Foundations: Data, Data, Everywhere

Introducing data analytics

The six steps of the data analysis process

ask, prepare, process, analyze, share and **act**.

1. **Ask**: Business Challenge/Objective/Question
2. **Prepare**: Data generation, collection, storage, and data management
3. **Process**: Data cleaning/data integrity
4. **Analyze**: Data exploration, visualization, and analysis
5. **Share**: Communicating and interpreting results
6. **Act**: Putting your insights to work to solve the problem

The analysts **asked** questions to define both the issue to be solved and what would equal a successful result. Next, they **prepared** by building a timeline and collecting data with employee surveys, which should be inclusive. They **processed** the data by cleaning it to make sure it was complete, correct, relevant, and free of errors and outliers. They **analyzed** the clean employee survey data. Then the analysts **shared** their findings and recommendations with team leaders. Afterward, leadership **acted** on the results and focused on improving key areas.

Data + business knowledge = mystery solved For data analysts, just trusting our gut instinct without any data to back it up can be a problem. Your business knowledge and experience may help you understand problems intuitively. But, unlike gut instinct, it will give you more than just a feeling to go on. Blending facts and data with your business knowledge will be a common part of your process. The key is figuring out the exact mix of data and business knowledge for each particular project. Try asking yourself these questions about a project:

- What kind of results are needed?
- Who will be informed?
- Am I answering the question being asked?
- How quickly does a decision need to be made?

Data analysis life cycle

The process of going from data to decision.

EMC's data analysis life cycle

EMC Corporation's data analytics life cycle is cyclical with six steps:

1. Discovery
2. Pre-processing data

3. Model planning
4. Model building
5. Communicate results
6. Operationalize

SAS' iterative life cycle

An iterative life cycle was created by a company called **SAS**, a leading data analytics solutions provider.

1. Ask
2. Prepare
3. Explore
4. Model
5. Implement
6. Act
7. Evaluate

Project-based data analytics life cycle

A project-based data analytics life cycle has five simple steps:

1. Identifying the problem
2. Designing data requirements
3. Pre-processing data
4. Data analysis
5. Data visualizing

Big data analytics life cycle

In the book, *Big Data Fundamentals: Concepts, Drivers & Techniques*, suggests phases divided into nine steps:

1. Business case evaluation
2. Data identification
3. Data acquisition and filtering
4. Data extraction
5. Data validation and cleaning
6. Data aggregation and representation
7. Data analysis
8. Data visualization
9. Utilization of analysis results

Data life cycle based on research

One final data life cycle informed by Harvard University research has eight phases:

1. Generation
2. Collection
3. Processing

4. Storage
5. Management
6. Analysis
7. Visualization
8. Interpretation

Thinking analytically

Data analyst skills

Analytical skills: Qualities and characteristics associated with solving problems using facts.

Five key data analyst(/Data-driven decision-making) skills:

1. **Curiosity** A desire to know more about something, asking the right questions.
2. **Understanding context** Understanding where information fits into the big picture and how you group things into categories.
3. **Having a technical mindset** The ability to break things down into smaller steps or pieces and work with them in an orderly and logical way.
4. **Data design** How you organize data and information. Typically has to do with an actual database.
5. **Data strategy** The analytical skill that involves managing the people, processes and tools used in data analysis.

All this part can help you tap into the data-driven decision making, which can lead to successful outcomes.

Analytical thinking

Identifying and defining a problem and then solving it by using data in an organized, step-by-step way.

Five key aspects of analytical thinking:

1. **Visualization** The graphical representation of information. e.g.: graphs, charts, maps, design elements. Can help data analysts understand and explain information more effectively.
2. **Strategy** Help analyst see what they what to chieve with the data and how they can get there. Also help impove the quality and usefulness of the data we collected.
3. **Problem-orientation** Analysts use a problem-oriented approach to data analysis in order to identify, describe, and solve problems.
4. **Correlation** Like a relationship. *"Correlation does not equal to causation.""*
5. **Big-picture and detail-oriented thinking** Like a jigsaw puzzle. Help you zoom out as well as considers the specifics.

