literacy

MAVEN ANALYTICS

If you'd like to continue learning about data analytics, AI and machine learning, join the **1M+ students who have trusted Maven Analytics** to launch or accelerate their careers

*Copyright Maven Analytics, LLC