FOUNDATIONS: DATA DATA EVERYWHERE

DATA ANALYSIS is the collection, transformation and organization of data in order to draw conclusions, make predictions, and drive informed decision-making

DATA ANALYSIS PROCESS
Ask
Prepare
Process
Analyze
Share
Act

TRANSFORMING DATA INTO INSIGHTS

ASK define what project would look like
what would qualify as successful result
ask effective questions
collaborate with stakeholders

PREPARE timeline
how to relay progress to stakeholders
identify data needed to achieve result
brainstorm possible issues and how to avoid them

PROCESS consent to participate
understand how data will be
COLLECTED STORED MANAGED PROTECTED

ANALYZE document what was found no matter the results

SHARE the full picture
communicate results with the right context

ACT work with stakeholders and decide how best to implement changes and take actions based on the findings

collecting and using data ethically is one of the responsibilities of data analysts

restrict access to data
clean data: COMPLETE CORRECT RELEVANT
secure storage

DECISION INTELLIGENCE is a combination of applied data science and the social and managerial sciences

DATA SCIENCE is the discipline of making data useful and is an umbrella term that encompasses 3 disciplines: machine learning, statistics, analytics

UNDERSTANDING THE DATA ECOSYSTEM

ECOSYSTEM is a group of elements that interact with one another

DATA SCIENCE is defined as creating new ways of modelling and understanding the unknown by using raw data

SCIENTISTS create new questions using data

DATA-DRIVEN DECISION-MAKING is defined as using facts to guide business strategy and the first step is figuring out the business need

GUT INSTINCT is an intuitive understanding of something with little or no explanation

If you ignore data by preferring to make decisions based on your own experience, your decisions may be biased
decisions without data to back them up can cause mistakes

DATA ECOSYSTEMS are made up of various elements that interact with one another in order to produce, manage, store, organize, analyze and share data which include hardware and software tools as well as the people who use them

DATA ANALYTICS in the simplest terms is the science of data

ANALYSTS find answers to existing questions by creating insights from data sources

SUBJECT MATTER EXPERTS are people who are familiar with the business problem and have the ability to look at the results of data analysis and identify any inconsistencies, make sense of gray areas and eventually validate choices being made

DATA ANALYSIS is rooted in statistics

DATA ANALYSIS LIFE CYCLE the process of going from data to decision; data goes through several phases as it gets created, consumed, tested, processed and reused

DATA ANALYST SKILLS

ANALYTICAL SKILLS are qualities and characteristics associated with solving problems using facts

FIVE ESSENTIAL SKILLS/ELEMENTS

Curiosity is all about wanting to learn something, seeking out new challenges and experiences leading to knowledge

Technical mindset involves the ability to break things down into smaller steps or pieces and work with them in an orderly and logical way

Context is the condition in which something exists or happens; listening and trying to understand the full picture is critical; group things

Data design is how you organize information; typically has to do with an actual database

Data strategy is the management of the people, processes, and tools used in data analysis:

MANAGE **people** by making sure they know how to use the right data to find solutions to the problems you are working on

processes by making sure the path to the solution is clear and accessible

tools by making sure the right technology is being used for the job

ANALYTICAL THINKING

Analytical thinking involves identifying and defining a problem and then solving it by using data in an organized, step by step manner

FIVE KEY ASPECTS TO ANALYTICAL THINKING

Visualization is the graphical representation of information. It is important because visuals can help data analysts understand and explain information more effectively

Problem-orientation data analysts use a problem-oriented approach in order to identify, describe and solve problems. It is all about keeping the problem top of mind throughout the entire project

Big-picture thinking means being able to see the big picture as well as the details. It is like looking at a complete puzzle where you can enjoy the whole picture without getting stuck on every tiny piece that went into making it. Helps you zoom out and see possibilities and opportunities

Important to think in different ways because in data analysis, solutions are almost never right in front of you. You need to think critically to find out the right questions to ask but you also need to think creatively to get new and unexpected answers

Strategy with so much data available, having a strategic mindset is key to staying focused and on track. Strategizing helps data analysts see what they want to achieve with the data and how they can get there. Strategy also helps improve the quality and usefulness of data we collect

Correlation is like a relationship. It does not equal causation so just because 2 pieces of data are both trending in the same direction, that doesn't necessarily mean they are all related

Detail-oriented thinking is all about figuring out all of the aspects that will help you execute a plan; the pieces that make up your puzzle

Root cause is the reason why a problem occurs

FIVE WHYS PROCESS is where you ask why 5 times to reveal the root cause. The fifth and final answer should give you some useful and sometimes surprising insights

GAP ANALYSIS lets you examine and evaluate how a process works currently in order to get where you want to be in the future.

FOLLOW THE DATA LIFE CYCLE

PLAN happens well before starting a project. During planning, a business decides what kind of data it needs, how it will be managed throughout its life cycle, who will be responsible for it and the optimal outcomes

MANAGE how we care for our data: how and where it is stored, the tools to keep it safe and secure and the actions taken to make sure that it is maintained properly

DESTROY we secure data erasure software; shred paper

CAPTURE this is where data is collected from a variety of different sources and brought into the organization

ANALYZE in this phase, the data is used to solve problems, make great decisions and support business goals

ARCHIVE storing data in a place where it's still available but may not be used again

Database is a collection of data stored in a computer system

OUTLINING THE DATA ANALYSIS PROCESS

Data analysis isn't a life cycle, it's the process of analyzing data

In the **ASK** phase, we define the problem to be solved and make sure that we fully understand stakeholder expectations. Defining a problem means you look at the current state and identify how it's different from the ideal state