Exploring core analytical skills Thinking analytically, critically and creatively.

Some questions data analysts ask when hunting for a solution:

- What is the root cause of the problem? (Root cause: the reason why a problem occurs.) Five "Why"s to approach the root cause: ask "why" five times to reveal the root cause.
- Where are the gaps in our process? The general approach to gap analysis is understanding where you are now comparing to where you want to be. Then you can identify the gaps that exists between the current and future state and determine how to bridge them.
- What did we not consider before? Great way to think about what information or procedure might be missing from a process.
- ...

Exploring the wonderful world of data

Data analysis tools: spreadsheets, databases, query languages, visualization software.

Data life cycle

Plan, capture, manage, analyze, archive, destory

1. Planning Decide what kind of data it needs, how it will be managed, who will be responsible for it, and the optimal outcomes.
2. Capture Gathering data from various sources and bringing it into the organization.
3. Manage How data is cared for, how and where it’s stored, the tools used to keep it safe and secure, and the actions taken to make sure it’s maintained properly. Very important to data cleansing.
4. Analyze Using data to solve problems, make smart decisions and support business goals.
5. Archive Storing data in a place where it's still available but may not be used again.
6. Destory Use secure data erase software. Paper files would be shredded. Protecting private information and private data.

Question asking type:

- Plan - What plans and decisions do you need to make? What data do you need to answer your question?
- Capture - Where does your data come from? How will you get it?
- Manage - How will you store your data? What should it be used for, and how do you keep this data secure and protected?
- Analyze - How will the company analyze the data? What tools should they use?
- Archive - What should they do with their data when it gets old? How do they know when it's time?
- Destroy - Should they ever dispose of any data? If so, when and how?

Data analysis steps

Ask, prepare, process, analyze, share, act

1. Ask Define the problem to be solved and make sure that we fully understand stakeholder expectations.

- ***Defining a problem***: look at the current state and identify how it's different from the ideal state. (Usually there is an obstacle we need to get rid of or something wrong that needs to be fixed.)
- ***Understanding stakeholder expectations***: first determine who the stakeholders are. (Various stakeholders all have in common: they help make decisions, influence actions and strategies, and have specific goals) Five "why"s are extremely helpful here.

2. Prepare Collect and store data they'll use for the upcoming analysis process.
3. Process Find and eliminate any errors and inaccuries. Means cleaning data, transforming it into a more useful format, combining dataset and removing outliers which are any data points that can skew the information. The process phase is all about getting the details right, so data analysts clean data by fixing typos, inconsistencies, and missing or inaccurate data.
4. Analyze Involves using tools to transform and organize that information so that you can draw useful conclusions, make predictions, and drive informed decision making. Spreadsheets, Structured query language(SQL, ["sequel"])
5. Share How to interpret results and share with others to help stakeholder make effective data-driven decisions. Visualization is the best friend. Using R language.
6. Act The business put all the insights to work in order to solve the business problem.

How t
process g

Ask

- Ask effective questions
- Define the problem
- Use structured thinking
- Communicate with others

Process

- Create and transform data
- Maintain data integrity
- Test data
- Clean data
- Verify and report on cleaning results

Data analyst tools

Spreadsheets, Query language for database, Visualization tools A career as a data analyst also involves using programming languages, like R and Python, which are used a lot for statistical analysis, visualization, and other data analysis.

Spreadsheets

A spreadsheets is a digital worksheet, which stores, organizes and sorts data. Also has useful features called formulas and functions. e.g. Microsoft Excel, Google sheets **Formula**: A set of instructions that performs a specific calculation using the data in the spreedsheet. **Function**: A preset command that automatically performs a specific process or task using the data in the spreadsheet.

Digital worksheets structure data in a meaningful way by letting you:

- Collect, store, organize, and sort information
- Identify patterns and piece the data together in a way that works for each specific data project
- Create excellent data visualizations, like graphs and charts.

Databases and query languages

A database is a collection of structured data stored in a computer system. The query language is a computer programming language that allows you to retrive and manipulate data from a database. e.g. Some popular SQL(Structured Query Language) programs include MySQL, Microsoft SQL Server, and BigQuery)

Query languages

- Allow analysts to isolate specific information from a database(s)
- Make it easier for you to learn and understand the requests made to databases
- Allow analysts to select, create, add, or download data from a database for analysis

Difference between spreadsheets and databases:

Spreadsheets	**Databases**
Software applications	Data stores - accessed using a query language (e.g. SQL)
Structure data in a row and column format	Structure data using rules and relationships
Organize information in cells	Organize information in complex collections
Provide access to a limited amount of data	Provide access to huge amounts of data
Manual data entry	Strict and consistent data entry

Spreadsheets	**Databases**
Generally one user at a time	Multiple users
Controlled by the user	Controlled by a database management system

Visualization tools

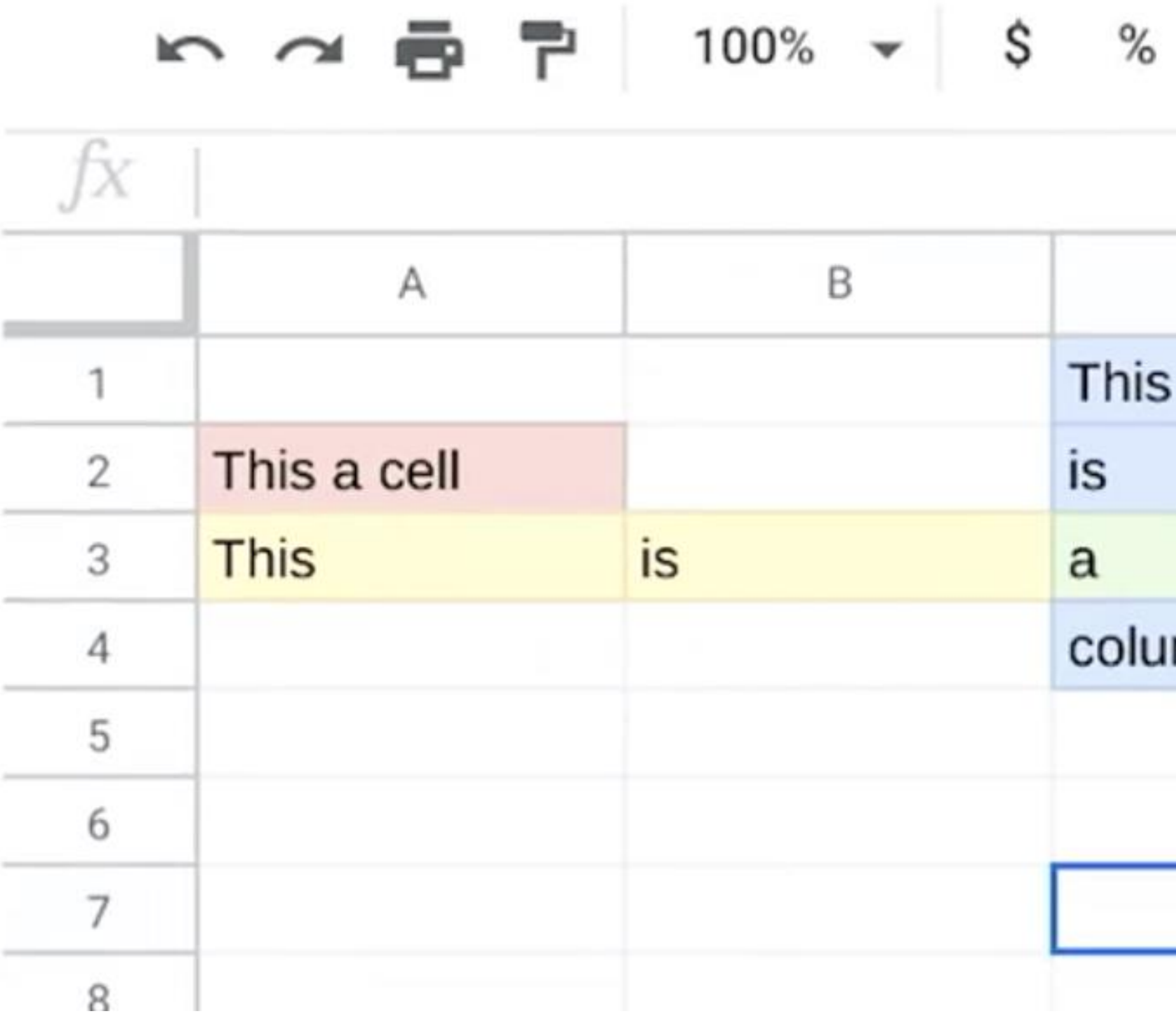
Visualization is the graphical representation of information. Data analysts use a number of visualization tools, like graphs, maps, tables, charts, and more. e.g. Tableau, Looker