In the **PROCESS** step, analysts find and eliminate any errors and inaccuracies that can get in the way of results which usually means cleaning data, transforming it into a more useful format, combining 2 or more datasets to make information more complete and removing outliers or any data points that could skew the information

In the **PREPARE** step, analysts collect and store data to use for the upcoming analysis process. This is where we look at different types of data and identify which kinds are most useful for solving a particular problem

In the **ANALYZE** phase, we use tools to transform and organize the information to draw useful conclusions, make predictions and drive informed decision-making

In the **SHARE** phase, analysts interpret results and share them with others to help stakeholders make effective data-driven decisions

DATA ANALYSIS TOOLBOX

SPREADSHEET is a digital worksheet that stores, organizes and sorts data. Important because the usefulness of data depends on how well it is structured

DATA VISUALIZATION is the graphical representation of information

QUERY LANGUAGE is a computer programming language that allows you to retrieve and manipulate data from a database

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

SPREADSHEET BASICS

COLUMNS are organized vertically and ordered by letter

ROWS are organized horizontally and ordered by number

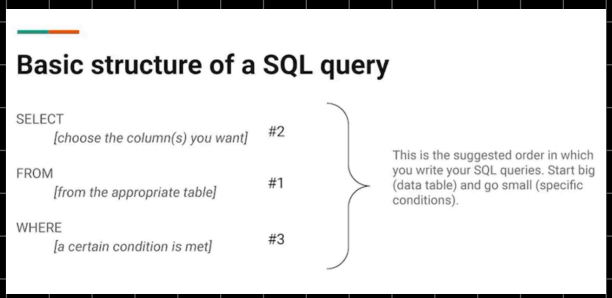
ATTRIBUTES is what column labels are usually called; is a characteristic or a quality of data used to label a column in a table; more commonly referred to as column names, column labels, headers or header row

FORMULA is a set of instructions that performs a specific action using the data in a spreadsheet

SQL

SQL can be used to store, organize and analyze data; structured query language; enables analysts to talk to their databases; can help investigate huge databases, track down text (referred to as **STRINGS**) and numbers, and filter for the exact kind of data you need; follows a unique set of guidelines known as **SYNTAX**

QUERY is a request for data or information from a database



SYNTAX is the predetermined structures of a language that includes all required words, symbols, and punctuation, as well as their proper placement. As soon as you enter your search criteria using the correct syntax, the query starts working to pull the data you've requested from the target database

```
SELECT
      first_name
FROM
      customer_data.customer_name
WHERE
      first_name = 'Tony'
```

SELECT to choose the columns you want to return

FROM to choose the tables where the columns you want are located

WHERE to filter for certain information

* from a table called `customer_name` in a dataset named `customer_data`

MULTIPLE columns that are chosen by the **SELECT** command can be indented and grouped together

```
SELECT
      customer_id,
      first_name,
      last_name
FROM
      customer_data.customer_name
WHERE
      first_name = 'Tony'
```

If you have multiple conditions in your **WHERE** clause:

```
SELECT
      customer_id,
      first_name,
      last_name
FROM
      customer_data.customer_name
WHERE
      customer_id > 0
      AND first_name = 'Tony'
      AND last_name = 'Magnolia'
```

SEMICOLON is a statement terminator and is part of **ANSI** (American National Standards Institute) SQL-92 standard

* using capitalization and indentation can help you read information more easily
* don't worry about extra spaces

WHERE CONDITION if you are looking for a specific word the WHERE clause would be: `WHERE field1 = 'Word'` but if looking for words that start with "Wo" then the WHERE clause would be: `WHERE field1 LIKE 'Wo%'`

LIKE clause allows you to tell the database to look for a certain pattern

% sign is used as a wildcard to match one or more characters

SELECT all columns by using `SELECT *`

COMMENTS are text placed between `/*, */` or `--`; used to help you remember what a name represents; can also be added outside of a statement as well as within

```
SELECT
      field1 /* this is the last name column */
FROM
      table -- this is the customer data table
WHERE
      field1 LIKE 'Ch%';
```

ALIASES are assigning a new name to the column or table name can make them easier to work with and is done with a **SQL AS** clause

```
field1 AS last_name -- Alias to make my work easier
table AS customers -- Alias to make my work easier

SELECT
      last_name
FROM
      customers
WHERE
      last_name LIKE 'Ch%';
```

```
-- This is an important query used later to join with the accounts table
      SELECT
            rowkey,  -- key used to join with account_id
            Info.date,  -- date is in string format YYYY-MM-DD HH:MM:SS
            Info.code  -- e.g., 'pub-###'

      FROM  Publishers
```

* these aliases are good for the duration of the query only

* an alias does not change the actual name of a table/column

<> does not equal

Data Visualization

the graphical representation of information; steps to plan:

EXPLORE THE DATA FOR PATTERNS by getting access to data → **PLAN YOUR VISUALS** by refining the data and presenting results of analysis → **CREATE YOUR VISUALS** by using:

- spreadsheets
- visualization software like TABLEAU
- programming language like R with RStudio

Importance of Fair Business Decisions

ISSUE is a topic to investigate

QUESTION is designed to discover information

PROBLEM is an obstacle or complication that needs to be worked out

BUSINESS TASK is the question or problem analysis answer

FAIRNESS means ensuring that your analysis doesn't create or reinforce bias by creating systems that are fair and inclusive to everyone

ETHICS is not just about minimizing harm but also a concept of beneficience in that we think of how do we improve the lives of people by using data