These tools

- Turn complex numbers into a story that people can understand
- Help stakeholders come up with conclusions that lead to informed decisions and effective business strategies
- Have multiple features
    - **Tableau**'s simple drag-and-drop feature lets users create interactive graphs in dashboards and worksheets
    - **Looker** communicates directly with a database, allowing you to connect your data right to the visual tool you choose

Setting up a data toolbox

Mastering the spreadsheet basics



	A	B	
1			This
2	This a cell		is
3	This	is	a
4			colu
5			
6			
7			
8			

Columns are ordered by letter, while rows are ordered by number. So when you talk about a specific cell, you name it by combining the column letter and row numebr like D3. Format --> Text wrapping:

- Overflow: text can show flowing over the cell when it's too long
- Wrap: text will change line in the cell if too long. (Automatically change cell height in order to allow all of the text to fit inside.)
- Clip: text showing will be cutted if too long

Column labels are usually called attributes. Loated at row 1. **Attribute**: A characteristic or quality of data used to label a column in a table.

A row is called an observation. **Observation**: All of the attributes for something contained in a row of a data table.

Formula: A set of instructions used to perform a calculation using the data in a spreadsheet. Start with `=` , e.g. In a cell, type `=average(C2:C4)` then press enter.

More spreadsheet resources

Google Sheets Training and Help **Google Sheets Cheat Sheet** **Microsoft Excel for Windows Training**

Work with functions.

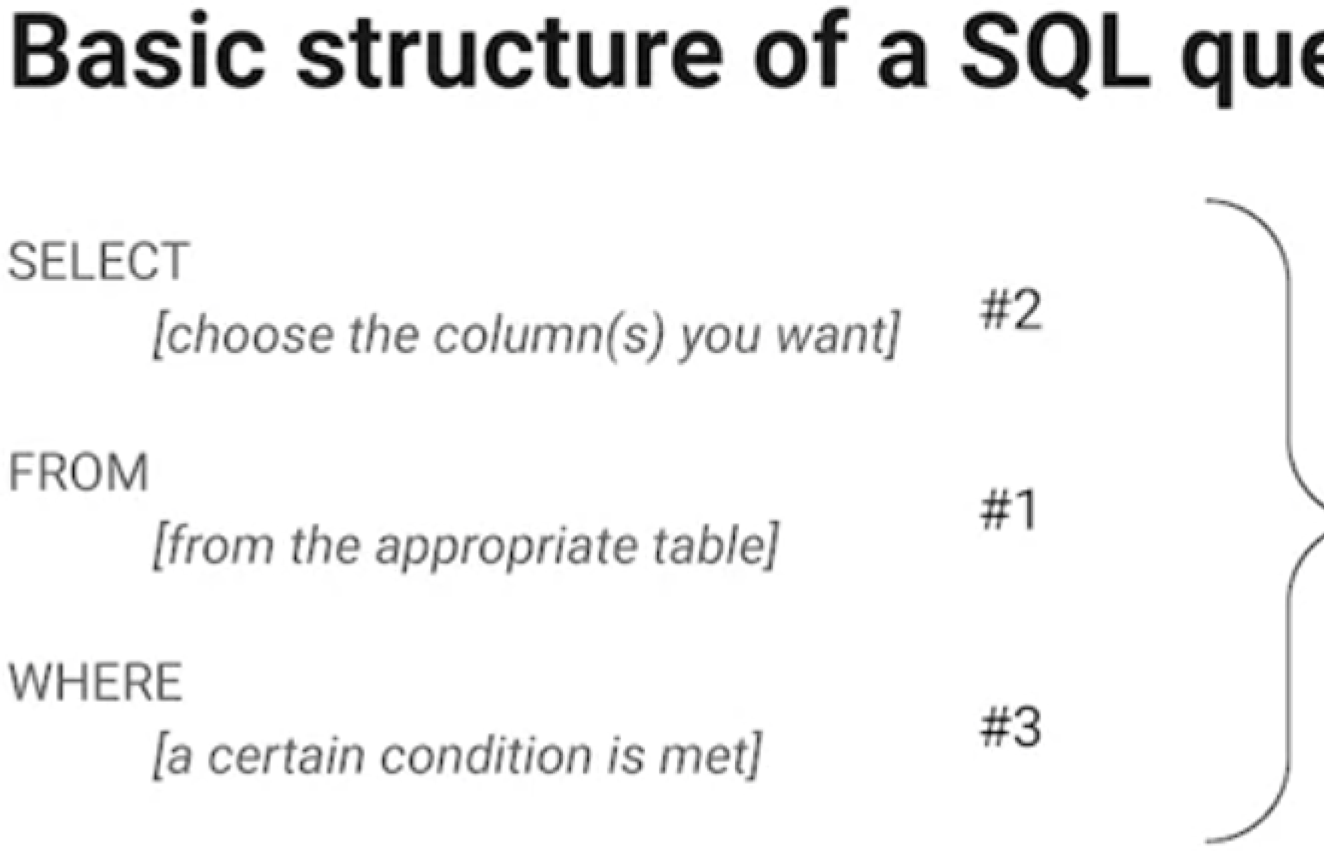
Most important Excel functions exist in Sheets.

AVERAGE	Statistical Returns the numerical average value in a dataset, ignoring text.
AVERAGEIFS	**Statistical** Returns the average of a range that depends upon multiple criteria.
CHOOSE	**Lookup** Returns an element from a list of choices based on index.
COUNT	**Statistical** Returns the count of the number of numeric values in a dataset.
COUNTIF	**Statistical** Returns a conditional count across a range.
DATE	**Date** Converts a provided year, month, and day into a date.
FIND	**Text** Returns the position at which a string is first found within text.
GETPIVOTDATA	**Text** Extracts an aggregated value from a pivot table that corresponds to the specified row and column headings.
IF	**Logical** Returns one value if a logical expression is true and another if it is false.
INDEX	**Lookup** Returns the content of a cell, specified by row and column offset.
INT	**Math** Rounds a number down to the nearest integer that’s less than or equal to it.
LOOKUP	**Lookup** Looks through a row or column for a key and returns the value of the cell in a result range located in the same position as the search row or column.
MATCH	**Lookup** Returns the relative position of an item in a range that matches a specified value.
MAX	**Statistical** Returns the maximum value in a numeric dataset.
MIN	**Statistical** Returns the minimum value in a numeric dataset.
NOW	**Date** Returns the current date and time as a date value.
ROUND	**Math** Rounds a number to a certain number of decimal places according to standard rules.
SUM	**Math** Returns the sum of a series of numbers and/or cells.
SUMIF	**Math** Returns a conditional sum across a range.
TODAY	**Date** Returns the current date as a date value.
VLOOKUP	**Lookup** Searches down the first column of a range for a key and returns the value of a specified cell in the row found.

Learn about SQL

SQL can store, organize and analyze data. Databases that use SQL: Oracle, MySQL, PostgreSQL, Microsoft SQL Server, MangoDB w3schools

Query: The way we use SQL to communicate with the databse, basically a request for data or information from a database.



Basic structure of a SQL que

SELECT
[choose the column(s) you want] #2

FROM
[from the appropriate table] #1

WHERE
[a certain condition is met] #3

```
table AS customers      --Alias to make the work easier

SELECT *                --choose the fields you want to return. Use
`*` to select all the fields in the table
FROM move_data.moveis   --choose the tables where the fields you want
are located
WHERE Genre = 'Action'; /*filter the data based on certain
conditions*/
```

percent sign (`%`) is used as a wildcard to match zero, one or more characters. In some databases the asterisk (`*`) is used as the wildcard instead of the percent sign (%). The underscore(_) represents a single character. The semicolon(`;`) is a statement terminator recommended in ANSI SQL-92. Use `/*` and `*/`, or after two dashes (`--`) to place comments alongside your SQL statements. Use `AS` to alias a field or table name.

```
original AS alias
```

Recommended writing format

Capitalization and indentation: Capitalize SELECT, FROM, and WHERE. Make sure to add a new line and indent when adding the fields.

```
-- never hurts to have comments in a query to remind yourself what
you're trying to do
SELECT
        Column A,    -- you want to look at,
        Column B,    -- provide an overall description
        Column C
FROM
        Table -- the data lives in
WHERE
        Condition 1 -- Certain condition is met
        AND Condition 2
        AND Condition 3
```

Visualizing the data

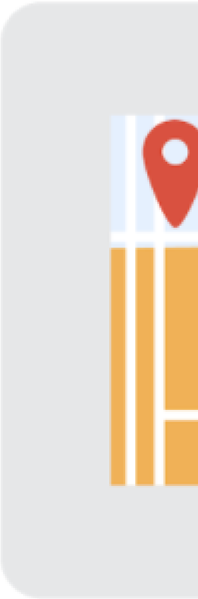
Data analysts use data visualizations to explain complex data quickly, reinforce data analysis, and create interesting graphs and charts.

Steps to plan a data visualization

1. Explore the data for patterns
2. Plan your visuals

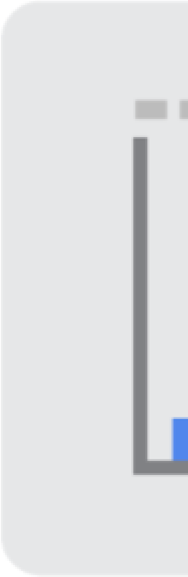
3. Create your visuals

Line charts can track sales over time



Maps

Donut charts can show customer segments



Bar cha
total visit
ma

Build your data visualization toolkit

There are many different tools you can use for data visualization:

- Spreadsheets (Microsoft Excel or Google Sheets) The built-in charts and graphs in spreadsheets made the process of creating visuals quick and easy.
- Visualization software (Tableau) A popular data visualization tool that lets you pull data from nearly any system and turn it into compelling visuals or actionable insights. start exploring Tableau from the How-to Video resources. To explore what other data analysts are sharing on Tableau, visit the Viz of the Day page where you will find beautiful visuals ranging from the Hunt for (Habitable) Planets to Who's Talking in Popular Films.
- Programming language (R with RStudio) RStudio Cheatsheets and the RStudio Visualize Data Primer are great places to start.

Let's get down to business

- The role of a Data Analyst
- Business tasks for Data Analyst
- Fairness in analysis
- Opportunities for you
- And your future success

Data Analyst job opportunities

Technology, Marketing, Finance, Healthcare

data example	industry	
Use geographic data to power GPS technology in cars.	**Technology**	**Technology** relies on software and hardware to function.
Use demographic data to target advertisements for a new consumer product for youths.	**Marketing**	**Marketing** uses audience insights to make decisions.
Use stock market data to determine which portfolios to invest in.	**Finance**	**Finance** relies on daily market trends for insight.
Use bed occupancy data to determine the number of nurses and orderlies to schedule on a given shift.	**Healthcare**	**Healthcare** involves reviewing hospital traffic to inform staff decisions.
Use past booking data to accurately anticipate levels of demand for hotel rooms.	**Hospitality**	**Hospitality** looks at seasonal trends to predict demand.
Use population data to determine which communities need federal funding.	**Government**	**Government** relies on demographic information in order to provide proper support.

The use of data-driven decision-making and the application of data strategy at a corporation are examples of a business task. A business task is the question or problem that data analysis answers for a business.

Defining fairness in data

Analyst should make sure their analysis fair. **Fairness**: Ensuring that your analysis doesn't create or reinforce bias.

Ensuring that analysis does not create or reinforce bias requires using processes and systems that are fair and inclusive to everyone.

As a data analyst, it's your responsibility to make sure your analysis is fair, and factors in the complicated social context that could create bias in your conclusions. (like different proportion in gender)

Considering inclusive sample populations, social context, and self-reported data enable fairness in data collection.

The women doing poorly in the boy company example: First, it doesn't even consider all of the available data on company culture; Second, it doesn't think about the other surrounding factors that impact the data.

Solution: invite a social scientist for consultations; ask women; oversample women in the data; probably apply weights and other positive bias to data for women (clearly communicate the applied bias to stakeholders); normalize; use other metrics while control for income

The fairness measures the team that track patients at risk of cardiovascular disease in a specific area took:

1. Teamed with social scientists who could provide insights on human bias and the social context that created them.
2. Collected self reported data in a separate system to avoid the potential for racial bias, which might skew the results of their study and unfairly represent patients.
3. Oversampled non-dominant groups to make sure their sample population was representative

Exploring your next job

Important factors to think about when searching for the dream job: Industry, Tools, Location, Travel, Culture

Every industry has specific data needs that have to be addressed differently by their data analysts. A great way to guide your search is to think first about what you're interested in. Remember, you want to enjoy what you do, so it's a good idea to think about how you want to use your skills. Then search for jobs that allow you to do that.

Decoding the job description

roles	description
Business Analyst	analyzes data to help businesses improve processes, products, or services

roles	description
Business Intelligence Analyst	analyzes data for finance or market insights
Data Analytics Consultant	analyzes the systems and models for using data
Data Engineer	prepares and integrates data from different sources for analytical use
Data Scientist	uses expert skills in technology and social science to find trends through data analysis
Data Specialist	organizes or converts data for use in databases or software systems
Operations Analyst	analyzes data to assess the performance of business operations and workflows

Decodin

	Data Analy
Problem solving	Use existing tools ar methods to solve pro with existing types o
Analysis	Analyze collected da help stakeholders m better decisions
Other relevant skills	• Database queries • Data visualization • Dashboards • Reports

Three core roles:

- Data analyst: work with SQL, spreadsheets, databases, create dashboard. (The data they used is get from data engineer)
- Data engineer: turn raw data into actionable pipelines
- Data scientist: turn data into cool machine learning models or statistical inferences

-

	industry-specific specialist positions
Marketing analyst	analyzes market conditions to assess the potential sales of products and services
HR/payroll analyst	analyzes payroll data for inefficiencies and errors
Financial analyst	analyzes financial status by collecting, monitoring, and reviewing data
Risk analyst	analyzes financial documents, economic conditions, and client data to help companies determine the level of risk involved in making a particular business decision
Healthcare analyst	analyzes medical data to improve the business aspect of hospitals and medical facilities

Data interview best practices

- Update Linkedln, Github
- Prepare questions for the interviewer, to understand the team and job
- Solve the business problem along with the sample data set. Analyze and come with a solution that relates back to that data.
- Look for the recruiter and hiring manager online, try to reach out to them and send them your resume directly